

---

**Mathematische Modelle für die  
kombinatorische Chemie und die  
molekulare Strukturaufklärung**

---

Der Universität Bayreuth  
zur Erlangung des Grades eines  
Doktors der Naturwissenschaften (Dr. rer. nat.)  
vorgelegte Dissertation

von  
Dipl.-Math.  
Markus Meringer  
aus Hof

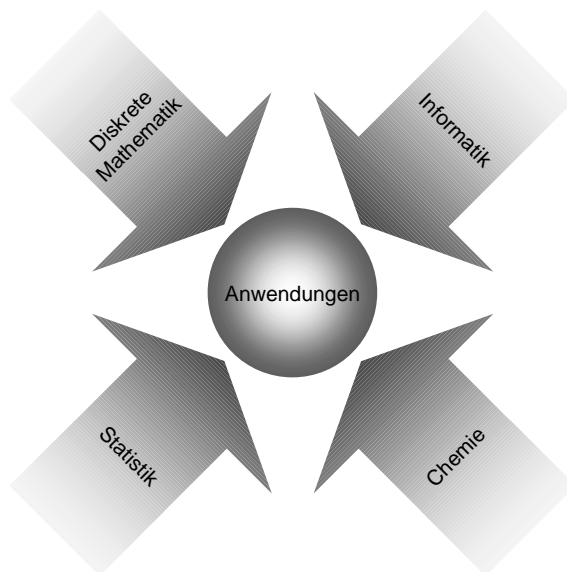
Von der Fakultät für Mathematik und Physik zur Erlangung des Grades  
eines Doktors der Naturwissenschaften genehmigte Dissertation

Erster Gutachter:	Prof. Dr. A. Kerber
Zweiter Gutachter:	Prof. Dr. R. Laue
Tag der Einreichung:	28. Mai 2004
Tag des Kolloquiums:	23. Juli 2004

# Vorwort

Die vorliegende Arbeit dokumentiert die im Rahmen von Forschungsprojekten<sup>1</sup> der DFG und des BMBF entwickelten *mathematischen Modelle* für die Computerchemie. Sie dienen der *kombinatorischen Chemie* und der *molekularen Strukturaufklärung*, und wurden von mir in den Software-Paketen *MOLGEN-QSPR* und *MOLGEN-MS* implementiert.

Ich verwende und erweitere Methoden aus verschiedenen Disziplinen der Diskreten Mathematik (Graphentheorie, konstruktive Kombinatorik), Statistik (explorative Datenanalyse, Lernverfahren), Informatik (Datenstrukturen, Algorithmen) und Chemie (kombinatorische Chemie, molekulare Strukturaufklärung) mit dem Ziel, Lösungen für konkrete Problemstellungen der mathematischen Chemie und der Chemoinformatik zu finden. Schwerpunkte bilden dabei die Suche nach quantitativen Struktur-Eigenschafts-Beziehungen (engl. *Quantitative Structure Property Relationships*, kurz *QSPR*) und die computer-unterstützte Strukturaufklärung (engl. *Computer Aided Structure Elucidation*, kurz *CASE*).



---

<sup>1</sup>DFG Ke201/16-1, Ke201/19-1, BMBF 03KE7BA1-4, 03CO318C

Bei der Erstellung dieser Arbeit wurden folgende Aspekte in besonderem Maße berücksichtigt:

- Die grundlegenden mathematischen Konzepte für die Darstellung und Konstruktion molekularer Strukturen wurden in kompakter, möglichst vollständiger Weise zusammengefasst: Molekulare Graphen, Substrukturen, Restriktionen, Reaktionen, Strukturgenerierung, Deskriptoren. Neben der Darstellung molekularer Strukturen bilden statistische Lernverfahren ein elementares Werkzeug.
- Die wichtigsten neuen Forschungsergebnisse und die Erweiterungen der *MOLGEN*-Klassenbibliothek wurden dokumentiert: Reaktionsbasierte Strukturgenerierung, vergleichende QSPR-Studien für verschiedene Arten molekularer Deskriptoren, verschiedene Methoden zur Suche von Vorhersagefunktionen, Rankingfunktionen und Klassifikatoren für Massenspektren, Spektren-Eigenschafts-Beziehungen, CASE mit hochauflösender Massenspektrometrie.
- Ansatzpunkte und Anregungen für weiterführende Forschungsarbeiten werden skizziert: Neue Ansätze zur Interpretation und Verifikation von Massenspektren, Normalformenproblem im Patentwesen der Chemie.

## Implementierung

Die beschriebenen Datenstrukturen und Algorithmen wurden in der Programmiersprache *C++* [138] implementiert und sind Bestandteil der *MOLGEN*-Klassenbibliothek. Für statistische Verfahren wurde eine Schnittstelle zu der Statistik-Software *R* [69] geschaffen. Numerische Methoden zur linearen Regression wurden der *C++* Bibliothek *TNT* [107] entnommen. Die graphische Benutzeroberfläche wurde unter Verwendung der *MFC* mit *Visual C++ 6* [82] erstellt. Die Visualisierung dreidimensionaler Molekülstrukturen wurde mit *OpenGL* [167] realisiert.

## Danksagungen

Ein interdisziplinäres Projekt des beschriebenen Umfangs ist nur in ständiger Zusammenarbeit mit hoch qualifizierten Teamkollegen erfolgreich realisierbar. Ihnen gebührt an dieser Stelle mein ausdrücklicher Dank.

Wichtige Software-Komponenten wurden von Dr. T. Grüner (zielgerichtete Generierung molekularer Graphen, inkl. Kanonisierer und Substruktursuche, ordnungstreue Erzeugung von Doppelnebenklassen) und Dipl.-Math. J. Braun (molekulare Deskriptoren, Aromatizitätserkennung) zur Verfügung

gestellt. Des Weiteren finden Implementationen von Dr. R. Grund (ordnungstreue Erzeugung molekularer Graphen) und Dipl.–Math. R. Hohberger (Berechnung von zwei- und dreidimensionalen Platzierungen molekularer Graphen) Anwendung. Der 3D–Platzierer wird derzeit von Dipl.–Math. R. Guggisch gepflegt und weiterentwickelt.

Mein Dank beschränkt sich nicht nur auf die „Autoren von Sourcecode“. Ebenso wichtig für die Entstehung dieser Arbeit war ein enormer Transfer chemoinformatischen Know–Hows. Diesbezüglich möchte ich neben vielen „Mitreitern“, deren Namen sich im Literaturverzeichnis wiederfinden, insbesondere PD Dr. C. Rücker sowie Dr. W. Werther und Prof. K. Varmuza hervorheben. Dr. A. Kohnert gebührt Dank für die ausgezeichnete informationstechnische Infrastruktur am Lehrstuhl Mathematik II, Dr. R. Neudert (Chemical Concepts) und Ph.D. S. Stein (National Institute of Standards and Technology) für die zur Verfügung gestellten MS–Datenbanken.

Prof. A. Kerber und Prof. R. Laue haben durch ihren unermüdlichen Einsatz als Initiatoren von Forschungsprojekten an der Schnittstelle von diskreter Mathematik und Computerchemie das wissenschaftliche Fundament und auch die finanzielle Grundlage für die Entstehung dieser Arbeit geschaffen.

Bayreuth, im Mai 2004

*Markus Meringer*



# Inhaltsverzeichnis

Abbildungsverzeichnis	ix
Tabellenverzeichnis	xiii
Symbolverzeichnis	xvii
Einleitung und Übersicht	xxv
<b>I Mathematische Grundlagen</b>	<b>1</b>
<b>1 Diskrete Strukturen in der Chemie</b>	<b>3</b>
1.1 Gruppenoperationen . . . . .	3
1.2 Graphen und Multigraphen . . . . .	7
1.3 Molekulare Graphen . . . . .	16
1.4 Molekulare Substrukturen . . . . .	27
1.5 Chemische Reaktionen . . . . .	32
1.6 Erweiterungen des Molekülmodells . . . . .	37
1.6.1 Mesomerie . . . . .	37
1.6.2 Geometrie . . . . .	41
1.7 Existente chemische Verbindungen . . . . .	43
1.8 Abstraktionsebenen des Molekülmodells . . . . .	47
<b>2 Molekulare Strukturgenerierung</b>	<b>49</b>
2.1 Bruttoformelbasierte Strukturgenerierung . . . . .	50
2.1.1 Ordnungstreue Erzeugung . . . . .	53
2.1.2 Zielgerichtete Erzeugung . . . . .	56
2.2 Reaktionsbasierte Strukturgenerierung . . . . .	57
2.2.1 Zentralmolekül-Ligand-Anlagerungen . . . . .	57
2.2.2 Konstruktion nach dem Netzwerkprinzip . . . . .	58
2.3 Generische Strukturformeln . . . . .	64

<b>3</b>	<b>Überwachtes statistisches Lernen</b>	<b>69</b>
3.1	Variablen und Vorhersagefunktionen . . . . .	71
3.1.1	Regression und Klassifikation . . . . .	72
3.1.2	Bewertung der Vorhersagefunktion . . . . .	74
3.1.3	Datenvorverarbeitung . . . . .	78
3.1.4	Variablen-Selektion . . . . .	79
3.2	Modelle für Vorhersagefunktionen . . . . .	83
3.2.1	Lineare Modelle . . . . .	83
3.2.2	Neuronale Netze . . . . .	85
3.2.3	Support-Vektor-Maschinen . . . . .	86
3.2.4	Entscheidungsbäume . . . . .	88
3.2.5	Nächste Nachbarn . . . . .	90
<b>II</b>	<b>Anwendungen in der Chemie</b>	<b>93</b>
<b>4</b>	<b>Kombinatorische Chemie</b>	<b>95</b>
4.1	Optimierungen in der kombinatorischen Chemie . . . . .	96
4.2	Generierung kombinatorischer Bibliotheken . . . . .	99
4.2.1	Beispiel: Amidierung von Säurechloriden . . . . .	99
4.2.2	Beispiel: Ugi's Siebenkomponentenreaktion . . . . .	104
4.3	Molekulare Deskriptoren . . . . .	108
4.3.1	Arithmetische, topologische und geometrische Indizes . . . . .	109
4.3.2	Substruktur-Vielfachheiten . . . . .	117
4.4	Struktur-Eigenschafts-Beziehungen . . . . .	122
4.4.1	Beispiel: Siedepunkte von Alkanen . . . . .	123
4.4.2	Beispiel: Physikalische Dichte von Propylacrylaten . . . . .	140
4.4.3	Beispiel: Antibakterielle Aktivität von Quinolonen . . . . .	153
4.5	Bemerkungen zur realen Bibliothek . . . . .	164
4.5.1	Diversität der realen Bibliothek . . . . .	164
4.5.2	Teilmengenrelation von realer und virtueller Bibliothek . . . . .	166
<b>5</b>	<b>Molekulare Strukturaufklärung</b>	<b>167</b>
5.1	Spektroskopische Methoden . . . . .	168
5.2	Automatisierte Strukturaufklärung . . . . .	169
5.3	Grundlagen der Massenspektrometrie . . . . .	173
5.3.1	Funktionsweise des EI-Massenspektrometers . . . . .	173
5.3.2	Problemstellungen der EI-Massenspektrometrie . . . . .	175
5.3.3	Massenspektrum und Peakcluster . . . . .	178
5.3.4	Isotope und theoretische Isotopenmuster . . . . .	180
5.3.5	Datenbasis aufgeklärter EI-Massenspektren . . . . .	187



5.4	Rankingfunktionen für Massenspektren . . . . .	191
5.4.1	Ranking von Bruttoformeln . . . . .	196
5.4.2	Ranking von Strukturformeln . . . . .	206
5.5	Klassifikation von Massenspektren . . . . .	218
5.5.1	MS-Deskriptoren . . . . .	218
5.5.2	MS-Klassifikatoren . . . . .	222
5.5.3	Suche nach Substrukturen für MS-Klassifikatoren . . .	236
5.6	Automatisierte Strukturaufklärung mit MS . . . . .	240
5.6.1	Beispiel: n-Pentansäuremethylester . . . . .	240
5.6.2	Beispiel: 3-Hydroxyphenyllessigsäureethylester . . . . .	245
5.7	Quantitative MS-Eigenschafts-Beziehungen . . . . .	249
5.8	Hochauflösende Massenspektrometrie . . . . .	261
<b>6</b>	<b>Patentwesen in der Chemie</b>	<b>271</b>
	<b>Anhang</b>	<b>275</b>
<b>A</b>	<b>Datenstrukturen</b>	<b>277</b>
<b>B</b>	<b>Molekulare Deskriptoren</b>	<b>281</b>
B.1	Arithmetische Indizes . . . . .	281
B.2	Topologische Indizes . . . . .	282
B.3	Geometrische Indizes . . . . .	287
<b>C</b>	<b>Substrukturen für MS-Klassifikatoren</b>	<b>289</b>
C.1	Alkyle . . . . .	291
C.2	Aromaten . . . . .	293
C.3	Bindungen . . . . .	303
C.4	Elemente . . . . .	304
C.5	Funktionelle Gruppen . . . . .	305
C.6	Ringe . . . . .	309
<b>D</b>	<b>Bruttoformeln nach Masse und Ionentyp</b>	<b>311</b>
<b>E</b>	<b>Isomere nach Bruttoformel und Masse</b>	<b>315</b>
	<b>Literaturverzeichnis</b>	<b>327</b>
	<b>Abkürzungsverzeichnis</b>	<b>343</b>
	<b>Index</b>	<b>347</b>



# Abbildungsverzeichnis

1.1	Konstitutionsisomere von $C_6H_6$ in der <i>Beilstein</i> -Datenbank zusammen mit den berechneten Werten für die sterische Energie	44
1.2	Sterische Energie der Konstitutionsisomere von $C_6H_6$	45
1.3	Van der Waals Volumen der Konstitutionsisomere von $C_6H_6$	46
2.1	Alkylreste mit 1–6 C-Atomen	66
3.1	Beispiele starker und schwacher Korrelationen	80
3.2	Schema eines neuronalen Netzes mit einer verborgen Schicht und Bias-Neuronen	85
3.3	Support-Vektor-Klassifikator für den separablen Fall	87
3.4	Schema eines Entscheidungsbaumes	89
4.1	Vorgehensweise bei der Vorhersage von Eigenschaften für vir- tuelle kombinatorische Bibliotheken	97
4.2	Strukturformeln von 20 natürlichen Aminosäuren	100
4.3	Schrittweise Darstellung der Siebenkomponentenreaktion	106
4.4	Atome aus $\mathcal{E}_{11}$ dargestellt als Kugeln mit Van der Waals Radien	115
4.5	Darstellung eines hochsubstituiertes Bernsteinsäureanhydrids als Kalottenmodell mit Van der Waals Radien	115
4.6	Reale Bibliothek von Decanen mit ihren Siedepunkten	119
4.7	Substrukturen mit 2–6 Kanten für die Bibliothek von Decanen aus Abbildung 4.6	121
4.8	Bestimmtheitsmaß für die BP-Modelle	132
4.9	Standardfehler für die BP-Modelle	132
4.10	Empirischer $F$ -Wert für die BP-Modelle	133
4.11	Scatterplot gemessener und vorhergesagter BP für das beste LM mit 3 Deskriptoren (nur TI)	134
4.12	Scatterplot gemessener und vorhergesagter BP für das beste LM mit 4 Deskriptoren (nur SC)	134

4.13	Scatterplot gemessener und vorhergesagter BP für das beste LM mit 3 Deskriptoren (TI und SC) . . . . .	135
4.14	Scatterplot gemessener und vorhergesagter BP für das beste LM mit 7 Deskriptoren (TI und SC) . . . . .	135
4.15	Rein virtuelle Bibliothek von Decanen mit vorhergesagten Siedepunkten . . . . .	139
4.16	Scatterplot von $R_{LS}^2$ und $R_{TS}^2$ für die besten LM bzgl. $R_{LS}^2$ mit $n = 1, \dots, 5$ Deskriptoren . . . . .	146
4.17	Scatterplot gemessener und vorhergesagter PD für das markierte Modell aus Abbildung 4.16 . . . . .	146
4.18	Scatterplot von $R_{LS}^2$ und $R_{TS}^2$ für die besten LM bzgl. $R_{TS}^2$ konstruiert nach 50-fachem schrittweisem Verfahren . . . . .	148
4.19	Scatterplot gemessener und vorhergesagter PD für das markierte Modell aus Abbildung 4.18 . . . . .	148
4.20	Beste Modelle bzgl. $R_{LS}^2$ (bzw. $R_{TS}^2$ ) und zugehörige Werte für $R_{TS}^2$ (bzw. $R_{LS}^2$ ) sowie $R_{CV}^2$ nach Anzahl von Deskriptoren . . . . .	149
4.21	Scatterplot von $R_{LS}^2$ und $R_{CV}^2$ für die besten LM bzgl. $R_{LS}^2$ konstruiert nach 50-fachem schrittweisem Verfahren . . . . .	151
4.22	$R_{LS}^2$ und $R_{TS}^2$ für LM bestimmt durch PCR in Abhängigkeit der Anzahl verwendeter Hauptkomponenten . . . . .	152
4.23	Substituenten für $R^1$ (obere Reihe) und $R^2$ . . . . .	154
4.24	Regressionsbaum für MIC mit 5 Deskriptoren . . . . .	156
4.25	Mehrklassen-CT für MIC mit 7 Deskriptoren . . . . .	158
4.26	2-Klassen-CT für ABA mit 3 Deskriptoren . . . . .	160
4.27	Dendrogramm der virtuellen Bibliothek von Decanen . . . . .	165
5.1	Vorgehensweise bei der automatisierten Strukturaufklärung . . . . .	171
5.2	Beispiel eines Elektronenstoß-Massenspektrums . . . . .	174
5.3	Funktionsweise eines EI-Massenspektrometers . . . . .	174
5.4	Vorgehensweise bei der Strukturaufklärung mittels MS . . . . .	177
5.5	Graphische Darstellung der natürlichen Isotopenverteilungen . . . . .	182
5.6	Molekülmassen der MS-Struktur-Datenbasis zu $\mathcal{E}_{11}$ . . . . .	189
5.7	Molekülmassen der MS-Struktur-Datenbasis zu $\mathcal{E}_4$ . . . . .	189
5.8	Vergleichswerte für die Bruttoformeln mit Masse 116 . . . . .	199
5.9	Histogramm der Bruttoformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren . . . . .	200
5.10	Verteilung der Bruttoformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren . . . . .	200
5.11	Histogramm der RRP für Bruttoformeln von 100 Massenspektren zu Verbindungen aus $\mathcal{E}_4$ . . . . .	203

5.12	Ranking-Position und Kandidaten-Anzahl bei Verlässlichkeit 0,9 für Bruttoformeln zu Verbindungen aus $\mathcal{E}_4$ . . . . .	203
5.13	Histogramm der RRP für Bruttoformeln von 100 Massenspektren zu Verbindungen aus $\mathcal{E}_{11}$ . . . . .	204
5.14	Ranking-Position und Kandidaten-Anzahl bei Verlässlichkeit 0,9 für Bruttoformeln zu Verbindungen aus $\mathcal{E}_{11}$ . . . . .	204
5.15	MS-Reaktionen von n-Pentansäuremethylester . . . . .	208
5.16	Mögliche Fragmentationen von n-Pentansäuremethylester . . . . .	210
5.17	Gemessenes Spektrum und Anteil erklärbarer Intensität im Vergleich . . . . .	211
5.18	Ranking von $C_6H_{12}O_2$ Isomeren bzgl. des Spektrums aus Beispiel 5.3.2 . . . . .	212
5.19	Histogramm der Strukturformel-Vergleichswerte für die Konstitutionsisomere von $C_6H_{12}O_2$ . . . . .	213
5.20	Verteilung der Strukturformel-Vergleichswerte für die Konstitutionsisomere von $C_6H_{12}O_2$ . . . . .	213
5.21	Histogramm der Strukturformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren . . . . .	215
5.22	Verteilung der Strukturformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren . . . . .	215
5.23	Histogramm der RRP für Strukturformeln von 100 Massenspektren . . . . .	217
5.24	Ranking-Position und Kandidaten-Anzahl bei Verlässlichkeit 0,9 für Strukturformeln . . . . .	217
5.25	Vorgehensweise bei der Vorhersage struktureller Eigenschaften durch Spektren-Klassifikation . . . . .	219
5.26	Klassifikationsbaum für Methylester . . . . .	224
5.27	Komplexität der Klassifikationsbäume . . . . .	226
5.28	Mittlere Missklassifikationsraten für Lern- und Testsatz bei Klassifikation durch CT . . . . .	228
5.29	Mittlere Missklassifikationsraten für den Testsatz bzgl. der beiden Klassen bei Klassifikation durch CT . . . . .	228
5.30	Mittlere Missklassifikationsraten für Lern- und Testsatz bei Klassifikation durch LDA . . . . .	231
5.31	Mittlere Missklassifikationsraten für den Testsatz bzgl. der beiden Klassen bei Klassifikation durch LDA . . . . .	231
5.32	Missklassifikationsraten für CT mit $\text{mindev} = 0,04$ und LM mit 13 Deskriptoren . . . . .	233
5.33	Missklassifikationsraten verschiedener Klassifikationsverfahren bei Deskriptoren-Selektion durch CT . . . . .	235

5.34	Missklassifikationsraten verschiedener Klassifikationsverfahren bei schrittweiser Deskriptoren–Selektion durch MLR . . . . .	235
5.35	Vergleichswerte der Strukturkandidaten für das Spektrum aus Beispiel 5.3.2 . . . . .	243
5.36	Ranking von Strukturkandidaten für das Spektrum aus Beispiel 5.3.2 . . . . .	244
5.37	Massenspektrum von 3–Hydroxyphenylethylsäureethylester . . .	245
5.38	Vorgehensweise bei der Vorhersage von Eigenschaften durch Spektren–Eigenschafts–Beziehungen . . . . .	250
5.39	Massenspektren, Strukturen und BP von 29 Decanen . . . . .	251
5.40	Scatterplot gemessener und vorhergesagter BP für das beste LM mit 4 MS–Deskriptoren . . . . .	253
5.41	Scatterplot gemessener und vorhergesagter BP für das beste LM mit 5 MS–Deskriptoren . . . . .	253
5.42	Vorgehensweise bei der Spektren–Verifikation durch quantitative Spektren–Struktur–Beziehungen . . . . .	257
5.43	Vorgehensweise bei der Spektren–Verifikation durch quantitative Struktur–Spektren–Beziehungen . . . . .	258
5.44	Vorgehensweise bei der Spektren–Verifikation durch quantitative Struktur×Spektren–Kompatibilitäts–Beziehungen . . . . .	260
5.45	Minima, Maxima und arithmetische Mittel der Massendifferenzen für Bruttoformeln aus $\mathcal{B}_{\mathcal{E}_4}^C$ . . . . .	264
5.46	Relative Häufigkeiten der MMD für Bruttoformeln aus $\mathcal{B}_{\mathcal{E}_4}^C$ . .	264
5.47	Minima, Maxima und arithmetische Mittel der Massendifferenzen für Bruttoformeln aus $\mathcal{B}_{\mathcal{E}_{11}}^C$ . . . . .	265
5.48	Relative Häufigkeiten der MMD für Bruttoformeln aus $\mathcal{B}_{\mathcal{E}_{11}}^C$ . .	265
5.49	Boxplot der Anzahlen von Bruttoformel–Kandidaten für eine Stichprobe von Verbindungen aus $\mathcal{E}_4$ . . . . .	268
5.50	Plot der Anzahlen von Bruttoformel–Kandidaten und der Molekülmasse für eine Stichprobe von Verbindungen aus $\mathcal{E}_4$ . . .	268
5.51	Boxplot der Anzahlen von Bruttoformel–Kandidaten für eine Stichprobe von Verbindungen aus $\mathcal{E}_{11}$ . . . . .	269
5.52	Plot der Anzahlen von Bruttoformel–Kandidaten und der Molekülmasse für eine Stichprobe von Verbindungen aus $\mathcal{E}_{11}$ . . .	269

# Tabellenverzeichnis

1.1	Gültige Atomzustände für die Elemente aus $\mathcal{E}_{11}$ im Massenspektrometer . . . . .	18
2.1	Reaktanden und Reaktionsschemata für die Generierung des durch $GS$ beschriebenen Strukturraums . . . . .	65
4.1	Amide aus $a$ Aminosäuren und den verschiedenen Zentralmolekülen I, II und III . . . . .	102
4.2	Mittlere Atommasse, Van der Waals Radius und Dichte für die Elemente aus $\mathcal{E}_{11}$ . . . . .	114
4.3	Berechnete Van der Waals Volumina einiger kleiner organischer Moleküle . . . . .	114
4.4	Substruktur-Vielfachheiten für die Bibliothek von Decanen aus Abbildung 4.6 bzgl. der Substrukturen aus Abbildung 4.7	120
4.5	Werte topologischer Indizes für die reale Bibliothek von Decanen aus Abbildung 4.6 . . . . .	124
4.6	Werte topologischer Indizes für die reale Bibliothek von Decanen aus Abbildung 4.6 (fortgesetzt) . . . . .	125
4.7	Ausschnitt der Korrelationsmatrix für Siedepunkte und topologische Indizes der realen Bibliothek von Decanen . . . . .	127
4.8	Charakteristika der besten LM mit $n$ Deskriptoren (aus 18 TI) für Siedepunkte von Decanen . . . . .	128
4.9	Charakteristika der besten LM mit $n$ Deskriptoren (aus 20 SC) für Siedepunkte von Decanen . . . . .	130
4.10	Charakteristika der besten LM mit $n$ Deskriptoren (aus 18 TI und 19 SC) für Siedepunkte von Decanen . . . . .	136
4.11	Beste Teilmengen von $n$ Deskriptoren für BP-Modelle . . . . .	137
4.12	$R^2$ für Modellierung des BP durch verschiedene Regressionsverfahren . . . . .	138
4.13	Atomares Profil der realen Bibliothek von Propylacrylaten . . . . .	141

4.14	PD, Atomanzahl, Molekulargewicht und Van der Waals Volumen für die reale Bibliothek von Propylacrylaten . . . . .	142
4.15	Ausschnitt der Korrelationsmatrix für die physikalische Dichte und Deskriptoren der realen Bibliothek von Propylacrylaten . . . . .	144
4.16	Bestimmtheitsmaß für LS und TS der besten PD-Modelle bzgl. $R_{LS}^2$ mit $n$ Deskriptoren . . . . .	145
4.17	Beste Teilmengen von $n$ Deskriptoren für PD-Modelle . . . . .	147
4.18	Deskriptoren der 25 besten PD-Modelle bzgl. $R_{TS}^2$ berechnet durch 50-fache schrittweise Selektion . . . . .	150
4.19	Gemessene MIC für die reale Bibliothek von Quinolonen . . . . .	153
4.20	$R^2$ für Modellierung von MIC durch verschiedene Deskriptoren und Regressionsverfahren . . . . .	157
4.21	Verteilung gemessener und berechneter MIC für den CT aus Abbildung 4.25 . . . . .	159
4.22	Verteilung gemessener und berechneter ABA für verschiedene Klassifikationsverfahren und Deskriptorensätze . . . . .	161
4.23	$TCE$ und $TCE_{CV}$ für verschiedene Klassifikationsverfahren und Deskriptorensätze . . . . .	162
5.1	Peaks des Massenspektrums aus Abbildung 5.2 . . . . .	180
5.2	Natürliche Isotopenverteilungen für die Elemente aus $\mathcal{E}_{11}$ . . . . .	181
5.3	Atomares Profil der MS-Struktur-Datenbasis zu $\mathcal{E}_{11}$ . . . . .	188
5.4	Atomanzahlen und Molekülmassen der MS-Struktur-Datenbasis . . . . .	188
5.5	Berechnung des Vergleichswertes für $C_6H_{12}O_2$ zu dem Spektrum aus Beispiel 5.3.2 . . . . .	197
5.6	Ranking von Bruttoformeln mit Masse 116 zu dem Spektrum aus Beispiel 5.3.2 . . . . .	198
5.7	Quantile $q_p$ für Bruttoformel-Vergleichswerte zu verschiedenen Wahrscheinlichkeiten $p$ . . . . .	201
5.8	Berechnung des Vergleichswertes für n-Pentansäuremethylester zu dem Spektrum aus Beispiel 5.3.2 . . . . .	209
5.9	Quantile $q_p$ für Strukturformel-Vergleichswerte zu verschiedenen Wahrscheinlichkeiten $p$ . . . . .	216
5.10	Details der Knoten des Klassifikationsbaums für Methylester . . . . .	225
5.11	Deskriptorenwerte für das Spektrum aus Beispiel 5.3.2 . . . . .	225
5.12	Missklassifikationsraten von MS-Klassifikatoren (CT) für 77 strukturelle Eigenschaften . . . . .	229
5.13	Missklassifikationsraten von MS-Klassifikatoren (LDA) für 77 strukturelle Eigenschaften . . . . .	232



5.14	Missklassifikationsraten verschiedener Klassifikationsverfahren bei Deskriptoren–Selektion durch CT . . . . .	234
5.15	Missklassifikationsraten verschiedener Klassifikationsverfahren bei schrittweiser Deskriptoren–Selektion durch MLR . . . . .	234
5.16	Missklassifikationsraten von MS–Klassifikatoren (SVM mit radialem Kernel) für 77 strukturelle Eigenschaften . . . . .	237
5.17	Missklassifikationsraten für gut klassifizierbare Substrukturen .	238
5.18	$R^2$ und $R_{CV}^2$ verschiedener MS–BP–Beziehungen . . . . .	254
5.19	$R^2$ und $R_{CV}^2$ verschiedener optimierter MS–BP–Beziehungen .	255
5.20	Hochaufgelöste Isotopenmassen und Isotopenverteilungen für die Elemente aus $\mathcal{E}_{11}$ . . . . .	262
5.21	Anzahlen von Bruttoformel–Kandidaten für eine Stichprobe von Verbindungen aus $\mathcal{E}_4$ bzw. $\mathcal{E}_{11}$ . . . . .	267
D.1	Anzahlen von Bruttoformeln zu nominalen Massen von 1 bis 100 mit Elementen aus $\mathcal{E}_4$ . . . . .	312
D.2	Anzahlen von Bruttoformeln zu nominalen Massen von 1 bis 100 mit Elementen aus $\mathcal{E}_{11}$ . . . . .	313
D.3	Anzahlen von Bruttoformeln zu nominalen Massen über 100 mit Elementen aus $\mathcal{E}_4$ . . . . .	314
D.4	Anzahlen von Bruttoformeln zu nominalen Massen über 100 mit Elementen aus $\mathcal{E}_{11}$ . . . . .	314



# Symbolverzeichnis

$id$	Identität, 3
$\Omega_G$	Operation von $G$ auf $\Omega$ , 4
$\mathbb{N}$	Menge $\{0, 1, 2, \dots\}$ der natürlichen Zahlen, 4
$\mathbb{N}^*$	Menge $\{1, 2, \dots\}$ der natürlichen Zahlen ohne 0, 4
$n$	Menge $\{0, \dots, n - 1\}$ oder ihre Kardinalität, 4
$S_\Omega$	Symmetrische Gruppe auf $\Omega$ , 4
$\omega^G$	Bahn von $\omega$ unter $\Omega_G$ , 4
$\Omega//G$	Menge der Bahnen von $\Omega_G$ , 4
$ \Omega $	Kardinalität der Menge $\Omega$ , 4
$\mathcal{T}(\Omega//G)$	Menge der Transversalen von $\Omega_G$ , 4
$A \leq G$	Untergruppe von $G$ , 5
$gA$	Linksnebenklasse, 5
$G/A$	Linksnebenklassen von $A$ in $G$ , 5
$Ag$	Rechtsnebenklasse, 5
$A \setminus G$	Rechtsnebenklassen von $A$ in $G$ , 5
$AgB$	Doppelnebenklasse, 5
$A \setminus G/B$	Doppelnebenklassen von $A$ und $B$ in $G$ , 5
$Y^X$	Menge der Abbildungen von $X$ nach $Y$ , 6
$\gamma$	eine diskrete Struktur, insbesondere ein Graph, 6
$\text{Aut}(\gamma)$	Automorphismengruppe von $\gamma$ , 6
$\binom{V}{2}$	Menge der 2-Teilmengen von $V$ , 7
$\mathcal{G}_{V,m}$	Menge der Graphen $m^{\binom{V}{2}}$ , 7
$\bar{\gamma}$	Isomorphieklasse von $\gamma$ , 7
$\kappa$	eine kanonische Form, 7
$A_\gamma$	Adjazenzmatrix von $\gamma$ , 7
$E_\gamma$	Menge der Kanten von $\gamma$ , 8

$\deg_\gamma(v)$	Grad des Knotens $v$ von $\gamma$ , 8
$\lambda \models n$	$\lambda$ ist Partition von $n$ , 8
$\lambda_\gamma$	Gradpartition von $\gamma$ , 8
$\text{hyb}_\gamma(v)$	Hybridisierung von $v$ in $\gamma$ , 9
$\text{len}_\gamma(W)$	Länge des Kantenzugs $W$ in $\gamma$ , 10
$\mathcal{G}_{V,m}^C$	Menge der zusammenhängenden Graphen in $\mathcal{G}_{V,m}$ , 10
$\text{dist}_\gamma(v, w)$	Abstand der Knoten $v, w$ in $\gamma$ , 10
$\text{girth}_\gamma$	Tailenweite von $\gamma$ , 11
$\gamma' \subseteq \gamma$	$\gamma'$ ist Subgraph von $\gamma$ , 12
$\gamma' \subseteq^c \gamma$	$\gamma'$ ist geschlossener Subgraph von $\gamma$ , 12
$\gamma' \subseteq^i \gamma$	$\gamma'$ ist induzierter Subgraph von $\gamma$ , 12
$\gamma _{V'}$	auf $V'$ induzierter Teilgraph von $\gamma$ , 12
$\text{Con}(\gamma)$	Menge der durch die Zusammenhangskomponenten von $\gamma$ induzierten Teilgraphen, 12
$\phi$	eine Einbettung von Graphen, 13
$Y_{\text{inj}}^X$	Menge der injektiven Abbildungen von $X$ nach $Y$ , 13
$\gamma' \subseteq_\phi \gamma$	$\gamma'$ ist Subgraph von $\gamma$ bzgl. $\phi$ , 13
$\gamma' \subseteq_\phi^c \gamma$	$\gamma'$ ist geschlossener Subgraph von $\gamma$ bzgl. $\phi$ , 13
$\gamma' \subseteq_\phi^i \gamma$	$\gamma'$ ist induzierter Subgraph von $\gamma$ bzgl. $\phi$ , 13
$\text{Emb}_\subseteq(\gamma', \gamma)$	Menge der Einbettungen von $\gamma'$ in $\gamma$ als Subgraph, 13
$\text{Emb}_\subseteq^c(\gamma', \gamma)$	Menge der Einbettungen von $\gamma'$ in $\gamma$ als geschlossener Subgraph, 13
$\text{Emb}_\subseteq^i(\gamma', \gamma)$	Menge der Einbettungen von $\gamma'$ in $\gamma$ als induzierter Subgraph, 13
$Z$	ein Atomzustand, 16
$v_Z$	Valenz von $Z$ , 16
$p_Z$	Anzahl freier Elektronenpaare von $Z$ , 16
$q_Z$	Ladung von $Z$ , 16
$r_Z$	Vorhandensein eines ungepaarten Elektrons bei $Z$ , 16
$\mathbb{Z}$	Menge der ganzen Zahlen, 16
$\mathbb{B}$	boolsche Algebra $\{\text{falsch}, \text{wahr}\}$ bzw. $\{0, 1\}$ , 16
$X$	ein chemisches Element, 17
$\mathcal{Z}_X$	Menge gültiger Atomzustände für $X$ , 17
$\text{H}$	chemisches Element Wasserstoff, 17
$\text{C}$	chemisches Element Kohlenstoff, 17

N	chemisches Element Stickstoff, 17
O	chemisches Element Sauerstoff, 17
$\mathcal{E}_4$	Menge {H, C, N, O}, 17
F	chemisches Element Fluor, 17
Si	chemisches Element Silizium, 17
P	chemisches Element Phosphor, 17
S	chemisches Element Schwefel, 17
Cl	chemisches Element Chlor, 17
Br	chemisches Element Brom, 17
I	chemisches Element Jod, 17
$\mathcal{E}_{11}$	Menge {H, C, N, O, F, Si, P, S, Cl, Br, I}, 17
$TE_X$	Ordnungszahl von $X$ , 17
$VE_X$	Anzahl der Valenzelektronen von $X$ , 17
$v_X$	Standardvalenz von $X$ , 19
$\mathcal{E}$	eine Menge chemischer Elemente, 20
$\mathcal{Z}_{\mathcal{E}}$	Menge gültiger Atomzustände für Elemente aus $\mathcal{E}$ , 20
$\zeta$	eine Zustandsverteilung, 20
$\varepsilon$	eine Elementverteilung, 20
$\mathcal{M}_n$	Menge der molekularen Graphen mit $n$ Atomen, 20
$\mathcal{M}$	Menge der molekularen Graphen, 20
$M$	ein molekularer Graph, 20
$\mathcal{M}_n^C$	Menge der zshg. molekularen Graphen in $\mathcal{M}_n$ , 20
$\mathcal{M}^C$	Menge der zshg. molekularen Graphen in $\mathcal{M}$ , 20
$\bar{M}$	Isomorphieklasse von $M$ , 22
$\beta$	eine Bruttoformel, 23
$\beta_M$	Bruttoformel eines molekularen Graphen $M$ , 23
$\bar{\mathcal{M}}_{\beta}^C$	Menge der Konstitutionsisomere zu $\beta$ , 23
$a \equiv b \pmod{c}$	$a$ hat Rest $b$ bei ganzzahliger Division durch $c$ , 24
$\mathcal{B}_{\mathcal{E}}^C$	Menge der Bruttoformeln zusammenhängender molekularer Graphen über $\mathcal{E}$ , 24
$\text{DBE}(\beta)$	Doppelbindungsäquivalent von $\beta$ , 25
$\beta' \subseteq \beta$	Teilmengenrelation für Bruttoformeln, 25
$\text{ggT}(\Omega)$	größter gemeinsamer Teiler der Elemente von $\Omega$ , 25
$MMG$	ein mehrdeutiger molekularer Graph, 27

$MM\mathcal{G}_n$	Menge der mehrdeutigen molekularen Graphen mit $n$ Atomen, 27
$\mathcal{P}(\Omega)$	Potenzmenge von $\Omega$ , 27
$\mathcal{P}^*(\Omega)$	Potenzmenge von $\Omega$ ohne die leere Menge, 27
$\underline{n}$	Menge $\{1, \dots, n\}$ , 27
$\emptyset$	leere Menge, 27
$MMG \subseteq_{\phi} M$	$MMG$ ist mehrdeutiger molekularer Subgraph von $M$ bzgl. $\phi$ , 29
$MMG \subseteq_{\phi}^i M$	$MMG$ ist mehrdeutiger molekularer Teilgraph von $M$ bzgl. $\phi$ , 29
$SR$	eine Substruktur–Restriktion, 30
$\mathcal{SR}_k$	Menge der Substruktur–Restriktionen auf $k$ Atomen, 30
$[a, b]$	Intervall ganzer Zahlen $\{a, \dots, b\}$ , 30
$SR_{\{i,j\},[a,b]}^{\text{Dist}}$	eine Substruktur–Restriktion Distanz, 30
$\eta$	eine Hybridisierung, 30
$SR_{\{i_j j \in h\},\eta}^{\text{Hybrid}}$	eine Substruktur–Restriktion Hybridisierung, 30
$S$	eine molekulare Substruktur, 31
$\mathcal{S}_k$	Menge der molekularen Substrukturen auf $k$ Atomen, 31
$S \subseteq_{\phi} M$	$S$ ist molekulare Substruktur von $M$ bzgl. $\phi$ , 31
$S \subseteq_{\phi}^i M$	$S$ ist molekulare Teilstruktur von $M$ bzgl. $\phi$ , 31
$\text{Emb}_{\subseteq}(S, M)$	Menge der Einbettungen von $S$ in $M$ als molekulare Substruktur, 31
$\text{Emb}_{\subseteq}^i(S, M)$	Menge der Einbettungen von $S$ in $M$ als molekulare Teilstruktur, 31
$C$	eine chemische Reaktion, 32
$\mathcal{C}_n$	Menge der chemischen Reaktionen auf $n$ Atomen, 32
$\Delta C$	ein Reaktionsänderungsgraph, 32
$\Delta \zeta$	eine Zustandsänderungsverteilung, 32
$\Delta \gamma$	ein Bindungsänderungsgraph, 32
$\Delta \mathcal{Z}$	Menge der Zustandsänderungen, 32
$\mathcal{G}_{n,[-3,3]}$	Menge der Bindungsänderungsgraphen, 32
$\Delta \mathcal{C}_n$	Menge der Reaktionsänderungsgraphen auf $n$ Atomen, 32
$a \dot{\vee} b$	ausschließendes <i>oder</i> zweier boolescher Ausdrücke, 33
$\Delta C \circ M$	Anwendung von $\Delta C$ auf $M$ , 33
$\text{Cen}(C)$	Reaktionszentrum von $C$ , 33
$\text{RCG}(C)$	Reaktionszentrumsgraph von $C$ , 34

$R$	ein Reaktionsschema, 35
$\mathcal{R}_k$	Menge der Reaktionsschemata auf $k$ Atomen, 35
$R \circ_\phi M$	Anwendung von $R$ auf $M$ bzgl. $\phi$ , 35
$\text{Prod}_R(M)$	Menge der Produktgraphen bei der Anwendung von $R$ auf $M$ , 36
$\xi$	3D-Koordinaten eines Moleküls, 41
$\mathbb{R}$	Menge der reellen Zahlen, 41
$\mathcal{I}(\mathbb{N})$	Menge der Intervalle natürlicher Zahlen, 50
$B$	eine weiche Bruttoformel, 50
$\mathcal{B}_B$	Menge der zu $B$ kompatiblen Bruttoformeln, 50
$\text{rep}_<(\Omega//G)$	Transversale minimaler Bahnrepräsentanten, 53
$\mathcal{L}$	eine Menge molekularer Graphen, 58
$m_{\leq}^n$	Menge der schwach monotonen Abbildungen, 59
$\text{depth}_{\mathcal{R}}$	Tiefe für Reaktionsschemata, 60
$\text{cha}(M)$	Summe der Ladungen der Atome von $M$ , 61
$\text{depth}_{\mathcal{L}}$	Tiefe für Reaktanden, 62
$\text{size}(M)$	Anzahl der Atome von $M$ , 62
$GS$	eine generische Strukturformel, 64
$X_j$	unabhängige Variable (Vorhersagevariable), 71
$Y$	abhängige Variable (Zielvariable), 71
$x_{ij}$	Wert der Vorhersagevariable $X_j$ für Beobachtung $i$ , 71
$y_i$	Wert der Zielvariable $Y$ für Beobachtung $i$ , 71
$\mathbf{X}$	$m \times n$ -Matrix der mit Einträgen $x_{ij}$ , 71
$\mathbf{x}_i$	$i$ -ter Zeilenvektor von $\mathbf{X}$ , 71
$\mathbf{Y}$	$m \times 1$ -Matrix mit Einträgen $y_i$ , 71
$f$	eine Vorhersagefunktion, 71
$RSS$	Summe der Quadrate der Residuen, 72
$\mathcal{C}$	Menge verschiedener Klassen oder ihre Kardinalität, 72
$L$	eine Kostenfunktion, 72
$\mathbb{R}_0^+$	Menge der nicht negativen reellen Zahlen, 72
$\delta$	Kroneckersche Deltafunktion, 73
$TCE$	Gesamt-Klassifikationsfehler, 73
$CE^{(k)}$	Klassifikationsfehler für Klasse $k$ , 73
$R$	multipler Korrelationskoeffizient, 74

$\bar{y}$	arithmetisches Mittel, 74
$R^2$	Bestimmtheitsmaß (einer Regression), 74
$S$	Standardfehler (einer Regression), 75
$d$	Anzahl von Freiheitsgraden, 75
$F$	empirischer $F$ -Wert (einer Regression), 75
$MCE$	mittlerer Klassifikationsfehler, 75
$MCE^{(k)}$	MCE für Klasse $k$ , 75
$LS$	Lernsatz, 76
$TS$	Testsatz, 76
$A \cup B$	disjunkte Vereinigung zweier Mengen, 76
$f_{LS}$	Vorhersagefunktion für den Lernsatz, 76
$RSS_{TS}$	$RSS$ für den Testsatz, 76
$R_{TS}^2$	Bestimmtheitsmaß für den Testsatz, 76
$TCE_{TS}$	$TCE$ für den Testsatz, 76
$MCE_{TS}$	$MCE$ für den Testsatz, 76
$CE_{TS}^{(k)}$	Klassifikationsfehler für Klasse $k$ im Testsatz, 76
$MCE_{TS}^{(k)}$	$MCE$ für Klasse $k$ im Testsatz, 76
$A \setminus B$	logische Differenz zweier Mengen, 77
$RSS_{kCV}$	$RSS$ für $k$ -fache Kreuzvalidierung, 77
$TCE_{kCV}$	$TCE$ für $k$ -fache Kreuzvalidierung, 77
$RSS_{CV}$	$RSS$ für LOO-Kreuzvalidierung, 77
$R_{CV}^2$	Bestimmtheitsmaß für LOO-Kreuzvalidierung, 77
$S_{CV}$	Standardfehler für LOO-Kreuzvalidierung, 77
$TCE_{CV}$	$TCE$ für LOO-Kreuzvalidierung, 78
$MCE_{CV}$	$MCE$ für LOO-Kreuzvalidierung, 78
$\ \cdot\ _2$	euklidische Norm, 79
$R(X, Z)$	Korrelationskoeffizient von $X$ und $Z$ , 80
$FR(X, Y)$	Fisher Ratio von $X$ und $Y$ , 81
$\text{diag}(a_i)$	Diagonalmatrix mit Einträgen $a_i$ , 84
$N_k(\mathbf{x})$	$k$ nächste Nachbarn von $\mathbf{x}$ , 91
$D$	ein molekularer Deskriptor, 108
amu	atomare Masseneinheit, 109
$\text{deg}_\gamma^d(i)$	Knotendistanzgrad des Knotens $i$ in $\gamma$ , 111
$\text{deg}_M^v(i)$	Knotenvalenzgrad des Atoms $i$ in $M$ , 112



$HC_M(i)$	Anzahl zu Atom $i$ benachbarter H-Atome in $M$ , 112
$\text{Å}$	Ångström, 114
$\Psi$	eine QSPR-Funktion, 122
$\ln$	natürlicher Logarithmus, 143
$I$	Massenspektrum, Isotopenmuster, 178
$\hat{m}$	größte Masse mit Intensität $> 0$ , 178
$\tilde{m}$	Masse größter Intensität, 178
$P$	Peak eines Massenspektrums, 178
$\tilde{P}$	Peak größter Intensität, Basispeak, 178
$\hat{P}$	Peak größter Masse, 178
$\mathcal{P}$	eine Menge von Peaks, ein Peakcluster, 179
${}^mX$	Isotop der Masse $m$ von $X$ , 181
$I_X$	natürliche Isotopenverteilung von $X$ , 181
$\hat{m}_X$	größte Isotopenmasse von $X$ , 181
$\tilde{m}_X$	nominale Masse von $X$ , 181
$\check{m}_X$	kleinste Isotopenmasse von $X$ , 181
$\mathcal{I}$	Menge der Isotopenmuster, 184
$I_\beta$	theoretisches Isotopenmuster von $\beta$ , 185
$m_\beta$	nominale Masse von $\beta$ , 186
$\tilde{m}_\beta$	Masse größter Intensität von $\beta$ , 186
$\hat{m}_\beta$	größte Masse von $\beta$ , 186
$\check{m}_\beta$	kleinste Masse von $\beta$ , 186
$K$	Kandidat für die Brutto- oder Strukturformel, 192
$MV(I, K)$	Vergleichswert für $K$ bzgl. $I$ , 192
$K^T$	korrekter Kandidat für die Brutto-/Strukturformel, 192
$K^F$	falscher Kandidat für die Brutto-/Strukturformel, 192
$q_p$	$p$ -Quantil, 199
$RRP$	relative Ranking-Position, 201
$\Phi$	ein MS-Klassifikator, 218
$MCE_{TS}^F$	$MCE_{TS}$ beschränkt auf Klasse $F$ , 227
$MCE_{TS}^W$	$MCE_{TS}$ beschränkt auf Klasse $W$ , 227
$\bar{m}_X$	mittlere Atommasse von $X$ , 262



# Einleitung und Übersicht

Mathematische Modelle bilden eine unentbehrliche Grundlage in nahezu allen Bereichen von Wissenschaft und Technik. Ohne sie wären Reisen zu entfernten Gestirnen ebenso wenig möglich wie die Entschlüsselung des menschlichen Erbguts.

Immer öfter werden auch Problemstellungen der organischen Chemie durch mathematische Modellierung simuliert und gelöst. Insbesondere kann die Darstellung chemischer Verbindungen durch molekulare Graphen als einer der Ursprünge der Graphentheorie betrachtet werden.

Zwei wichtige Teildisziplinen in der organischen Chemie bilden die *Synthese* und die *Analytik*. Die Aufgabe der Synthese besteht darin, aus bekannten chemischen Bausteinen über Reaktionen neue Verbindungen zu erschaffen. Die Analytik beschäftigt sich unter anderem damit, Eigenschaften chemischer Verbindungen zu bestimmen.

Motivation ist dabei oft die Suche nach neuen Wirkstoffen und Materialien mit angestrebten biologisch-pharmazeutischen oder physiko-chemischen Eigenschaften. Wurde eine entsprechende Verbindung synthetisiert und gefunden, besteht eine weitere wichtige Aufgabe darin, die *molekulare Struktur* der oft noch unbekannt Substanz zu bestimmen. Zu diesem Zweck verwendet man hauptsächlich spezielle physiko-chemische Eigenschaften, die aus *spektroskopischen* Methoden gewonnen werden.

Bei der Suche nach neuen Wirkstoffen und Materialien finden immer häufiger Techniken der *kombinatorischen Chemie* Verwendung. Dabei werden aus mehreren Sätzen chemischer Bausteine sämtliche Kombinationen synthetisiert und anschließend auf ihre biologisch-pharmazeutische Wirksamkeit getestet oder hinsichtlich angestrebter physiko-chemischer Eigenschaften untersucht.

Die innerhalb eines kombinatorisch-chemischen Experiments synthetisierten Verbindungen werden *kombinatorische Bibliothek* genannt, die Testung allgemein als *Screening* bezeichnet. Der enorme Vorteil dieser Methode besteht darin, dass sowohl die Synthese als auch das Screening in hohem Maße automatisiert und parallelisiert werden kann. Obwohl somit erstaunliche Durch-

satzraten erzielt werden, mahnen Kosten- und Zeitgründe zur gewissenhaften Planung und zur automatischen Auswertung derartiger Experimente.

Die Optimierung kombinatorisch-chemischer Experimente und die automatisierte molekulare Strukturbestimmung werfen vielfältige Probleme auf, die nach mathematischer Modellierung mit Hilfe von algebraisch-kombinatorischen Konstruktionsalgorithmen, graphentheoretischen Invarianten und statistischen Lernverfahren gelöst werden können. Die entscheidenden Problemstellungen betreffen die

- **Strukturgenerierung:**

In der kombinatorischen Chemie benötigt man Methoden, um virtuelle kombinatorische Bibliotheken zu konstruieren. Meist werden solche Strukturräume durch *Reaktanden* und *Reaktionen* definiert. In dieser Arbeit werden Algorithmen zur *reaktionsbasierten* Strukturgenerierung beschrieben.

Für die molekulare Strukturaufklärung werden Algorithmen verwendet, die ausgehend von der *Bruttoformel* eines Analyten unter Berücksichtigung struktureller *Restriktionen* mögliche Strukturformeln generieren können. Strategien zur bruttoformelbasierten Strukturgenerierung werden vorgestellt.

Wichtige Methoden für beide Problemkreise bilden *kanonische Nummerierung* und *ordnungstreue Erzeugung*.

- **Suche nach Beziehungen zwischen Struktur und Eigenschaft:**

Um Eigenschaften für die Strukturen virtueller kombinatorischer Bibliotheken vorhersagen zu können, verwendet man *quantitative Struktur-Eigenschafts-Beziehungen (QSPR)*. Diese werden zuvor anhand einer tatsächlich synthetisierten und gescreenten kleineren *realen* Bibliothek ermittelt.

Die *computer-unterstützte molekulare Strukturaufklärung (CASE)* verfolgt das Ziel, ausgehend von spektroskopisch gemessenen Eigenschaften einer unbekanntem chemischen Verbindung ihre molekulare Struktur zu bestimmen.

Die mathematischen Werkzeuge für diese Aufgaben sind *molekulare* bzw. *spektrale Deskriptoren* und *statistische Lernverfahren*.

Das Innovationspotential dieser Arbeit liegt hauptsächlich in der Kombination von Methoden zur Lösung der beiden obigen Problemstellungen. Im Folgenden wird der Inhalt der einzelnen Kapitel kurz skizziert. Kapitel 1–3 enthalten grundlegende mathematische Modelle und Lösungen. In den nachfolgenden Kapiteln werden diese auf konkrete Probleme aus der Chemie angewendet.

## 1 Diskrete Strukturen in der Chemie

In der vorliegenden Arbeit werden chemische Verbindungen als *molekulare Graphen* dargestellt. Ein molekularer Graph ist ein ungerichteter schleifenfreier Multigraph mit speziell beschrifteten Knoten. Die Knoten entsprechen den Atomen, die Kanten den kovalenten Bindungen. Jeder Knoten ist bewertet mit einem Elementsymbol und einem Atomzustand. Der Atomzustand setzt sich zusammen aus der Valenz, der Anzahl freier Elektronenpaare<sup>2</sup>, der Ladung des Atoms und der Information über das Vorhandensein eines einsamen Elektrons. Die Kantenvielfachheiten spiegeln die verschiedenen Bindungstypen wieder. Eine eindeutige Zuordnung zwischen chemischen Verbindungen und mathematischen Strukturen wird erreicht, indem man *Strukturformeln* chemischer Verbindungen mit *Isomorphieklassen* molekularer Graphen identifiziert.

*Molekulare Substrukturen* dienen dazu, strukturelle Eigenschaften molekularer Graphen zu beschreiben. Dabei können Alternativen für Elemente und Atomzustände<sup>3</sup> sowie Kantenvielfachheiten berücksichtigt werden. Zusätzliche topologische Nebenbedingungen wie Abstände zwischen Atomen, Hybridisierungen oder die An- und Abwesenheit von Ringen vorgegebener Länge ermöglichen in Form von *Substruktur-Restriktionen* eine sehr exakte Spezifizierung struktureller Eigenschaften.

Molekulare Substrukturen werden unter anderem zur Definition von *Reaktionsschemata*<sup>4</sup> verwendet. Zusammen mit den durch eine chemische Reaktion bewirkten Änderungen der Atomzustände und der Bindungsvielfachheit bilden sie eine geeignete Syntax, um chemische Reaktionen graphisch beschreiben und qualitativ simulieren zu können.

Erweiterungen des Molekülmodells zur Darstellung mesomerer Strukturen und zur Berücksichtigung geometrischer Aspekte werden kurz besprochen.

In der Regel ist nur eine geringe Menge mathematisch möglicher Konstitutionen als real existent nachgewiesen und in Datenbanken erfasst. Wir betrachten diesen Sachverhalt am Beispiel von  $C_6H_6$  und ziehen exemplarisch berechnete Energiewerte und van der Waals Volumina als mögliche Erklärungskomponenten für diese Tatsache heran.

---

<sup>2</sup>Dieses Merkmal wurde im Vergleich zu [53] zusätzlich in die Beschreibung des Atomzustands aufgenommen. Es kann aus Valenz, einsamen Elektron, Ladung und Element berechnet werden und ist wichtig zur Simulation bestimmter Reaktionen.

<sup>3</sup>Im Zuge dieser Arbeit wurden hier ein spezieller Atomtyp eingeführt, der die Simulation von Fragmentierungsreaktionen im Massenspektrometer wesentlich vereinfacht.

<sup>4</sup>Als Erweiterung zu einer früheren Arbeit [164] können nunmehr auch Zerfalls- und Umlagerungsreaktionen durch Reaktionsschemata dargestellt werden.

## 2 Molekulare Strukturgenerierung

Die Generierung molekularer Graphen mit vorgegebenen Eigenschaften bis auf *Isomorphie* bildet ein wichtiges Werkzeug im Rahmen dieser Arbeit. Prinzipiell unterscheiden wir dabei zwei Problemstellungen:

- Generierung aller Strukturformeln zu einer gegebenen (optional *weichen*) *Bruttoformel*.
- Generierung aller Reaktionsprodukte zu gegebenen *Reaktionen* und *Reaktanden*.

*Restriktionen* können diese Strukturgenerierungsprobleme genauer spezifizieren und verschiedene Strategien zu ihrer Lösung erforderlich machen.

Die Generierung aller Konstitutionsisomere zu einer Bruttoformel wurde in [51] behandelt. Die dort verwendete Strategie der *ordnungstreuen Erzeugung* stößt aber an ihre Grenzen, wenn *inkonsistente* Restriktionen berücksichtigt werden sollen. Für diesen Fall bietet die *zielgerichtete Erzeugung* nach T. Grüner [53] eine geeignetere Lösung, welche auf möglichst effizientem Backtracking des Konstruktionsbaumes und kanonischer Nummerierung der gefundenen Strukturen beruht.

Für die Anlagerung verschiedener *Liganden* an ein *Zentralmolekül* wird in [164] gezeigt, wie sich dieses Problem auf die Erzeugung von Symmetrie- bzw. Doppelnebenklassen überführen und mit ordnungstreuer Erzeugung lösen lässt. Oft trifft man jedoch in der Chemie auf Situationen, wo die sukzessive Anwendung von Reaktionen nicht durch Zentralmolekül-Ligand-Anlagerungen beschrieben werden kann. Ringschlüsse, Umlagerungs- und Zerfallsreaktionen können zu verschiedenartigsten *Reaktionsnetzwerken* führen. Hier wird ein allgemeiner Konstruktionsalgorithmus erarbeitet, welcher das Reaktionsnetzwerk durchläuft und durch kanonische Nummerierung der gefundenen Produkte derartige Strukturgenerierungsprobleme löst.

## 3 Überwachtes statistisches Lernen

In Kapitel 3 werden einige grundlegende Prinzipien des *überwachten statistischen Lernens* erläutert. Statistische Lernverfahren werden in der Computerchemie immer dann eingesetzt, wenn der ursächliche Zusammenhang zwischen Struktur und Eigenschaft nicht bekannt ist oder nur mit extrem hohem Aufwand berechnet werden kann. Auf derartige Problemstellungen trifft man sowohl in der *kombinatorischen Chemie* als auch der *molekularen Strukturaufklärung*.

Beim überwachten Lernen wird anhand bekannter Fälle eine *Vorhersagefunktion* trainiert, die dann zur Prognose für unbekanntere Fälle herangezogen

werden kann. Die bekannten Fälle  $i \in m$  werden dabei repräsentiert durch Werte  $x_{ij}$  der *Vorhersagevariablen*  $X_j$ ,  $j \in n$  und Werte  $y_i$  der *Zielvariablen*  $Y$ . Die  $X_j$  sind für unsere Zwecke reellwertig,  $Y$  kann reellwertig oder diskret sein. Gesucht wird eine Vorhersagefunktion  $f : \mathbb{R}^n \rightarrow Y$ , deren Funktionswerte  $f((x_{ij})_{j \in n})$  für die bekannten Fälle  $i \in m$  gut mit den Werten  $y_i$  der Zielvariablen übereinstimmen.

Im Falle einer reellwertigen Zielvariablen bestimmt man Vorhersagefunktionen durch *Regression*, im diskreten Fall durch *Klassifikation*. Die Bestimmung der Güte von Vorhersagefunktionen kann durch *Resubstitution*, eine *Teststichprobe* oder *Kreuzvalidierung* erfolgen. Es kann sinnvoll sein, die Variablen vor Anwendung des Lernverfahrens vorverarbeitenden Prozeduren wie *Zentrierung*, *Bereichsskalierung* oder *Autoskalierung* zu unterziehen. Zur Vermeidung von Overfitting ist es wichtig, die Anzahl der Vorhersagevariablen zu beschränken. Möglichkeiten zur Variablen-Selektion bieten *Korrelationsanalyse* sowie vollständige oder schrittweise Suche nach geeigneten Teilmengen von Vorhersagevariablen.

Die inferentielle Statistik hat in den letzten Jahrzehnten diverse Verfahren zum Trainieren verschiedener Typen von Vorhersagefunktionen hervorgebracht [64]. Als wichtigste Vertreter werden *lineare Modelle*, *künstliche neuronale Netze*, *Support-Vektor-Maschinen*, *Entscheidungsbäume* sowie die Methode der *k-nächsten Nachbarn* kurz beschrieben.

Sie komplettieren damit den Pool mathematischer Werkzeuge, die wir im anschließenden anwendungsbezogenen Teil der Arbeit zur Modellierung von Problemstellungen der chemischen Synthese und Analytik heranziehen.

#### 4 Kombinatorische Chemie

Bei der Suche nach neuen Wirkstoffen und Materialien finden immer häufiger Techniken der kombinatorischen Chemie Verwendung. Dabei werden aus mehreren Sätzen chemischer *Bausteine* sämtliche Kombinationen synthetisiert und anschließend auf ihre biologisch-pharmazeutische Wirksamkeit getestet oder hinsichtlich angestrebter physiko-chemischer Eigenschaften untersucht. Die innerhalb eines kombinatorisch-chemischen Experiments synthetisierten Verbindungen werden *kombinatorische Bibliothek* genannt, die Testung allgemein als *Screening* bezeichnet.

Der enorme Vorteil dieser Methode besteht darin, dass sowohl die Synthese als auch das Screening in hohem Maße automatisiert und parallelisiert werden können. Obwohl somit erstaunliche Durchsatzraten erzielt werden, mahnen Kosten- und Zeitgründe zur gewissenhaften Planung derartiger Experimente. Die Optimierung kombinatorisch-chemischer Experimente wirft vielfältige, mathematisch anspruchsvolle Teilprobleme auf, die mit Hilfe von diskreten

Konstruktionsalgorithmen (Kapitel 2), graphentheoretischen Invarianten und statistischen Lernverfahren (Kapitel 3) gelöst werden können. Diese allgemeinen Methoden werden in Kapitel 4 auf die Anforderungen der kombinatorischen Chemie konkretisiert. Dazu zählt

- (i) die dublettenfreie, vollständige *Generierung* der durch Reaktanden und Reaktionen definierten virtuellen Bibliothek.

Je nach Situation muss

- (ii) die Bestimmung einer zu synthetisierenden realen Teilbibliothek hoher *Diversität* oder
- (ii') die Überprüfung der *Teilmengenrelation* einer bereits bestehenden realen zu einer generierten virtuellen Bibliothek

erfolgen. Weitere Schritte umfassen

- (iii) die Berechnung *molekularer Deskriptoren*, welche molekulare Strukturen vermöge graphentheoretischer Invarianten auf reelle Zahlen abbilden,
- (iv) die Bestimmung von *Vorhersagefunktionen* durch überwachtetes statistisches Lernen, basierend auf experimentell bestimmten Eigenschaftswerten der realen Bibliothek sowie
- (v) die Anwendung der Vorhersagefunktion auf die virtuelle Bibliothek zur Prognose aussichtsreicher Kandidaten für eine gezielte Synthese.

Schritte (iii) und (iv) werden als Suche nach quantitativen Struktur–Eigenschafts–Beziehungen (QSPR) bezeichnet.

Ein klassisches Beispiel für QSPR–Studien bilden *Siedepunkte* von Alkanen. Erstmals werden in dieser Arbeit Modelle für *Decane* aufgestellt. Der methodische Schwerpunkt liegt hierbei auf dem Vergleich verschiedener Arten molekularer Deskriptoren. 30 topologische Indizes werden mit 20 Substruktur–Vielfachheiten verglichen. Für beide Deskriptorensätze sowie deren Vereinigung werden beste lineare Modelle mit  $n = 1, \dots, 5$  Deskriptoren ermittelt. Entscheidendes praktisches Ergebnis ist, dass schon ab  $n = 3$  Deskriptoren Modelle, die sowohl topologische Indizes als auch Substruktur–Vielfachheiten verwenden, *bessere* Übereinstimmung mit den gemessenen Daten zeigen als solche, die jeweils nur auf einen Deskriptorensatz beschränkt sind. Bemerkenswert ist hierbei, dass ausgehend von einem numerischen Experiment ein allgemein gültiger mathematischer Zusammenhang gefunden wurde: Bei



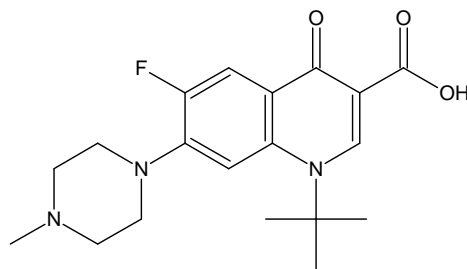
einer Korrelationsanalyse der Deskriptorenwerte wurde die zuvor unbekannte lineare Abhängigkeit von *Zagreb Index*  $M_2$  und *Molecular Walk Count*  $mwc^{(3)}$  für Länge 3

$$mwc^{(3)} = 2M_2$$

entdeckt.

Als zweites Beispiel wird eine QSPR–Untersuchung der *physikalischen Dichte* von *Propylacrylaten* durchgeführt. Hier wird mit disjunktem Lern- und Test-satz gearbeitet. Als Deskriptoren werden neben arithmetischen und topologischen auch geometrische Indizes herangezogen. Die statistische Modellierung wird mit kleinster Quadrate–Regression in Kombination mit schrittweiser Variablen–Selektion vorgenommen. Überraschendes Ergebnis war hierbei, dass die besten Modelle *keine* geometrischen Deskriptoren verwenden. Zudem wird ein Vergleich zur Modellierung mit Hauptkomponenten–Regression und ein Vergleich zur Modell–Bewertung mit Kreuzvalidierung angestellt.

In einem dritten Beispiel wurde eine biologisch–pharmazeutische Eigenschaft untersucht, die *antimycobakterielle Aktivität* von *Quinolonen*. Dieses Problem wurde bereits in früheren Arbeiten [113, 125] behandelt. Nun konnten Modelle gefunden werden, welche aktive und inaktive Strukturen vollständig separieren. Es handelt sich dabei um drei Modelle mit linearer Diskriminanzfunktion unter Verwendung von je drei Deskriptoren. In der rein virtuellen Bibliothek von 39 Verbindungen wird eine bislang noch nicht synthetisierte Struktur von allen drei Modellen übereinstimmend als aktiv prognostiziert:



Sie bildet somit einen aussichtsreichen Kandidaten für einen neuen Wirkstoff gegen *Mycobacterium fortuitum*.

## 5 Molekulare Strukturaufklärung

Die Aufgabe der molekularen Strukturaufklärung besteht darin, anhand von spektroskopischen Daten einer unbekanntem chemischen Verbindung ihre molekulare Struktur zu bestimmen. Für diesen Problemkreis ist neben NMR– und IR–Spektroskopie vor allem die *Massenspektrometrie* (MS) aufgrund ihrer hohen Sensitivität von neu erwecktem Interesse. Insbesondere ist man an einer Automatisierung des Strukturaufklärungsprozesses interessiert.

Nach einem klassischen Ansatz [87] wird die automatisierte Strukturaufklärung in drei Stufen gegliedert, die wiederum mit konstruktiven Methoden der diskreten Mathematik und statistischen Lernverfahren modelliert werden können:

- (i) Als erster Schritt wird durch *Interpretation* versucht, aus den spektroskopischen Daten Informationen über strukturelle Eigenschaften des Analyten zu extrahieren.
- (ii) Im zweiten Schritt werden alle molekularen Graphen bis auf Isomorphie *generiert*, die diesen strukturellen Eigenschaften entsprechen. Man erhält die Strukturkandidaten.
- (iii) Der dritte Schritt umfasst die *Simulation* virtueller Spektren für die Strukturkandidaten, den *Vergleich* mit dem experimentellen Spektrum, das *Ranking* und die *Selektion* relevanter Kandidaten.

Die Strukturaufklärung anhand von MS findet auf mehreren Ebenen statt, die sich vor allem hinsichtlich ihrer Komplexität unterscheiden: Die Bestimmung der Molekülmasse, der Summenformel und der Strukturformel. Mit der besonders verbreiteten, niedrig auflösenden Elektronenstoß-Massenspektrometrie (EI-MS) ist es im Allgemeinen weder möglich, die Molekülmasse noch Summen- oder Strukturformel eindeutig zu bestimmen. So müssen oft mehrere Kandidaten für die Molekülmasse als Eingabe für die Berechnung der Summenformel und ebenso mehrere Summenformel-Kandidaten zur Bestimmung der Strukturformel berücksichtigt werden. Die Bestimmung von Molekülmasse und insbesondere Summenformel wird analog zu obiger Vorgehensweise ebenfalls in die drei Teilprobleme Interpretation, Generierung und Selektion aufgegliedert.

Charakteristisch für die hier betrachtete EI-MS ist, dass nicht die unbekannt Substanz selbst vermessen wird, sondern ein Ionengemisch, das aus ihr durch Elektronenbeschuss hervorgeht. Diese Ionen entstehen durch bekannte Ionisations-, Fragmentierungs- und Umlagerungsreaktionen, die durch Reaktionsschemata dargestellt und für einen gegebenen Strukturkandidaten generiert werden.

Verantwortlich für die Signale im Massenspektrum sind die theoretischen Isotopenmuster der Ionen und die Häufigkeiten, mit denen sie in dem Ionengemisch auftreten. Während theoretische Isotopenmuster über die Bruttoformel eines Fragments exakt berechnet werden können, sind Intensitäten im Massenspektrum nach derzeitigem Kenntnisstand nicht allgemein berechenbar.

Aufgrund der im Ionengemisch auftretenden Bruttoformeln wird ein Vergleichswert für die Übereinstimmung eines Bruttoformel- oder Strukturkandidaten  $K$  mit einem experimentellen Spektrum  $I$  kausal hergeleitet:

$$\text{MV}(I, K) = 1 - \sqrt{\left(\sum_m (I(m))^2\right)^{-1} \min_{\mathbf{x} \geq 0} \sum_m \left(I(m) - \sum_{i=1}^n x_i I_{\beta_i}(m)\right)^2}$$

$I$  ist dabei gegeben durch die Intensität  $I(m)$  bei Masse  $m$ , die theoretischen Isotopenmuster der virtuellen Fragmente mit Bruttoformel  $\beta_i$  werden mit  $I_{\beta_i}$  bezeichnet und  $x_i$  steht für deren unbekannte Intensitäten. Mit Hilfe dieses Vergleichswertes kann ein Ranking von Bruttoformel- sowie Strukturkandidaten vorgenommen werden.

Die Güte von Rankingfunktionen wird in der vorliegenden Arbeit erstmals statistisch evaluiert. Für die vorgestellte Rankingfunktion von Bruttoformeln wurde eine Stichprobe von 100 aufgeklärten Massenspektren aus einer Spektren-Datenbank<sup>5</sup> entnommen. Für jedes Spektrum wurden alle Bruttoformeln zu der bekannten Molekülmasse generiert, Werte der Rankingfunktion bestimmt und die Kandidaten gemäß dieses Wertes abfallend sortiert. Die Güte der Rankingfunktion wird über die *relative Ranking Position*

$$RRP = \frac{\text{Position des korrekten Kandidaten} - 1}{\text{Gesamtzahl der Kandidaten} - 1}$$

bestimmt. Als Mittelwert über die 100 Spektren unserer Stichprobe erhält man dabei  $\overline{RRP} = 0,1037$ . In einem weiteren Schritt wurde die Verwendbarkeit der Rankingfunktion zur *Kandidaten-Selektion* untersucht.

Die Kenntnis über den Verlauf von Fragmentierungs- und Umlagerungsreaktionen im Massenspektrum wird zur Berechnung einer Rankingfunktion für Strukturformeln herangezogen. Dabei werden für einen Strukturkandidaten zunächst die möglichen Fragmente generiert. Die theoretischen Isotopenmuster der virtuellen Fragmente werden zur Erklärung des experimentellen Spektrums verwendet. Wie zuvor für Bruttoformeln werden auch für Strukturformeln die Güte der Rankingfunktion für 100 Datenbank-Spektren<sup>6</sup> ermittelt. Dazu werden alle Konstitutionsisomere zur korrekten Bruttoformel als mögliche Kandidaten herangezogen.

Für die Extraktion struktureller Eigenschaften aus MS werden MS-Klassifikatoren verwendet. Vorhersagevariablen für MS-Klassifikatoren bilden MS-Deskriptoren. Die Zielvariable ist ein binärer molekularer Deskriptor zu einer

<sup>5</sup>Hierbei wurden nur Spektren berücksichtigt, deren zugehörige Molekülmasse nicht größer als 200 amu ist und zu der mindestens 2 Bruttoformeln existieren.

<sup>6</sup>Hierbei wurden nur Spektren berücksichtigt, deren zugehörige Molekülmasse nicht größer als 200 amu ist und für deren Bruttoformel mindestens 2 und höchstens 10000 Strukturformeln existieren

strukturellen Eigenschaft  $S$ , der besagt, ob die zum Spektrum gehörige chemische Verbindung Eigenschaft  $S$  besitzt. Für 77 strukturelle Eigenschaften werden verschiedene Klassifikationsverfahren (CART, LDA, ANN, SVM) auf ihre Vorhersagefähigkeit getestet. Die kleinste durchschnittliche Missklassifikationsrate (0,23160) liefern dabei SVM mit radialem Kernel.

Als eine Möglichkeit zur Weiterentwicklung von MS-Klassifikatoren wird die systematische Suche nach neuen strukturellen Eigenschaften für MS-Klassifikation vorgestellt. Anhand zweier Beispiele wird die Kopplung der drei Schritte Interpretation, Konstruktion und Verifikation demonstriert. Eine weitere Studie zeigt, wie MS direkt mit strukturellen Eigenschaften in Beziehung gebracht werden und beschreibt Wege, wie diese Erkenntnisse für die molekulare Strukturaufklärung genutzt werden können.

Der letzte Abschnitt in Kapitel 5 gibt einen Ausblick auf hoch auflösende Massenspektrometrie. HR-MS kann insbesondere dazu verwendet werden, um Kandidatenmengen für die Bruttoformel einzuschränken. Dazu werden die exakten Atommassen der chemischen Elemente herangezogen. Eine oft gestellte Frage des chemischen Analytikers lautet: Wie genau muss die Molekülmasse gemessen werden können, um eine eindeutige Bestimmung der Bruttoformel zu gewährleisten. Diese Frage wird mit einer mathematischen Simulation beantwortet.

**Teil I**

**Mathematische Grundlagen**



# Kapitel 1

## Diskrete Strukturen in der Chemie

Dieses erste Kapitel umfasst einige grundlegende Definitionen und Sätze über endliche Gruppenoperationen und diskrete Strukturen [72] sowie elementare Begriffe aus der Graphentheorie [63]. Es bildet somit das Fundament für die nachfolgenden Betrachtungen. Wir werden molekulare Graphen zur Darstellung chemischer Verbindungen einführen. Molekulare Substrukturen sollen dazu dienen, strukturelle Eigenschaften molekularer Graphen zu beschreiben. Chemische Reaktionen werden wir in diesem Zusammenhang durch Reaktionsschemata graphisch modellieren. Schließlich wollen wir einige Erweiterungen des Molekülmodells diskutieren.

### 1.1 Gruppenoperationen

Diskrete Strukturen liegen zunächst in nummerierter Form vor. Dies ist insbesondere notwendig, um sie im Speicher eines Rechners darstellen zu können. Gruppenoperationen bilden das mathematische Werkzeug, um den für unsere Anwendungen entscheidenden Übergang von nummerierten zu nicht-nummerierten diskreten Strukturen zu realisieren.

#### 1.1.1 Definition:

Sei  $G$  eine Gruppe und  $\Omega$  eine nichtleere Menge. Eine Abbildung

$$\varphi : \Omega \times G \longrightarrow \Omega, \quad (\omega, g) \longmapsto \varphi(\omega, g) =: \omega^g$$

heißt *Operation* von  $G$  auf  $\Omega$ , falls  $\varphi$  mit der Verknüpfung in  $G$  verträglich ist und *id* die Elemente von  $\Omega$  fest lässt. Es müssen also folgende Bedingungen erfüllt sein:

- (i)  $\forall \omega \in \Omega \forall g, g' \in G : \omega^{gg'} = (\omega^g)^{g'}$ .
- (ii)  $\forall \omega \in \Omega : \omega^{id} = \omega$ .

Operiert  $G$  auf  $\Omega$ , so wird diese Operation mit  $\Omega_G$  bezeichnet. Sind sowohl  $G$  als auch  $\Omega$  endlich, so spricht man von einer *endlichen* Operation. Im Folgenden werden ausschließlich endliche Operationen betrachtet.

### 1.1.2 Schreibweise:

Sei  $n \in \mathbb{N} \setminus \{0\} =: \mathbb{N}^*$  eine natürliche Zahl ungleich 0. Falls es aus dem Zusammenhang eindeutig hervorgeht, bezeichnen wir mit  $n$  auch die Menge  $\{0, \dots, n-1\}$ . Eine häufig verwendete Gruppe ist die symmetrische Gruppe  $S_n$  der Bijektionen von  $n$  nach  $n$  mit Komposition als Verknüpfung. Ein sehr einfaches Beispiel einer Gruppenoperation liefert die Operation der symmetrischen Gruppe  $S_n$  auf  $n$  mit  $x^\pi := \pi^{-1}(x)$ .

### 1.1.3 Definition:

Sei  $\Omega_G$  eine Operation und  $\omega \in \Omega$ . Dann ist

- $\omega^G := \{\omega^g \in \Omega \mid g \in G\}$  die *Bahn* von  $\omega$  unter  $\Omega_G$  und
- $\Omega//G := \{\omega^G \mid \omega \in \Omega\}$  die Menge der Bahnen von  $\Omega_G$ .

Eine Menge  $T \subseteq \Omega$  heißt *Transversale* der Bahnen von  $\Omega_G$ , falls

$$\forall B \in \Omega//G : |B \cap T| = 1$$

Eine Transversale  $T$  enthält also aus jeder Bahn genau ein Element. Die Menge der Transversalen von  $\Omega_G$  wird mit  $\mathcal{T}(\Omega//G)$  bezeichnet. Bekanntlich gilt:

### 1.1.4 Satz:

Sei  $\Omega_G$  eine Operation. Dann gilt:

- Je zwei Bahnen von  $\Omega_G$  sind entweder gleich oder disjunkt:

$$\forall \omega, \omega' \in \Omega : \omega^G \cap \omega'^G \neq \emptyset \iff \omega' \in \omega^G.$$

- $\Omega$  ist die disjunkte Vereinigung der Bahnen von  $\Omega_G$ :

$$\forall T \in \mathcal{T}(\Omega//G) : \Omega = \bigcup_{\omega \in T} \omega^G.$$

- *Klassengleichung*:

$$\forall T \in \mathcal{T}(\Omega//G) : |\Omega| = \sum_{\omega \in T} |\omega^G|.$$



**1.1.5 Beispiele und Definitionen:**

Sei  $G$  eine Gruppe und  $A \leq G$  eine Untergruppe. Dann lassen sich wie folgt Gruppenoperationen auf  $G$  erklären:

- Die Abbildung

$$G \times A \longrightarrow G, \quad (g, a) \longmapsto g^a := ga$$

induziert eine Operation von  $A$  auf  $G$ . Für  $g \in G$  heißt die Bahn

$$gA := g^A = \{g^a \mid a \in A\} = \{ga \mid a \in A\}$$

*Linksnebenklasse* von  $A$  in  $G$ . Die Menge aller Linksnebenklassen von  $A$  in  $G$  wird mit  $G/A$  bezeichnet. Nach Satz 1.1.4 sind je zwei Linksnebenklassen gleich oder disjunkt und  $G$  ist die disjunkte Vereinigung der Linksnebenklassen von  $A$  in  $G$ .

- Ebenso definiert

$$G \times A \longrightarrow G, \quad (g, a) \longmapsto g^a := a^{-1}g$$

eine Operation von  $A$  auf  $G$ . Dabei heißt für  $g \in G$  die Bahn

$$Ag := g^A = \{g^a \mid a \in A\} = \{ag \mid a \in A\}$$

*Rechtsnebenklasse* von  $A$  in  $G$ . Die Menge der Rechtsnebenklassen wird mit  $A \backslash G$  bezeichnet.

- Sei  $B \leq G$  eine weitere Untergruppe. Dann erklärt

$$G \times (A \times B) \longrightarrow G, \quad (g, (a, b)) \longmapsto g^{(a,b)} := a^{-1}gb$$

eine Operation von  $A \times B$  auf  $G$ . Die Bahn

$$AgB := g^{A \times B} = \{g^{(a,b)} \mid (a, b) \in A \times B\} = \{agb \mid a \in A, b \in B\}$$

heißt *Doppelnebenklasse*. Man schreibt  $A \backslash G / B$  für die Menge der Doppelnebenklassen von  $A$  und  $B$  in  $G$ .

**1.1.6 Definition:**

Sei  $\Omega_G$  eine Operation und  $T \in \mathcal{T}(\Omega // G)$  eine Transversale der Bahnen dieser Operation. Eine Abbildung  $\rho : \Omega \longrightarrow G$  heißt *fusionierende Abbildung* zu  $T$ , falls

$$\forall \omega \in \Omega : \quad \omega^{\rho(\omega)} \in T.$$

Das *fusionierende Element*  $\rho(\omega)$  bildet somit  $\omega$  auf den zugehörigen Bahnrepräsentanten ab.

**1.1.7 Bemerkung:**

Eine fusionierende Abbildung  $\rho$  wird insbesondere dann wichtig, wenn zwei Bahnen  $\omega^G$  und  $\omega'^G$  verglichen werden sollen. Anstatt zu testen, ob  $\omega'$  in  $\omega^G$  liegt, braucht man nur die beiden Elemente  $\omega^{\rho(\omega)}$  und  $\omega'^{\rho(\omega')}$  zu bestimmen. Dann ist

$$\omega^G = \omega'^G \iff \omega^{\rho(\omega)} = \omega'^{\rho(\omega')}.$$

**1.1.8 Definition:**

Seien  $X$  und  $Y$  zwei nichtleere endliche Mengen und

$$Y^X := \{\gamma : X \longrightarrow Y\}$$

die Menge der Abbildungen von  $X$  nach  $Y$ . Eine (*nummerierte*) *diskrete Struktur* ist ein Tripel  $(X, Y, \gamma)$ , wobei  $\gamma \in Y^X$ .

**1.1.9 Bemerkung:**

Sei  $X_G$  eine Operation und  $Y$  eine nichtleere endliche Menge. Dann lässt sich auf folgende Weise eine Operation von  $G$  auf  $Y^X$  erklären:

$$Y^X \times G \longrightarrow Y^X, \quad (\gamma, g) \longmapsto \gamma^g,$$

wobei für  $x \in X$

$$(\gamma^g)(x) := \gamma(x^{g^{-1}}).$$

Die Bahnen  $B \in Y^X // G$  heißen auch *Isomorphieklassen von Abbildungen*  $Y^X$ . Von entscheidender Bedeutung für uns ist, dass auf diese Weise der Übergang von nummerierten zu nicht-nummerierten diskreten Strukturen erreicht wird. In den nächsten Abschnitten findet diese Technik Anwendung auf Graphen, Multigraphen und insbesondere molekulare Graphen.

**1.1.10 Definition:**

Mit den Bezeichnungen aus Bemerkung 1.1.9 ist

$$\text{Aut}(\gamma) := \{g \in G \mid \gamma^g = \gamma\}$$

eine Untergruppe von  $G$  und wird *Automorphismengruppe* von  $\gamma$  genannt.

**1.1.11 Definition:**

Sei  $\Omega_G$  eine Operation und  $W$  eine nichtleere Menge. Dann heißt eine Abbildung  $f : \Omega \rightarrow W$  *Invariante* von  $\Omega$  bezüglich  $\Omega_G$ , wenn sie auf den Bahnen von  $\Omega_G$  konstant ist.

Invarianten diskreter Strukturen, insbesondere Grapheninvarianten werden im Rahmen dieser Arbeit eine wichtige Rolle spielen.

## 1.2 Graphen und Multigraphen

### 1.2.1 Definition:

Sei  $m \in \mathbb{N}^*$ ,  $m \geq 2$  und  $V$  eine nichtleere Menge und  $\binom{V}{2}$  die Menge der 2-Teilmengen von  $V$ . Dann heißen die Elemente von

$$\mathcal{G}_{V,m} := m^{\binom{V}{2}}$$

(nummerierte) Graphen. Für  $m = 2$  spricht man von *schlichten* Graphen, anderenfalls von *Multigraphen*.

### 1.2.2 Bemerkung:

Nach 1.1.9 erklärt folgende Abbildung eine Operation von  $S_V$  auf  $\mathcal{G}_{V,m}$ :

$$\mathcal{G}_{V,m} \times S_V \longrightarrow \mathcal{G}_{V,m}, \quad (\gamma, \pi) \longmapsto \gamma^\pi,$$

wobei für  $v, w \in V$

$$\gamma^\pi(\{v, w\}) := \gamma(\{\pi(v), \pi(w)\}).$$

Zwei Graphen  $\gamma, \gamma' \in \mathcal{G}_{n,m}$  sind zueinander *isomorph*, wenn sie in der gleichen Bahn unter dieser Operation liegen. Die Elemente von  $\mathcal{G}_{n,m} // S_n$  heißen *Isomorphieklassen von Graphen* oder auch *nicht-nummerierte Graphen* und werden im Folgenden mit  $\bar{\gamma} := \gamma^{S_n}$  bezeichnet.

### 1.2.3 Definition:

Sei  $T \in \mathcal{T}(\mathcal{G}_{n,m} // S_n)$  eine Transversale und  $\rho : \mathcal{G}_{n,m} \longrightarrow S_n$  eine fusionierende Abbildung zu  $T$ . Dann ist die Abbildung

$$\kappa : \mathcal{G}_{n,m} \longrightarrow T, \quad \gamma \longmapsto \gamma^{\rho(\gamma)}$$

eine *kanonische Form* auf  $\mathcal{G}_{n,m}$  und zu  $\gamma \in \mathcal{G}_{n,m}$  heißt  $\rho(\gamma) \in S_n$  *kanonische Nummerierung* von  $\gamma$ .

### 1.2.4 Beispiel:

Die durch  $\gamma$  definierte Matrix

$$A_\gamma = (a_{ij}) \in m^{n \times n} \quad \text{mit} \quad a_{ij} := \begin{cases} \gamma(\{i, j\}) & \text{falls } i \neq j, \\ 0 & \text{sonst} \end{cases}$$

heißt *Adjazenzmatrix* von  $\gamma$ . Eine Möglichkeit zur Bestimmung einer Transversale von  $\mathcal{G}_{n,m} // S_n$  besteht darin, als Bahnrepräsentanten denjenigen Graphen auszuwählen, dessen Adjazenzmatrix, zeilenweise betrachtet, lexikographisch minimal ist (vgl. Abschnitt 2.1.1).

**1.2.5 Definition:**

Sei  $\gamma \in \mathcal{G}_{V,m}$  ein Graph. Dann ist  $V$  die Menge der *Knoten* und

$$E_\gamma := \left\{ e \in \binom{V}{2} \mid \gamma(e) > 0 \right\}$$

die Menge der *Kanten* von  $\gamma$ . Für einer Kante  $e \in E_\gamma$  ist  $\gamma(e)$  die *Kanten-  
vielfachheit*. Der *Grad* eines Knoten  $v \in V$  ist

$$\deg_\gamma(v) := \sum_{w \in V \setminus \{v\}} \gamma(\{v, w\}).$$

Ist  $e = \{v, w\} \in E_\gamma$ , dann sind  $v$  und  $w$  die Endpunkte von  $e$ . Man sagt, dass  $v$  und  $w$  mit  $e$  *inzident* und dass  $v$  und  $w$  *adjazent* sind.

**1.2.6 Bemerkung:**

Für schlichte Graphen  $\gamma \in \mathcal{G}_{V,2}$  gibt es folgenden einfachen Zusammenhang zwischen den Knotengraden und der Mächtigkeit der Kantenmenge:

$$|E_\gamma| = \frac{1}{2} \sum_{v \in V} \deg_\gamma(v).$$

Allgemein gilt für  $\gamma \in \mathcal{G}_{V,m}$

$$\sum_{e \in E_\gamma} \gamma(e) = \frac{1}{2} \sum_{v \in V} \deg_\gamma(v).$$

Sei  $k := \max\{\deg_\gamma(v) \mid v \in V\}$  der maximale Knotengrad von  $\gamma$  und  $\lambda_\gamma(i) := |\{v \in V \mid \deg_\gamma(v) = i\}|$  für  $i = 0, \dots, k$  die Anzahl der Knoten vom Grad  $i$ . Dann ist die *Gradpartition* von  $\gamma$  definiert als

$$\lambda_\gamma := (\lambda_\gamma(0), \dots, \lambda_\gamma(k)) \models |V|.$$

Die Gradpartition eines Graphen ist invariant unter der Operation von  $S_V$  auf  $\gamma \in \mathcal{G}_{V,m}$ .

**1.2.7 Satz:**

Sei  $k \in \mathbb{N}$ ,  $n \in \mathbb{N}^*$  und  $\lambda = (\lambda(0), \dots, \lambda(k)) \models n$  eine *Partition* von  $n$ , d.h.  $\sum_{i=0}^k \lambda(i) = n$ .  $\lambda$  ist genau dann Gradpartition mindestens eines Graphen  $\gamma$ , wenn

$$\exists q \geq k : \sum_i i\lambda(i) = 2q.$$

**1.2.8 Bemerkung:**

$q$  entspricht dabei der Anzahl der Kanten von  $\gamma$ . Einen Beweis für obigen Satz findet man in [51]. Partitionen, die der Bedingung aus 1.2.7 genügen, heißen *multigraphisch*. Eine stärkere Aussage über die Existenz eines solchen zusammenhängenden Graphen liefert Satz 1.2.15.

**1.2.9 Definition:**

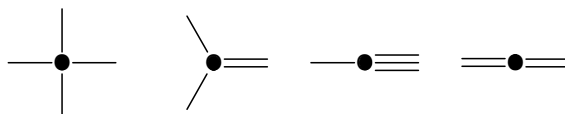
Sei  $\gamma \in \mathcal{G}_{V,m}$ ,  $v \in V$  mit  $\deg_\gamma(v) > 0$  und  $\mu_i = |\{w \in V \mid \gamma(\{v, w\}) = i\}|$  für  $i = 1, \dots, m-1$ . Die Verteilung der Vielfachheiten der zu  $v$  inzidenten Kanten

$$\text{hyb}_\gamma(v) := (\mu_1, \dots, \mu_{m-1})$$

bezeichnen wir als *Hybridisierung*<sup>1</sup> von  $v$ .

**1.2.10 Beispiel:**

Sei  $m = 4$ . Dann sind für einen Knoten vom Grad 4 folgende Hybridisierungen möglich:

**1.2.11 Definition:**

Sei  $\gamma \in \mathcal{G}_{V,m}$  ein Graph,  $k \in \mathbb{N}^*$  und  $v_0, \dots, v_k \in V$ .

- $(v_0, \dots, v_k)$  ist ein *Kantenzug* zwischen  $v_0$  und  $v_k$ , wenn

$$\forall i \in k : \{v_i, v_{i+1}\} \in E_\gamma.$$

$v_0$  und  $v_k$  sind die Endpunkte des Kantenzugs.

- Ein Kantenzug  $(v_0, \dots, v_k)$  heißt *offen*, falls  $v_0 \neq v_k$ , anderenfalls *geschlossen*.
- Ein offener Kantenzug  $(v_0, \dots, v_k)$  heißt *Weg*, wenn seine Knoten paarweise verschieden sind.

<sup>1</sup>Diese rein mathematische Definition weicht ab von dem chemischen Begriff der Hybridisierung. So kann ein sp-hybridisiertes Kohlenstoff-Atom (Grad 4) die mathematischen Hybridisierungen (1,0,1) und (0,2,0) haben. Um mit der Nomenklatur früherer Arbeiten konform zu bleiben wurde von einer neuen Begriffsbildung abgesehen.

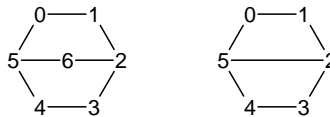
- Ein geschlossener Kantenzug  $(v_0, \dots, v_k)$  heißt *Kreis*, wenn seine Knoten bis auf die Endpunkte paarweise verschieden sind.
- Ein Kreis  $(v_0, \dots, v_k)$  heißt *Ring*, wenn

$$|\{\{v_i, v_j\} \in E_\gamma \mid i, j \in k\}| = k.$$

Enthält ein Graph keinen Kreis, so spricht man von einem *Baum*.

### 1.2.12 Beispiel:

In den folgenden beiden Graphen ist  $W = (v_0, \dots, v_6) = (0, 1, 2, 3, 4, 5, 0)$  jeweils ein Kreis:



Im linken Graphen ist  $W$  auch Ring, im rechten nicht.

### 1.2.13 Definition:

Sei  $\gamma \in \mathcal{G}_{V,m}$  ein Graph.

- Die Anzahl der Kanten eines Kantenzugs  $W = (v_0, \dots, v_k)$  wird als seine *Länge*  $\text{len}_\gamma(W) := k$  bezeichnet.
- Zwei Knoten  $v, w \in V$  heißen *verbindbar*, wenn in  $\gamma$  ein Kantenzug zwischen  $v$  und  $w$  existiert.
- Sind je zwei Knoten von  $\gamma$  verbindbar, dann heißt  $\gamma$  *zusammenhängend*. Die Menge der zusammenhängenden Graphen wird bezeichnet mit

$$\mathcal{G}_{V,m}^C := \{\gamma \in \mathcal{G}_{V,m} \mid \gamma \text{ zusammenhängend}\}$$

- Der *Abstand* zweier verbindbarer Knoten  $v, w \in V$  ist

$$\text{dist}_\gamma(v, w) := \min\{\text{len}_\gamma(W) \mid W \text{ Kantenzug zwischen } v \text{ und } w\}.$$

Sind  $v$  und  $w$  nicht verbindbar, definiert man  $\text{dist}_\gamma(v, w) := \infty$ .

**1.2.14 Bemerkung:**

Die durch Verbindbarkeit erklärte Relation auf der Knotenmenge  $V$  eines Graphen  $\gamma$  ist eine Äquivalenzrelation. Die Äquivalenzklassen dieser Relation heißen *Zusammenhangskomponenten*.  $\gamma$  ist also genau dann zusammenhängend, wenn nur eine Zusammenhangskomponente existiert. Einen Knoten vom Grad 0 nennt man *triviale* Zusammenhangskomponente. Für zusammenhängende Graphen zu gegebener Gradpartition gilt folgendes Existenzkriterium:

**1.2.15 Satz:**

Sei  $k \in \mathbb{N}$ ,  $n \in \mathbb{N}^*$  und  $\lambda = (\lambda(0), \dots, \lambda(k)) \models n$  eine multigraphische Partition mit  $\sum_i i\lambda(i) = 2q$ .  $\lambda$  ist genau dann Gradpartition mindestens eines zusammenhängenden Graphen  $\gamma$ , wenn

$$q \geq n - 1.$$

*Beweis:* Siehe [51], 2.1.7. □

**1.2.16 Definition:**

Sei  $\gamma \in \mathcal{G}_{V,m}$  und  $K$  ein Kreis in  $\gamma$ .  $K$  heißt *Taillenkreis*, wenn es in  $\gamma$  keinen Kreis kleinerer Länge gibt. Die Länge eines Taillenkreises bezeichnet man als *Taillenweite* von  $\gamma$ :

$$\text{girth}_\gamma := \min\{\text{len}_\gamma(K) \mid K \text{ Kreis in } \gamma\}.$$

Besitzt  $\gamma$  mindestens einen Kreis, so ist  $\text{girth}_\gamma$  als Minimum einer endlichen Menge wohldefiniert. Anderenfalls setzt man  $\text{girth}_\gamma := \infty$ .

**1.2.17 Bemerkung:**

Die Anzahl der Zusammenhangskomponenten und die Taillenweite sind invariant unter der Operation von  $S_V$  auf  $\gamma \in \mathcal{G}_{V,m}$ . Deshalb sind diese Begriffe auch für Isomorphieklassen von Graphen verwendbar.

**1.2.18 Definition:**

Sei  $\gamma \in \mathcal{G}_{V,m}$ ,  $V' \subseteq V$  eine nichtleere Teilmenge von  $V$  und  $\gamma' \in \mathcal{G}_{V',m}$ .  $\gamma'$  ist *Subgraph* von  $\gamma$ , wenn

$$\forall e \in E_{\gamma'} : \gamma'(e) \leq \gamma(e).$$

Wir schreiben  $\gamma' \subseteq \gamma$ . Ist die stärkere Bedingung

$$\forall e \in E_{\gamma'} : \gamma'(e) = \gamma(e)$$

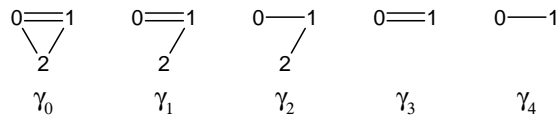
erfüllt, dann bezeichnen wir  $\gamma'$  als *geschlossenen Subgraph* von  $\gamma$  ( $\gamma' \subseteq^c \gamma$ ). Gilt schließlich

$$\forall e \in \binom{V'}{2} : \gamma'(e) = \gamma(e),$$

dann heißt  $\gamma'$  *induzierter Subgraph* oder *Teilgraph* von  $\gamma$  und wir schreiben  $\gamma' \subseteq^i \gamma$ .

**1.2.19 Beispiel:**

Wir wollen die folgenden Graphen nach Subgraph-Beziehungen zu  $\gamma_0$  untersuchen:



Es ist  $\gamma_1 \subseteq^c \gamma_0$ , aber nicht induzierter Subgraph;  $\gamma_2 \subseteq \gamma_0$ , aber nicht geschlossener Subgraph;  $\gamma_3 \subseteq^i \gamma_0$ ;  $\gamma_4 \subseteq \gamma_0$ , aber nicht geschlossener Subgraph.

**1.2.20 Bemerkung:**

Zu gegebenem  $\gamma \in \mathcal{G}_{V,m}$  und  $\emptyset \neq V' \subseteq V$  ist der Teilgraph  $\gamma' \in \mathcal{G}_{V',m}$  von  $\gamma$  eindeutig bestimmt. Man nennt  $\gamma'$  deshalb auch den auf  $V'$  *induzierten* Teilgraph von  $\gamma$  und schreibt  $\gamma' = \gamma|_{V'}$ . Die Menge der auf den Zusammenhangskomponenten  $V_1, \dots, V_k$  von  $\gamma$  induzierten Teilgraphen  $\gamma|_{V_1}, \dots, \gamma|_{V_k}$  von  $\gamma$  wird im Folgenden mit  $\text{Con}(\gamma)$  bezeichnet. Sie sind zusammenhängende Graphen und die Vereinigung ihrer Kantenmengen ist gleich der Kantenmenge von  $\gamma$ . Als nächstes werden die Sub- und Teilgraph-Relationen für Graphen auf Isomorphieklassen von Graphen übertragen.



**1.2.21 Definition:**

Sei  $\gamma \in \mathcal{G}_{V,m}$ ,  $V' \subseteq V$  eine nichtleere Teilmenge von  $V$  und  $\gamma' \in \mathcal{G}_{V',m}$ . Eine injektive Abbildung  $\phi \in V_{\text{inj}}^{V'}$  heißt *Einbettung* von  $\gamma'$  in  $\gamma$

- als Subgraph, wenn

$$\forall \{i, j\} \in E_{\gamma'} : \quad \gamma'(\{i, j\}) \leq \gamma(\{\phi(i), \phi(j)\}),$$

- als geschlossener Subgraph, wenn

$$\forall \{i, j\} \in E_{\gamma'} : \quad \gamma'(\{i, j\}) = \gamma(\{\phi(i), \phi(j)\}),$$

- als induzierter Subgraph oder Teilgraph, wenn

$$\forall \{i, j\} \in \binom{V'}{2} : \quad \gamma'(\{i, j\}) = \gamma(\{\phi(i), \phi(j)\}).$$

Wir schreiben  $\gamma' \subseteq_{\phi} \gamma$  bzw.  $\gamma' \subseteq_{\phi}^c \gamma$ ,  $\gamma' \subseteq_{\phi}^i \gamma$ . Mit

$$\begin{aligned} \text{Emb}_{\subseteq}(\gamma', \gamma) &:= \{\phi \in V_{\text{inj}}^{V'} \mid \gamma' \subseteq_{\phi} \gamma\}, \\ \text{Emb}_{\subseteq^c}(\gamma', \gamma) &:= \{\phi \in V_{\text{inj}}^{V'} \mid \gamma' \subseteq_{\phi}^c \gamma\}, \\ \text{Emb}_{\subseteq^i}(\gamma', \gamma) &:= \{\phi \in V_{\text{inj}}^{V'} \mid \gamma' \subseteq_{\phi}^i \gamma\} \end{aligned}$$

bezeichnen wir die Mengen von Einbettungen.

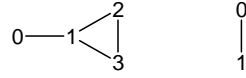
**1.2.22 Bemerkung:**

Mit Hilfe von Einbettungen können Sub- und Teilgraph-Relationen für Isomorphieklassen von Graphen erklärt werden. Für  $\bar{\gamma} \in \mathcal{G}_{V,m} // S_V$ ,  $V' \subseteq V$  und  $\bar{\gamma}' \in \mathcal{G}_{V',m} // S_{V'}$  ist

$$\begin{aligned} \bar{\gamma}' \subseteq \bar{\gamma} &: \iff \exists \phi \in V_{\text{inj}}^{V'} : \gamma' \subseteq_{\phi} \gamma, \\ \bar{\gamma}' \subseteq^c \bar{\gamma} &: \iff \exists \phi \in V_{\text{inj}}^{V'} : \gamma' \subseteq_{\phi}^c \gamma, \\ \bar{\gamma}' \subseteq^i \bar{\gamma} &: \iff \exists \phi \in V_{\text{inj}}^{V'} : \gamma' \subseteq_{\phi}^i \gamma. \end{aligned}$$

**1.2.23 Beispiel:**

Wir betrachten die beiden schlichten Graphen  $\gamma \in \mathcal{G}_{4,2}$  und  $\gamma' \in \mathcal{G}_{2,2}$ :



Dann gibt es 8 Einbettungen von  $\gamma'$  in  $\gamma$ :

$$\text{Emb}_{\subseteq}(\gamma', \gamma) = \{\phi_i \mid i \in 8\}$$

wobei die  $\phi_i$  wie folgt definiert sind:

$i$	0	1	2	3	4	5	6	7
$\phi_i(0)$	0	1	1	2	1	3	2	3
$\phi_i(1)$	1	0	2	1	3	1	3	2

In manchen Situationen ist man daran interessiert, wie oft ein Subgraph in einem Graphen auftritt. Möchte man eine *Vielfachheit* von  $\gamma'$  in  $\gamma$  angeben, muss man sich zunächst darüber klar werden, in wie weit man die Nummerierung der Graphen berücksichtigen will. Es werden vier Möglichkeiten unterschieden:

- i) Berücksichtigung der Nummerierung von  $\gamma'$  und  $\gamma$ : Die Vielfachheit von  $\gamma'$  in  $\gamma$  ist dann  $|\text{Emb}_{\subseteq}(\gamma', \gamma)| = 8$ .
- ii) Vernachlässigung der Nummerierung von  $\gamma'$ : Dazu betrachtet man folgende Operation:

$$\text{Emb}(\gamma', \gamma) \times \text{Aut}(\gamma') \longrightarrow \text{Emb}(\gamma', \gamma), \quad (\phi, \pi) \longmapsto \phi \circ \pi^{-1}.$$

Unter dieser Operation zerfällt  $\text{Emb}(\gamma', \gamma)$  in 4 Bahnen

$$\text{Emb}(\gamma', \gamma) // \text{Aut}(\gamma') = \{\{\phi_0, \phi_1\}, \{\phi_2, \phi_3\}, \{\phi_4, \phi_5\}, \{\phi_6, \phi_7\}\}$$

und die Vielfachheit von  $\gamma'$  in  $\gamma$  unter Berücksichtigung der Symmetrie von  $\gamma'$  ist 4.

- iii) Vernachlässigung der Nummerierung von  $\gamma$ :

$$\text{Emb}(\gamma', \gamma) \times \text{Aut}(\gamma) \longrightarrow \text{Emb}(\gamma', \gamma), \quad (\phi, \tau) \longmapsto \tau \circ \phi.$$

Unter dieser Operation zerfällt  $\text{Emb}(\gamma', \gamma)$  in 5 Bahnen

$$\text{Emb}(\gamma', \gamma) // \text{Aut}(\gamma) = \{\{\phi_0\}, \{\phi_1\}, \{\phi_2, \phi_4\}, \{\phi_3, \phi_5\}, \{\phi_6, \phi_7\}\}$$

und die Vielfachheit von  $\gamma'$  in  $\gamma$  unter Berücksichtigung der Symmetrie von  $\gamma$  ist 5.

iv) Vernachlässigung der Nummerierung von  $\gamma'$  und  $\gamma$ :

$$\begin{aligned} \text{Emb}(\gamma', \gamma) \times (\text{Aut}(\gamma') \times \text{Aut}(\gamma)) &\longrightarrow \text{Emb}(\gamma', \gamma), \\ (\phi, (\pi, \tau)) &\longmapsto \tau^{-1} \circ \phi \circ \pi. \end{aligned}$$

Man erhält 3 Bahnen

$$\text{Emb}(\gamma', \gamma) // \text{Aut}(\gamma') \times \text{Aut}(\gamma) = \{\{\phi_0, \phi_1\}, \{\phi_2, \phi_3, \phi_4, \phi_5\}, \{\phi_6, \phi_7\}\}$$

und die Vielfachheit von  $\gamma'$  in  $\gamma$  unter Berücksichtigung der Symmetrien beider Graphen ist 3.

Vielfachheiten von Subgraphen werden für unsere Anwendungen in der Chemie des Öfteren eine Rolle spielen. Falls nicht explizit angegeben werden wir Variante ii) verwenden.

## 1.3 Molekulare Graphen

Chemische Verbindungen werden in dieser Arbeit als Multigraphen mit speziellen Knoten dargestellt. *Kovalente Bindungen* entsprechen den Kanten des Multigraphen. Doppel- und Dreifachbindungen werden anhand der Kantenvielfachheiten modelliert. Die Knoten sind *Atome*. Atome besitzen als Attribute ein *chemisches Element* und einen *Atomzustand*.

### 1.3.1 Vorbemerkung:

Die chemischen Elemente werden eindeutig identifiziert durch ihre *Ordnungszahl*. Diese entspricht der Anzahl positiv geladener Elementarteilchen, so genannten *Protonen* im Atomkern. Atome besitzen im elementaren Zustand genauso viele *Elektronen* wie Protonen. Elektronen sind negativ geladene Elementarteilchen, Atome sind also zunächst ladungsmäßig neutral.

Manche Elektronen ist in der Lage, Wechselwirkungen mit anderen Atomen einzugehen. Elektronen, die diese Eigenschaft besitzen, heißen *Valenzelektronen*. Die Anzahl der Valenzelektronen ist abhängig vom chemischen Element des Atoms. Wechselwirkungen mit anderen Atomen manifestieren sich in Form chemischer Bindungen. Bei einer kovalenten Bindungen teilen sich zwei Atome zwei, vier oder sechs Elektronen. Solche Bindungen bezeichnet man gemäß ihrer Vielfachheit als Einfach-, Doppel- oder Dreifachbindungen. Andere Arten von Bindungen werden wir in Abschnitt 1.6.1 kennen lernen. Valenzelektronen, die nicht an Bindungen beteiligt sind, schließen sich paarweise zu *freien Elektronenpaaren* zusammen. Verbleibende einzelne Valenzelektronen werden als *ungepaarte Elektronen* bezeichnet. Summiert man für ein Atom einer chemischen Verbindung anteilig kovalenten Bindungen zugeordnete Elektronen, Elektronen in freien Elektronenpaaren und ungepaarte Elektronen, und vergleicht das Ergebnis mit der für das Element charakteristischen Anzahl von Valenzelektronen, so kann mitunter eine Differenz resultieren. Diese bezeichnet man als *Ladung* des Atoms.

### 1.3.2 Definition:

Ein Atomzustand ist ein 4-Tupel  $Z = (v_Z, p_Z, q_Z, r_Z)$ , wobei

- $v_Z \in \mathbb{N}$  die *Valenz* des Atoms bezeichnet,
- $p_Z \in \mathbb{N}$  die Anzahl *freier Elektronenpaare* am Atom angibt,
- $q_Z \in \mathbb{Z}$  die *Ladung* am Atom kodiert, und
- $r_Z \in \mathbb{B} := \{0, 1\} := \{\text{falsch}, \text{wahr}\}$  das Vorhandensein eines *ungepaarten Elektrons* am Atom beschreibt (*Radikalstelle*).

Einen Atomzustand nennen wir *Grundzustand*, wenn  $q_Z = 0$  und  $r_Z = 0$ .

**1.3.3 Bemerkung:**

Die Valenz oder Wertigkeit eines Atoms ist die Summe der von ihm ausgehenden kovalenten Bindungen unter Berücksichtigung ihrer Vielfachheit. Sie entspricht also dem Grad eines Knotens im Multigraphen. Beispielsweise ist die Valenz für H-Atome 1, für O-Atome 2, für N-Atome 3 und für C-Atome 4. Dies gilt allerdings nur, solange sich Atome im Grundzustand befinden, d.h. keine Ladungen oder ungepaarte Elektronen besitzen. Es gibt Elemente, wie beispielsweise Phosphor oder Schwefel, deren Atome auch im Grundzustand mehrere verschiedene Valenzen annehmen können, d.h. sie besitzen mehrere Grundzustände. So gibt es Verbindungen mit 3- und 5-valentem Phosphor. Schwefel kann mit Valenzen 2, 4 und 6 auftreten. Dies wird dadurch ermöglicht, dass bei diesen Elementen die Anzahl der freien Elektronenpaare in den Grundzuständen variiert. Verlässt man die Grundzustände, so sind eine ganze Reihe weiterer Valenzen für die verschiedenen Elemente möglich. Mathematisch erfasst man diese Vielfalt, indem man für jedes chemische Element  $X$  eine Menge  $\mathcal{Z}_X$  gültiger Atomzustände definiert. Diese Definition ist natürlich abhängig von der jeweiligen Situation, auf die das Modell angewendet werden soll.

**1.3.4 Beispiel:**

Die vier wichtigsten Elemente in der organischen Chemie sind Kohlenstoff, Wasserstoff, Stickstoff und Sauerstoff. Wir setzen deshalb

$$\mathcal{E}_4 := \{\text{H, C, N, O}\}.$$

Darüber hinaus treten in der organischen Chemie weitere Elemente auf. Ohne Beschränkung der Allgemeinheit werden wir in vielen Beispielen Fluor, Silizium, Phosphor, Schwefel, Chlor, Brom und Jod hinzuziehen. Wir bezeichnen diese Menge mit

$$\mathcal{E}_{11} := \{\text{H, C, N, O, F, Si, P, S, Cl, Br, I}\}.$$

Tabelle 1.1 enthält für die Elemente  $X \in \mathcal{E}_{11}$  die Ordnungszahl  $TE_X$ , die Anzahl von Valenzelektronen  $VE_X$  sowie eine Aufstellung (nach [162]) aller Atomzustände  $Z$ , die für das Verhalten organischer Verbindungen in einem Massenspektrometer relevant sind. Nach Bemerkung 1.3.1 muss für den Atomzustand  $Z$  eines Atoms vom Element  $X$  gelten:

$$v_Z + 2p_Z + q_Z + r_Z = VE_X.$$

Aufgrund dieser Eigenschaft wurde bei der Implementierung [56] von Atomzuständen auf die explizite Speicherung der Anzahl freier Elektronenpaare verzichtet.

$X (TE_X, VE_X)$	$v_Z$	$p_Z$	$q_Z$	$r_Z$	RC	CSC
H (1, 1)	1	0	0	0	x	x
	0	0	1	0		
	0	0	0	1		
C (6, 4)	4	0	0	0	x	x
	3	0	1	0		
	3	0	0	1		
	2	0	1	1		
N (7, 5)	4	0	1	0		
	3	1	0	0	x	x
	3	0	1	1		
	2	1	0	1		
O (8, 6)	3	1	1	0		
	2	2	0	0	x	x
	2	1	1	1		
	1	2	0	1		
F (9, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		
Si (14, 4)	4	0	0	0	x	x
	3	0	1	0		
	3	0	0	1		
	2	0	1	1		
P (15, 5)	5	0	0	0		x
	4	0	1	0		
	4	0	0	1		
	3	1	0	0	x	x
	3	0	1	1		
	2	1	0	1		
S (16, 6)	6	0	0	0		x
	5	0	1	0		
	5	0	0	1		
	4	1	0	0		x
	4	0	1	1		
	3	1	1	0		
	3	1	0	1		
	2	2	0	0	x	x
	2	1	1	1		
1	2	0	1			
Cl (17, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		
Br (35, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		
I (53, 7)	2	2	1	0		
	1	3	0	0	x	x
	1	2	1	1		

Tabelle 1.1: Gültige Atomzustände für die Elemente aus  $\mathcal{E}_{11}$  im Massenspektrometer

**1.3.5 Bemerkung:**

Die Menge der gültigen Atomzustände  $\mathcal{Z}_X$  für Element  $X$  ist abhängig von der zugrunde liegenden *Chemie* für die jeweils betrachtete Situation. Die in [33] beschriebene hierarchische Klassifizierung chemischer Verbindungen lässt sich sehr treffend anhand der gültigen Atomzustände nachvollziehen.

Unter *eingeschränkter Chemie* (engl. *Restricted Chemistry*, kurz *RC*) werden alle Verbindungen zusammengefasst, deren Atome keine Ladungen oder ungepaarte Elektronen besitzen, und zudem der *Oktettregel* genügen. Diese besagt, auf die Nomenklatur der Atomzustände übertragen, dass die Summe der Elektronen, die an Bindungen beteiligt sind und der, die freie Elektronenpaare bilden, für den Zustand  $Z$  jedes Atoms acht sein muss, also insgesamt

$$q_Z = 0 \wedge r_Z = 0 \wedge 2v_Z + 2p_Z = 8.$$

Einzigste Ausnahme bildet dabei Wasserstoff. Die Atomzustände der *RC* sind in Tabelle 1.1 mit einem Kreuz in der Spalte *RC* gekennzeichnet. Eine Ausnahme bildet dabei Wasserstoff. Im Rahmen der *RC* ist es insbesondere möglich, jedem Element  $X \in \mathcal{E}_{11}$  *genau eine* Valenz  $v_X$  zuzuordnen. Wir bezeichnen  $v_X$  als *Standardvalenz* von  $X$ .

Verzichtet man auf die Einhaltung der Oktettregel, so begibt man sich auf das Gebiet der *Chemie der abgeschlossenen Schalen* (engl. *Closed Shell Chemistry*, kurz *CSC*). Nach wie vor muss für Atomzustände  $Z$  gelten:  $q_Z = 0$  und  $r_Z = 0$ . Für diese Zustände findet man in Spalte *CSC* von Tabelle 1.1 eine Markierung.

Lässt man auch diese Regel fallen, spricht man von *ganzzahliger Chemie* (engl. *Integral Chemistry*, kurz *IC*). Diese begrenzt zugleich den Darstellungsbereich des hier verwendeten Modells. Charakteristisch für diese Chemie ist, dass alle Bindungsvielfachheiten durch ganze Zahlen beschrieben werden können. Es gibt in der Chemie jedoch Phänomene, die eine derartige Darstellung nicht mehr erlauben (vgl. Abschnitt 1.6.1). Mesomerie und Multizentren-Bindungen müssen durch Erweiterungen unseres Molekülmodells beschrieben werden. Fasst man diese Objekte unter dem Begriff *Multizentren-Chemie* (engl. *Multicenter Chemistry*, kurz *MC*) zusammen, kann man folgende Inklusionen notieren:

$$RC \subset CSC \subset IC \subset MC.$$

Mit den obigen Bezeichnungen konnten durch den Strukturgenerator *MOLGEN* bis Version 3.5 [13] Verbindungen aus *RC* beschrieben werden. Ab Version 4.0 [56] ist es möglich, Moleküle aus *IC* darzustellen. Im Folgenden präzisieren wir diese Darstellungsweise.

**1.3.6 Definition:**

Sei  $n \in \mathbb{N}^*$ ,  $\mathcal{E}$  eine Menge chemischer Elemente und

$$\mathcal{Z}_{\mathcal{E}} := \bigcup_{X \in \mathcal{E}} \mathcal{Z}_X$$

die Menge gültiger Atomzustände der Elemente aus  $\mathcal{E}$ . Dann bezeichnet man  $\zeta \in \mathcal{Z}_{\mathcal{E}}^n := (\mathcal{Z}_{\mathcal{E}})^n$  als *Zustandsverteilung* und  $\varepsilon \in \mathcal{E}^n$ . Ein Tripel

$$(\varepsilon, \zeta, \gamma) \in \mathcal{E}^n \times \mathcal{Z}_{\mathcal{E}}^n \times \mathcal{G}_{n,4}$$

heißt *molekularer Graph*, wenn

- (i)  $\forall i \in n : \zeta(i) \in \mathcal{Z}_{\varepsilon(i)}$ ,
- (ii)  $\forall i \in n : \text{grad}_{\gamma}(i) = v_{\zeta(i)}$ .

$\mathcal{M}_n$  bezeichne die Menge der molekularen Graphen mit  $n$  Atomen und

$$\mathcal{M} := \bigcup_{n \in \mathbb{N}^*} \mathcal{M}_n$$

die Menge aller molekularen Graphen.

**1.3.7 Bemerkung:**

Molekulare Graphen mit  $n$  Atomen sind also gerade die Elemente

$$(\varepsilon, \zeta, \gamma) \in \mathcal{E}^n \times \mathcal{Z}_{\mathcal{E}}^n \times \mathcal{G}_{n,4},$$

bei denen (i) die Zustandsverteilung  $\zeta$  gültige Atomzustände bzgl. der Elementverteilung  $\varepsilon$  angibt und (ii) die Knotengrade des Multigraphen  $\gamma$  übereinstimmen mit den Valenzen, die  $\zeta$  für die Atome vorschreibt. Die in Abschnitt 1.2 eingeführten Bezeichnungen und Eigenschaften für Graphen  $\gamma$  werden auch für molekulare Graphen  $M = (\varepsilon, \zeta, \gamma)$  verwendet.  $(\varepsilon, \zeta, \gamma)$  ist ein *zusammenhängender molekularer Graph*, wenn  $\gamma$  zusammenhängend ist.  $\mathcal{M}_n^C$  bezeichnet die Menge der zusammenhängenden molekularen Graphen mit  $n$  Atomen,  $\mathcal{M}^C$  die aller zusammenhängenden molekularen Graphen. Für die reaktionsbasierte Strukturgenerierung (Abschnitt 2.2) werden wir die Zerlegung molekularer Graphen bzgl. ihrer Zusammenhangskomponenten sowie die Summe molekularer Graphen benötigen:



**1.3.8 Definition:**

Sei  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$ ,  $l := |\text{Con}(\gamma)|$  und  $\text{Con}(\gamma) = \{\gamma|_{V_0}, \dots, \gamma|_{V_{l-1}}\}$ , wobei  $n_k := |V_k|$  und  $V_k = \{v_i^k \mid i \in n_k\}$  für  $k \in l$ . Dann sind die durch die Zusammenhangskomponenten von  $\gamma$  induzierten molekularen Teilgraphen von  $M$  für  $k \in l$ :

$$M_k = (\varepsilon_k, \zeta_k, \gamma_k) \in \mathcal{E}^{n_k} \times \mathcal{Z}_{\mathcal{E}}^{n_k} \times \mathcal{G}_{n_k,4},$$

wobei für  $i, j \in n_k$ ,  $i \neq j$ :

$$\begin{aligned} \varepsilon_k(i) &:= \varepsilon(v_i^k), \\ \zeta_k(i) &:= \zeta(v_i^k), \\ \gamma_k(\{i, j\}) &:= \gamma(\{v_i^k, v_j^k\}). \end{aligned}$$

**1.3.9 Definition:**

Sei  $l \in \mathbb{N}^*$  und  $M_k = (\varepsilon_k, \zeta_k, \gamma_k) \in \mathcal{M}_{n_k}$  für  $k \in l$  molekulare Graphen. Dann ist die Summe der  $M_k$  erklärt als

$$\bigoplus_{k \in l} M_k := (\varepsilon, \zeta, \gamma) \in \mathcal{E}^n \times \mathcal{Z}_{\mathcal{E}}^n \times \mathcal{G}_{n,4},$$

wobei  $n := \sum_{k \in l} n_k$  und für  $v_i^k := i + \sum_{\nu \in k} n_\nu$ ,  $k \in l$ ,  $i \in n_k$ :

$$\begin{aligned} \varepsilon(v_i^k) &:= \varepsilon_k(i), \\ \zeta(v_i^k) &:= \zeta_k(i), \end{aligned}$$

sowie für  $h \in l$ ,  $j \in n_h$ ,  $v_i^h \neq v_j^k$ :

$$\gamma(\{v_i^k, v_j^h\}) := \begin{cases} \gamma_k(i, j) & \text{falls } h = k, \\ 0 & \text{sonst.} \end{cases}$$

**1.3.10 Bemerkung:**

Die durch Zusammenhangskomponenten induzierten molekularen Teilgraphen  $M_k$ ,  $k \in l$  eines molekularen Graphen  $M$  sind ihrerseits molekulare Graphen. Wir bezeichnen diese Familie mit  $\text{Con}(M)$ . Ebenso ist die Summe molekularer Graphen wieder ein molekularer Graph.

Ohne darauf im Einzelnen eingehen zu wollen, werden wir die in Abschnitt 1.2 eingeführten topologischen Eigenschaften für Graphen sowie deren Schreibweisen auf molekulare Graphen übertragen.

**1.3.11 Bemerkung:**

Als *molekulare Graphen einer chemischen Verbindung* bezeichnen wir diejenigen Elemente aus  $\mathcal{M}$ , welche den Bindungsverhältnissen und Atomzuständen der Atome eines Moleküls der Verbindung entsprechen. Um eine eindeutige Zuordnung von chemischen und mathematischen Strukturen zu erreichen, bedienen wir uns erneut einer Operation der symmetrischen Gruppe  $S_n$ .

**1.3.12 Definition:**

Sei  $n \in \mathbb{N}^*$ ,  $\mathcal{E}$  eine Menge chemischer Elemente und  $\mathcal{Z}_{\mathcal{E}} = \bigcup_{X \in \mathcal{E}} \mathcal{Z}_X$  die gültigen Atomzustände zu den Elementen aus  $\mathcal{E}$ . Dann ist durch

$$\begin{aligned} (\mathcal{E}^n \times \mathcal{Z}_{\mathcal{E}}^n \times \mathcal{G}_{n,4}) \times S_n &\longrightarrow \mathcal{E}^n \times \mathcal{Z}_{\mathcal{E}}^n \times \mathcal{G}_{n,4}, \\ ((\varepsilon, \zeta, \gamma), \pi) &\longmapsto (\varepsilon, \zeta, \gamma)^\pi, \end{aligned}$$

wobei

$$(\varepsilon, \zeta, \gamma)^\pi := (\varepsilon^\pi, \zeta^\pi, \gamma^\pi)$$

und für  $i, j \in n$ ,  $i \neq j$

$$\begin{aligned} \varepsilon^\pi(i) &:= \varepsilon(\pi(i)), \\ \zeta^\pi(i) &:= \zeta(\pi(i)), \\ \gamma^\pi(\{i, j\}) &:= \gamma(\{\pi(i), \pi(j)\}) \end{aligned}$$

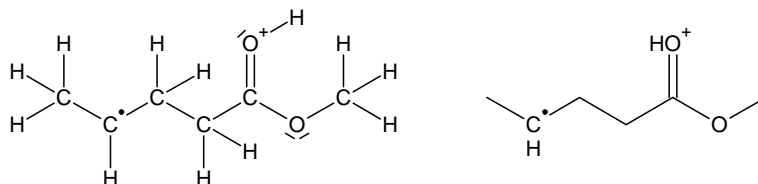
eine Operation von  $S_n$  auf  $\mathcal{E}^n \times \mathcal{Z}_{\mathcal{E}}^n \times \mathcal{G}_{n,4}$  erklärt. Die Eigenschaften (i) und (ii) aus Definition 1.3.6 bleiben unter dieser Operation erhalten und somit operiert  $S_n$  auch auf  $\mathcal{M}_n$ . Zwei molekulare Graphen  $M, M' \in \mathcal{M}_n$  sind zueinander *isomorph*, wenn sie in der gleichen Bahn unter dieser Operation liegen. Die Elemente von  $\mathcal{M}_n // S_n$  heißen *Isomorphieklassen molekularer Graphen*. Die Isomorphieklasse von  $M$  soll mit  $\bar{M}$  bezeichnet werden.

**1.3.13 Bemerkung:**

Isomorphieklassen zusammenhängender molekularer Graphen können mit *Strukturformeln* chemischer Verbindungen identifiziert werden. Gemäß [90] können molekulare Graphen mit polynomialem Aufwand auf Isomorphie getestet werden. In [20] wird eine *kanonische Form* auf  $\mathcal{M}_n$  beschrieben. Diese werden wir im Folgenden mit  $\kappa$  bezeichnen, und verwenden, um kanonische Bahnrepräsentanten zu berechnen und molekulare Graphen auf Isomorphie zu testen.

**1.3.14 Schreibweise:**

Die Darstellung chemischer Strukturformeln unterliegt einer Reihe sinnvoller Konventionen. Zunächst können wir einen molekularen Graphen als Multi-graphen mit Knotenbeschriftungen darstellen (links):



Die Knoten tragen die Elementsymbole, Ladungen werden links oberhalb des Elementsymbols angefügt, ebenso ungepaarte Elektronen, die durch einen Punkt dargestellt werden. Freie Elektronenpaare können als Striche am Elementsymbol eingetragen werden. Zur besseren Übersicht vernachlässigt man meist H-Atome, Elementsymbole von C-Atomen und freie Elektronenpaare (rechts). Für *Heteroatome*, d.h. Atome außer C und H, werden H-Atome direkt neben dem Elementsymbol geschrieben. Ebenso verfährt man bei C-Atomen mit Ladungen oder ungepaarten Elektronen.

**1.3.15 Definition:**

Sei  $\mathcal{E}$  eine Menge chemischer Elemente. Eine *Bruttoformel* ist eine Abbildung  $\beta \in \mathbb{N}^{\mathcal{E}}$ . Die Bruttoformel eines molekularen Graphen  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$  ist definiert durch

$$\beta_M(X) := |\{i \in n \mid \varepsilon(i) = X\}|.$$

Synonym verwendet man auch den Begriff *Summenformel*. Die Menge der Strukturformeln

$$\bar{\mathcal{M}}_{\beta}^{\mathcal{C}} := \{\bar{M} \mid M \in \mathcal{M}^{\mathcal{C}}, \beta_M = \beta\}$$

zur Bruttoformel  $\beta$  bezeichnet man als *Konstitutionsisomere* (oder einfach nur *Isomere*) von  $\beta$ .

**1.3.16 Beispiel:**

In Anhang E sind Anzahlen von Konstitutionsisomeren zu Summenformeln mit Elementen aus  $\mathcal{E}_4$  zusammengestellt.

**1.3.17 Schreibweise:**

Die Bruttoformel einer chemischen Verbindung umfasst also nur Information über die Anzahlen der in einem Molekül der Verbindung enthaltenen Atome der einzelnen Elemente. Dementsprechend ist es in der Chemie üblich, die Symbole der in der Verbindung enthaltenen Elemente gefolgt von der jeweiligen Anzahl von Atomen hintereinander zu schreiben. Tritt dabei ein Element nur einmal auf, wird auf die Angabe der Anzahl verzichtet. Beispielsweise gehört die Bruttoformel



zu einer Verbindung mit 2 Kohlenstoff-, 6 Wasserstoff-Atomen und einem Sauerstoff-Atom. Zuerst wird Kohlenstoff gefolgt von Wasserstoff aufgeführt. Danach folgen die übrigen Elemente in alphabetischer Reihenfolge ihrer Symbole. Um Verwechslungen mit mathematischer Indizierung auszuschließen schreiben wir Bruttoformeln stets *sans serif*.

**1.3.18 Bemerkung:**

Wie zu Beginn dieses Abschnitts erwähnt wurde, ist es im Allgemeinen nicht möglich, jedem chemischen Element eine feste Valenz zuzuordnen. Deshalb können wir die Existenzsätze 1.2.7 und 1.2.15 für Multigraphen zu gegebener Gradpartition nicht ohne weiteres auf molekulare Graphen und Bruttoformeln übertragen. In Kapitel 5 werden wir aber dennoch derartige Kriterien benötigen. Dazu bedienen wir uns der Valenzen  $v_X$  aus Bemerkung 1.3.5 für RC. Auf diese Zahlenwerte greifen unter anderem auch [91] und [108] zurück, wenn sie Regeln für die Plausibilität von Bruttoformeln im Rahmen molekularer Strukturaufklärungsprobleme angeben. Aus 1.2.7 und 1.2.15 folgt für Bruttoformeln unter Bedingungen der eingeschränkten Chemie:

**1.3.19 Satz:**

Zu  $\beta \in \mathbb{N}^{\mathcal{E}}$  existiert genau dann mindestens ein molekularer Graph  $M$  mit  $\beta_M = \beta$ , wenn

$$\text{(Gr1)} \quad \sum_{X \in \mathcal{E}} v_X \beta(X) \equiv 0 \pmod{2} \text{ und}$$

$$\text{(Gr2)} \quad \sum_{X \in \mathcal{E}} v_X \beta(X) - 2 \max \{v_X \mid X \in \mathcal{E}, \beta(X) > 0\} \geq 0.$$

Für die Existenz mindestens eines zusammenhängenden molekularen Graphen muss zudem

$$\text{(Con)} \quad \sum_{X \in \mathcal{E}} v_X \beta(X) - 2 \sum_{X \in \mathcal{E}} \beta(X) + 2 \geq 0$$

gelten. Wir bezeichnen die Menge der Bruttoformeln, die (Gr1), (Gr2) und (Con) erfüllen mit  $\mathcal{B}_{\mathcal{E}}^C$ .

**1.3.20 Bemerkung:**

Möchte man mehrere Valenzen  $v_{X,i}$ ,  $i \in n_X$  für ein Element zulassen, werden die Tests aufwändiger. Dann müssen Partitionen  $\mu_X = (\mu_X(i))_{i \in n_X} \models \beta(X)$  betrachtet werden.  $\mu_X(i)$  steht dabei für die Anzahl von Atomen des Elements  $X$  mit Valenz  $v_{X,i}$ . Es gibt genau dann mindestens einen molekularen Graphen mit Bruttoformel  $\beta$ , wenn Partitionen  $\mu_X \models \beta(X)$ ,  $X \in \mathcal{E}$  existieren, so dass

$$(\text{Gr1}') \quad \sum_{X \in \mathcal{E}} \sum_{i \in n_X} v_{X,i} \mu_X(i) \equiv 0 \pmod{2} \text{ und}$$

$$(\text{Gr2}') \quad \sum_{X \in \mathcal{E}} \sum_{i \in n_X} v_{X,i} \mu_X(i) - 2 \max \{v_{X,i} \mid X \in \mathcal{E}, i \in n_X, \mu_X(i) > 0\} \geq 0.$$

Wie gehabt muss für die Existenz eines zusammenhängenden molekularen Graphen zusätzlich

$$(\text{Con}') \quad \sum_{X \in \mathcal{E}} \sum_{i \in n_X} v_{X,i} \mu_X(i) - 2 \sum_{X \in \mathcal{E}} \beta(X) + 2 \geq 0$$

erfüllt sein.

In der Literatur wird in diesem Zusammenhang oft der Begriff des *Doppelbindungsäquivalents* (engl. *Double Bond Equivalent*, kurz *DBE*) genannt. Das DBE zur Bruttoformel  $\beta$  wird berechnet als

$$\text{DBE}(\beta) = \frac{1}{2} \left( 2 + \sum_{X \in \mathcal{E}} \beta(X)(v_X - 2) \right).$$

Die Bedingungen (Gr1) und (Con) werden dann meist unter Verwendung des DBE formuliert.

**1.3.21 Definition:**

Sei  $\mathcal{E}$  eine Menge chemischer Elemente. Auf  $\mathbb{N}^{\mathcal{E}}$  ist die *Teilmengenrelation für Bruttoformeln* wie folgt definiert: Für  $\beta, \beta' \in \mathbb{N}^{\mathcal{E}}$  ist

$$\beta' \subseteq \beta \quad :\iff \quad \forall X \in \mathcal{E} : \beta'(X) \leq \beta(X).$$

„ $\subseteq$ “ erklärt eine Halbordnung auf  $\mathbb{N}^{\mathcal{E}}$ .

**1.3.22 Definition:**

Sei  $\mathcal{E}$  eine Menge chemischer Elemente und  $\beta \in \mathbb{N}^{\mathcal{E}}$ ,  $\beta \neq 0$  eine Bruttoformel. Die *empirische Formel* zu  $\beta$  ist  $\beta' \in \mathbb{N}^{\mathcal{E}}$  mit

$$\beta'(X) = \frac{\beta(X)}{\text{ggt}(\beta(\mathcal{E}))}$$

für  $X \in \mathcal{E}$ . Dabei bezeichnet  $\text{ggt}(\beta(\mathcal{E}))$  den größten gemeinsamen Teiler der Elemente von  $\{\beta(X) \mid X \in \mathcal{E}\}$ .

**1.3.23 Bemerkung:**

Die empirische Formel einer chemischen Verbindung ist die empirische Formel ihrer Bruttoformel. Sie gibt lediglich Auskunft über die Verhältnisse der Anzahlen von Atomen der in einem Molekül der Verbindung auftretenden Elemente.

## 1.4 Molekulare Substrukturen

In Abschnitt 1.2 haben wir Sub- und Teilgraphen kennen gelernt. Wir wollen diese Begriffe im Folgenden auf molekulare Graphen übertragen.

### 1.4.1 Definition:

Sei  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$  und  $k \leq n$ . Ein Tripel

$$S = (\varepsilon', \zeta', \gamma') \in \mathcal{E}^k \times \mathcal{Z}_{\mathcal{E}}^k \times \mathcal{G}_{k,4}$$

ist *Substruktur* von  $M$ , wenn

$$\exists \phi \in \text{Emb}_{\subseteq}(\gamma', \gamma) : \forall i \in k : \varepsilon(\phi(i)) = \varepsilon'(i) \wedge \zeta(\phi(i)) = \zeta'(i).$$

Analog werden geschlossene und induzierte Substrukturen erklärt. Um strukturelle Eigenschaften molekularer Graphen möglichst genau und flexibel beschreiben zu können, bedarf es einiger weiterer Definitionen.

### 1.4.2 Definition:

Sei  $n \in \mathbb{N}^*$ ,  $\mathcal{E}$  eine Menge chemischer Elemente und  $\mathcal{Z}_{\mathcal{E}} = \bigcup_{X \in \mathcal{E}} \mathcal{Z}_X$  die gültigen Atomzustände zu den Elementen aus  $\mathcal{E}$ . Ein Tripel

$$MMG := (E, Z, \Gamma) \in \mathcal{P}^*(\mathcal{E})^n \times \mathcal{P}^*(\mathcal{Z}_{\mathcal{E}})^n \times \mathcal{P}(\underline{\mathbb{3}})^{\binom{n}{2}} =: \mathcal{MMG}_n$$

heißt *mehrdeutiger molekularer Graph* (kurz *MMG*). Dabei bezeichnet  $\mathcal{P}$  die Potenzmenge,  $\mathcal{P}^*$  die Potenzmenge ohne die leere Menge  $\emptyset$  und  $\underline{m}$  die Menge  $\{1, \dots, m\}$ . Die Kantenmenge von *MMG* ist definiert als

$$E_{\Gamma} := \left\{ e \in \binom{n}{2} \mid \Gamma(e) \neq \emptyset \right\}.$$

### 1.4.3 Bemerkung:

Während bislang die Implementation der Datenstrukturen recht offensichtlich erschien, wird an dieser Stelle eine erste ergänzende Bemerkung notwendig. Natürlich könnte man die Verteilung der Elemente und Atomzustände eines mehrdeutigen molekularen Graphen jeweils durch ein Feld oder eine Liste aller erlaubten Tupel von Elementen und Atomzuständen  $(E(i), Z(i))$  realisieren. Dies wäre aber spätestens dann höchst ineffizient, wenn  $E(i) = \mathcal{E}$  und  $Z(i) = \mathcal{Z}_{\mathcal{E}}$  dargestellt werden soll. Dieses Problem wurde durch die Einführung einer abstrakten Basisklasse *Atomtyp* gelöst. Dieser Datentyp verfügt über eine Funktion

$$\text{Kompatibel} : \mathcal{E} \times \mathcal{Z}_{\mathcal{E}} \longrightarrow \mathbb{B},$$

die im Rahmen einer Substruktursuche oder Strukturgenerierung aufgerufen wird, um festzustellen, ob für den Knoten  $j$  eines molekularen Graphen  $(\varepsilon, \zeta, \gamma) \in \mathcal{M}$  gilt:  $\varepsilon(j) \in E(i)$  und  $\zeta(j) \in Z(i)$ . Folgende Atomtypen sind derzeit implementiert:

- Atomtyp *Standard* kann genau ein Element und einen Atomzustand darstellen, und deckt damit die Fälle ab, wenn  $|E(i)| = |Z(i)| = 1$ .
- Atomtyp *Multi* kann ein Feld von Tupeln aus Elementen und Atomzuständen darstellen. Dieser Atomtyp kann allgemein verwendet werden.
- Atomtyp *Any* findet Verwendung, wenn  $E(i) = \mathcal{E}$  und  $Z(i) = \mathcal{Z}_{\mathcal{E}}$ . Die Funktion *Kompatibel* liefert bei diesen Atomtyp immer das Ergebnis *wahr*.

Neben der Effizienzsteigerung ist der wesentliche Vorteil dieser Technik die Tatsache, dass nachträgliche Erweiterungen einfach hinzugefügt werden können. So wäre es beispielsweise denkbar, dass ein Atomtyp *Element* benötigt wird, der alle Atomzustände eines Elements darstellen kann. Dann muss lediglich der neue Atomtyp inklusive der Funktion *Kompatibel* definiert werden. Veränderungen an der Substruktursuche oder gar an den Algorithmen zur Strukturgenerierung bleiben dem Programmierer erspart. Auf diesem Wege wurde ein Atomtyp hinzugefügt, der speziell der Darstellung von Fragmentierungsreaktionen im Massenspektrometer gerecht wird (vgl. Abschnitt 5.4.2).

- Atomtyp *MS* unterscheidet nach Zugehörigkeit von  $\varepsilon(j)$  zu einer der folgenden Mengen von Elementen:
  - alle Elemente,
  - alle *schweren* Elemente, d.h. Elemente außer H,
  - alle schweren Elemente außer C,
  - alle Elemente mit freien Elektronenpaaren (N, O, P, S, Halogene).

$\zeta(j)$  wird dabei hinsichtlich der Existenz einer positiven Ladung bzw. einer Radikalstelle untersucht.



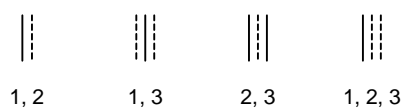
**1.4.4 Schreibweise:**

Bei mehrdeutigen molekularen Graphen werden die Knoten mit den Alternativen für Elemente und Atomzustände beschriftet bzw. mit Symbolen, welche die verschiedenen Alternativen darstellen. Typische Schreibweisen sind dabei

A für ein beliebiges Atom und

Q für ein Heteroatom.

Weitere Beispiele findet man in Abschnitt 5.4.2. Alternativen für Bindungen werden folgendermaßen kodiert:

**1.4.5 Definition:**

Sei  $k, n \in \mathbb{N}^*$ ,  $k \leq n$ ,  $\mathcal{E}$  eine Menge chemischer Elemente,  $\mathcal{Z}_{\mathcal{E}} = \bigcup_{X \in \mathcal{E}} \mathcal{Z}_X$  die gültigen Atomzustände zu den Elementen aus  $\mathcal{E}$  und  $M \in \mathcal{M}_n$ . Ein mehrdeutiger molekularer Graph  $MMG = (E, Z, \Gamma) \in \mathcal{MMG}_k$  heißt *mehrdeutiger molekularer Subgraph* von  $M$  wenn eine injektive Abbildung  $\phi \in n_{\text{inj}}^k$  existiert, so dass

- (i)  $\forall i \in k : \varepsilon(\phi(i)) \in E(i),$
- (ii)  $\forall i \in k : \zeta(\phi(i)) \in Z(i),$
- (iii)  $\forall \{i, j\} \in E_{\Gamma} : \gamma(\{\phi(i), \phi(j)\}) \in \Gamma(\{i, j\}).$

$\phi$  heißt dann *Einbettung* von  $MMG$  in  $M$  als mehrdeutiger molekularer Subgraph und man schreibt  $MMG \subseteq_{\phi} M$ . Gilt zudem

- (iv)  $\forall \{i, j\} \in \binom{k}{2} \setminus E_{\Gamma} : \gamma(\{\phi(i), \phi(j)\}) = 0,$

so ist  $MMG$  *mehrdeutiger molekularer Teilgraph* von  $M$ .  $\phi$  ist dann *Einbettung* von  $MMG$  in  $M$  als mehrdeutiger molekularer Teilgraph und man schreibt  $MMG \subseteq_{\phi}^i M$ .

**1.4.6 Bemerkung:**

Oft müssen strukturelle Eigenschaften verarbeitet werden, die sich nicht alleine durch mehrdeutige molekulare Graphen ausdrücken lassen. So ist beispielsweise denkbar, dass zwei Atome eines mehrdeutigen molekularen Graphen bzgl. ihrer Einbettung in einen molekularen Graphen vorgegebene Abstände einhalten müssen, oder auf Ringen gegebener Länge liegen sollen, oder nicht liegen dürfen. Zu diesem Zweck können mehrdeutige molekulare Graphen mit Substruktur-Restriktionen versehen werden.

**1.4.7 Definition:**

Sei  $k \in \mathbb{N}^*$ . Eine *Substruktur-Restriktion* ist eine Abbildung

$$SR : \bigcup_n (\mathcal{M}_n \times n_{\text{inj}}^k) \longrightarrow \mathbb{B}.$$

$\mathcal{SR}_k$  bezeichne die Menge der Substruktur-Restriktionen auf  $k$  Atomen.

**1.4.8 Bemerkung:**

Ebenso wie der Atomtyp ist die Substruktur-Restriktion als abstrakte Basisklasse implementiert. Derzeit sind folgende Typen von Substruktur-Restriktionen verfügbar:

- Substruktur-Restriktion *Distanz*: Für zwei Atome  $i, j \in k$  wird ein Intervall  $[a, b] \subset \mathbb{N}^*$  angegeben, welches den Abstand der beiden Atome bezüglich der Einbettung  $\phi$  des mehrdeutigen molekularen Graphen in  $M$  festlegt.

$$SR_{\{i,j\},[a,b]}^{\text{Dist}} : (M, \phi) \longmapsto \begin{cases} \text{wahr,} & \text{falls } \text{dist}_M(\phi(i), \phi(j)) \in [a, b], \\ \text{falsch,} & \text{sonst.} \end{cases}$$

- Substruktur-Restriktion *Hybridisierung*: Für eine nichtleere Teilmenge  $\{i_j \mid j \in h\} \subseteq k$  von Atomen des mehrdeutigen molekularen Graphen wird eine Hybridisierung  $\eta$  angegeben mit der diese Atome bezüglich der Einbettung  $\phi$  des mehrdeutigen molekularen Graphen in  $M$  vorliegen müssen.

$$SR_{\{i_j \mid j \in h\}, \eta}^{\text{Hybrid}} : (M, \phi) \longmapsto \begin{cases} \text{wahr,} & \text{falls } \forall j \in h : \text{hyb}_M(\phi(i_j)) = \eta, \\ \text{falsch,} & \text{sonst.} \end{cases}$$

- Substruktur-Restriktion *Nachbarschaft*: Für eine nichtleere Teilmenge von Atomen des mehrdeutigen molekularen Graphen wird eine Substruktur angegeben, zu der die gewählten Atome bezüglich der Einbettung des mehrdeutigen molekularen Graphen in  $M$  einen bestimmten, in Form eines Intervalls spezifizierten Abstand einnehmen müssen.
- Substruktur-Restriktion *Ring*: Für eine nichtleere Teilmenge von Atomen des mehrdeutigen molekularen Graphen wird eine Ringlänge in Form eines Intervalls angegeben. Die gewählten Atome müssen oder dürfen nicht bezüglich der Einbettung des mehrdeutigen molekularen Graphen in  $M$  auf einem Ring der vorgegebenen Länge liegen.

**1.4.9 Definition:**

Sei  $k \in \mathbb{N}^*$  und  $h \in \mathbb{N}$ . Eine *molekulare Substruktur*

$$S := (MMG, \{SR_i \mid i \in h\}) \in MMG_k \times \mathcal{P}(\mathcal{SR}_k) =: \mathcal{S}_k$$

ist ein Tupel bestehend aus einem mehrdeutigen molekularen Graphen und einer Menge von Substruktur–Restriktionen.

**1.4.10 Definition:**

Sei  $k, n \in \mathbb{N}^*$ ,  $k \leq n$ ,  $S = (MMG, \{SR_i \mid i \in h\}) \in \mathcal{S}_k$  eine molekulare Substruktur und  $M \in \mathcal{M}_n$  ein molekularer Graph. Eine injektive Abbildung  $\phi \in n_{\text{inj}}^k$  heißt Einbettung von  $S$  in  $M$  als *molekulare Substruktur*, wenn

- (i)  $MMG \subseteq_{\phi} M$  und
- (ii)  $\forall i \in h : SR_i(M, \phi) = \text{wahr}$ .

Wir schreiben dann  $S \subseteq_{\phi} M$  und nennen  $S$  molekulare Substruktur von  $M$ . Gilt zudem

- (iii)  $MMG \subseteq_{\phi}^i M$ ,

so ist  $\phi$  Einbettung von  $S$  in  $M$  als *molekulare Teilstruktur*. Wir schreiben dann  $S \subseteq_{\phi}^i M$  und nennen  $S$  molekulare Teilstruktur von  $M$ . Mit

$$\begin{aligned} \text{Emb}_{\subseteq}(S, M) &:= \{\phi \in n_{\text{inj}}^k \mid S \subseteq_{\phi} M\} \\ \text{Emb}_{\subseteq^i}(S, M) &:= \{\phi \in n_{\text{inj}}^k \mid S \subseteq_{\phi}^i M\} \end{aligned}$$

bezeichnen wir wie gehabt die entsprechenden Mengen von Einbettungen.

**1.4.11 Bemerkung:**

Ein Algorithmus, der nach Einbettungen einer Substruktur in einem molekularen Graphen sucht, heißt *Substruktursuche*. Substruktursuche wird für verschiedene Belange der Computerchemie benötigt. So erhalten Strukturgeneratoren [3, 13, 14, 51, 53, 56, 99] Listen von Substrukturen (so genannte *Good-* und *Badlist*) als Eingabe, um nur solche molekulare Graphen zu konstruieren, die diese Substrukturen enthalten bzw. nicht enthalten. In [53] findet man eine algorithmische Beschreibung der Substruktursuche. Die Aussage über das Vorhandensein einer Substruktur in einem molekularen Graphen kann als *binärer molekularer Deskriptor* (vgl. Abschnitt 4.3.2) aufgefasst werden. In [152] wird ein Vektor solcher binärer Deskriptoren verwendet, um die Ähnlichkeit molekularer Graphen zu bestimmen. Im Abschnitt 1.5 wird beschrieben, wie mit Hilfe molekularer Substrukturen Reaktionsschemata definiert und chemische Reaktionen im Computer simuliert werden können.

## 1.5 Chemische Reaktionen

Veränderungen an chemischen Verbindungen werden in der Chemie durch Reaktionen beschrieben. Wir wollen diese Veränderungen mathematisch spezifizieren, um sie auf unser graphisches Modell molekularer Strukturen anwenden zu können. Wir folgen dabei der Vorgehensweise aus [42] und [140].

### 1.5.1 Definition:

Sei  $n \in \mathbb{N}^*$ ,  $\mathcal{E}$  eine Menge chemischer Elemente und  $\mathcal{Z}_{\mathcal{E}} = \bigcup_{X \in \mathcal{E}} \mathcal{Z}_X$  die gültigen Atomzustände zu den Elementen aus  $\mathcal{E}$ . Ein Tupel

$$C := (M, M') \in \mathcal{M}_n \times \mathcal{M}_n$$

bestehend aus zwei molekularen Graphen  $M = (\varepsilon, \zeta, \gamma)$  und  $M' = (\varepsilon', \zeta', \gamma')$  heißt *chemische Reaktion*, wenn  $\varepsilon = \varepsilon'$ .  $M$  heißt *Eduktgraph* und  $M'$  *Produktgraph*. Die Elemente von  $\text{Con}(M)$  sind *Edukte* oder synonym *Reaktanden*, die von  $\text{Con}(M')$  *Produkte*.  $\mathcal{C}_n$  bezeichne im Folgenden die Menge der chemischen Reaktionen auf  $n$  Atomen.

Ist  $|\text{Con}(M)| = 1$  so bezeichnet man  $C$  als *unimolekulare* Reaktion oder *Einkomponentenreaktion*. Ist zusätzlich  $|\text{Con}(M')| = 1$  nennt man  $C$  *Umlagerungsreaktion*, falls  $|\text{Con}(M')| > 1$  spricht man von einer *Zerfallsreaktion*. Eine Reaktion mit  $|\text{Con}(M)| = 2$  bezeichnet man als *bimolekular* oder *Zweikomponentenreaktion*. Reaktionen mit  $|\text{Con}(M)| \geq 2$  nennen wir auch *Synthesereaktionen*.

### 1.5.2 Bemerkung:

Von besonderem Interesse sind die durch die chemische Reaktion  $C$  entstandenen Veränderungen der Atomzustände und Bindungen. Dazu betrachtet man den *Reaktionsänderungsgraph*

$$\Delta C := (\Delta\zeta, \Delta\gamma) \in \Delta\mathcal{Z}^n \times \mathcal{G}_{n,[-3,3]} =: \Delta\mathcal{C}_n,$$

wobei für  $i \in n$

$$\Delta\zeta(i) := (\Delta v_i, \Delta p_i, \Delta q_i, \Delta r_i) \in \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \times \mathbb{B} =: \Delta\mathcal{Z}$$

die *Zustandsänderung* von Atom  $i$  angibt. Dabei beschreibt

- $\Delta v_i := v_{\zeta'(i)} - v_{\zeta(i)}$  die Änderung der Valenz,
- $\Delta p_i := p_{\zeta'(i)} - p_{\zeta(i)}$  die Änderung der Anzahl freier Elektronenpaare,
- $\Delta q_i := q_{\zeta'(i)} - q_{\zeta(i)}$  die Ladungsänderung und

- $\Delta r_i := r_{\zeta'(i)} \dot{\vee} r_{\zeta(i)}$  die Änderung des Radikalcharakters an Atom  $i$ , wobei „ $\dot{\vee}$ “ die Verknüpfung zweier boolescher Ausdrücke durch „ausschließendes oder“ bezeichnet.

$\Delta\zeta$  ist die *Zustandsänderungsverteilung* von  $C$ . Weiterhin gibt

$$\Delta\gamma \in \mathcal{G}_{n,[-3,3]} := [-3, 3]^{\binom{n}{2}}$$

für  $i, j \in \underline{n}$ ,  $i \neq j$  mit

$$\Delta\gamma(\{i, j\}) := \gamma'(\{i, j\}) - \gamma(\{i, j\}) \in [-3, 3]$$

die *Bindungsänderung* zwischen Atom  $i$  und  $j$  an.  $\Delta\gamma$  ist der *Bindungsänderungsgraph* von  $C$ . Eine chemische Reaktion  $C$  wird durch ihren Eduktgraphen  $M$  und ihren Reaktionsänderungsgraphen  $\Delta C$  vollständig beschrieben. Wir bezeichnen deshalb den 5-Tupel  $(\varepsilon, \zeta, \gamma, \Delta\zeta, \Delta\gamma)$  als *Reaktionsgraph*. Mit den Bezeichnungen aus Definition 1.5.1 schreiben wir:

$$M' = \Delta C \circ M,$$

wobei

$$\Delta C \circ M = (\Delta\zeta, \Delta\gamma) \circ (\varepsilon, \zeta, \gamma) := (\varepsilon, \Delta\zeta \circ \zeta, \Delta\gamma \circ \gamma)$$

und für  $i, j \in \underline{n}$ ,  $i \neq j$

$$\begin{aligned} (\Delta\zeta \circ \zeta)(i) &:= \Delta\zeta(i) \circ \zeta(i) \\ (\Delta\gamma \circ \gamma)(\{i, j\}) &:= \gamma(\{i, j\}) + \Delta\gamma(\{i, j\}). \end{aligned}$$

Die Zustandsverteilung ist dabei

$$\Delta\zeta(i) \circ \zeta(i) := (v_{\zeta(i)} + \Delta v_i, p_{\zeta(i)} + \Delta p_i, q_{\zeta(i)} + \Delta q_i, r_{\zeta(i)} \dot{\vee} \Delta r_i).$$

### 1.5.3 Definition:

Sei  $C = ((\varepsilon, \zeta, \gamma), (\varepsilon, \zeta', \gamma')) \in \mathcal{C}_n$  eine chemische Reaktion. Dann ist

$$\text{Cen}(C) := \{i \in \underline{n} \mid \zeta(i) \neq \zeta'(i) \vee \exists j : \gamma(\{i, j\}) \neq \gamma'(\{i, j\})\}$$

das *Reaktionszentrum* von  $C$ .

### 1.5.4 Bemerkung:

Im Reaktionszentrum liegen also genau die Atome, deren Zustand sich bei der Durchführung der Reaktion ändert, oder deren inzidente Bindungen ihre Vielfachheit ändern bzw. neu gebildet oder gebrochen werden. Man kann eine chemische Reaktion auch durch Angabe des Eduktgraphen, des Reaktionszentrums und der Veränderung von Atomzuständen und Bindungen der im Reaktionszentrum gelegenen Atome beschreiben.

**1.5.5 Definition:**

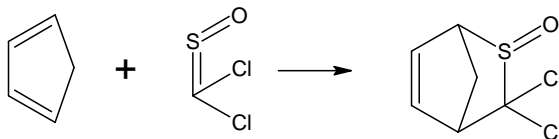
Sei  $C$  eine chemische Reaktion. Der auf dem Reaktionsgraphen durch das Reaktionszentrum induzierte Teilgraph

$$\text{RCG}(C) := (\varepsilon|_{\text{Cen}(C)}, \zeta|_{\text{Cen}(C)}, \gamma|_{\text{Cen}(C)}, \Delta\zeta|_{\text{Cen}(C)}, \Delta\gamma|_{\text{Cen}(C)})$$

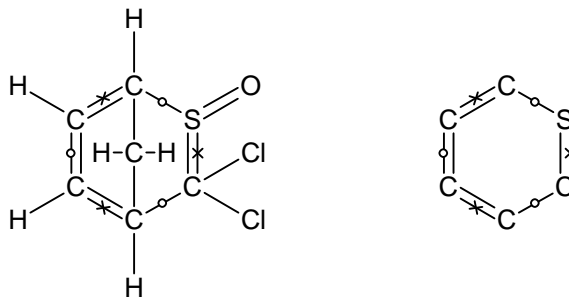
heißt *Reaktionszentrumsgraph* von  $C$ .

**1.5.6 Beispiel:**

Folgende Abbildung zeigt die Diels–Alder–Reaktion nach [140]. Auf der linken Seite des Reaktionspfeils stehen dabei die Edukte, rechts die Produkte:



Nachfolgend ist links der Reaktionsgraph und rechts der Reaktionszentrumsgraph dargestellt:



Neu geschlossene Bindungen sind dabei durch Kreise, gebrochene Bindungen durch Kreuze ausgezeichnet.

**1.5.7 Bemerkung:**

In der Chemie beobachtet man oft, dass verschiedene Reaktionen nach gleichen oder ähnlichen Schemata ablaufen. Dies manifestiert sich, indem ihre Reaktionszentrumsgraphen „ähnlich“ sind. Diese „Ähnlichkeit“ wollen wir nutzen, um eine Datenstruktur zu erklären, die zum einen „ähnliche“ Reaktionszentrumsgraphen repräsentiert und zum anderen ermöglicht, bei einem gegebenen Eduktgraph die chemische Reaktion, aus der der Reaktionszentrumsgraph gewonnen wurde, zu rekonstruieren.

**1.5.8 Definition:**

Sei  $k \in \mathbb{N}^*$ . Ein *Reaktionsschema* ist ein Tripel

$$R := (S, \Delta\zeta, \Delta\gamma) \in \mathcal{S}_k \times \Delta\mathcal{Z}^k \times \mathcal{G}_{k,[-3,3]} := \mathcal{R}_k$$

bestehend aus der *Reaktionssubstruktur*  $S$ , der *Zustandsänderungsverteilung*  $\Delta\zeta$  und dem *Bindungsänderungsgraph*  $\Delta\gamma$ .

Je nach Anzahl der Zusammenhangskomponenten des  $S$  zugrunde liegenden MMG bezeichnen wir  $R$  als *unimolekular*, *bimolekular* u.s.w.

**1.5.9 Bemerkung:**

Diese Definition weicht ab von [164]. Dafür sind nunmehr auch Zerfalls- und Umlagerungsreaktionen sowie Synthesereaktionen von mehr als zwei Komponenten darstellbar.

Die Anwendung eines Reaktionsschemas  $R = (S, \Delta\zeta, \Delta\gamma) \in \mathcal{R}_k$  auf einen molekularen Graphen  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$  erfolgt in zwei Schritten. Zuerst wird eine Einbettung der Reaktionssubstruktur  $S$  in  $M$  als molekulare Teilstruktur gesucht. Ist eine solche Einbettung  $\phi \in \text{Emb}_{\subseteq}^i(S, M)$  gefunden, werden Zustandsänderungsverteilung und Bindungsänderungsgraph folgendermaßen auf  $M$  angewendet:  $\phi$  induziert eine Abbildung

$$\phi : \Delta\mathcal{C}_k \longrightarrow \Delta\mathcal{C}_n, \quad (\Delta\zeta, \Delta\gamma) \longmapsto (\Delta\zeta, \Delta\gamma)^\phi := (\Delta\zeta^\phi, \Delta\gamma^\phi),$$

wobei für  $i \in n$

$$\Delta\zeta^\phi := \begin{cases} \Delta\zeta(\phi^{-1}(i)) & \text{falls } i \in \phi(k), \\ (0, 0, 0, \text{falsch}) & \text{sonst,} \end{cases}$$

und für  $i, j \in n, i \neq j$

$$\Delta\gamma^\phi := \begin{cases} \Delta\gamma(\{\phi^{-1}(i), \phi^{-1}(j)\}) & \text{falls } i, j \in \phi(k), \\ 0 & \text{sonst.} \end{cases}$$

Die Anwendung von  $R$  auf  $M$  bzgl.  $\phi$  können wir nun vermöge  $(\Delta\zeta, \Delta\gamma)^\phi$  definieren als

$$R \circ_\phi M := (\Delta\zeta, \Delta\gamma) \circ_\phi M := (\Delta\zeta, \Delta\gamma)^\phi \circ M.$$

Dabei ist jedoch nicht sichergestellt, dass es sich bei  $M \circ_\phi R$  um einen molekularen Graphen handelt.

**1.5.10 Definition:**

Seien  $k, n \in \mathbb{N}^*$ ,  $k \leq n$ ,  $R = (S, \Delta\zeta, \Delta\gamma) \in \mathcal{R}_k$  ein Reaktionsschema und  $M \in \mathcal{M}_n$  ein molekularer Graph. Die Menge der Produktgraphen bei der Anwendung von  $R$  auf  $M$  ist

$$\text{Prod}_R(M) := \{R \circ_\phi M \in \mathcal{M}_n \mid \phi \in \text{Emb}_{\subseteq^i}(S, M)\}.$$

**1.5.11 Bemerkung:**

Das beschriebene Modell erlaubt, chemische Reaktionen qualitativ zu simulieren, d.h. zu gegebenen Edukten und Reaktionsschemata *alle möglichen* Produkte zu bestimmen. Ein weiterer wichtiger Aspekt ist der quantitative Verlauf chemischer Reaktionen, d.h. in welchen Konzentrationen die Produkte auftreten. Dies ist weit schwieriger zu modellieren und die Modelle sind oft nur auf eng begrenzte Substanzklassen anwendbar. So wurde beispielsweise in [62] und [130] versucht, Reaktionen im Massenspektrometer quantitativ vorherzusagen. Dabei kommen Methoden des maschinellen Lernens zum Einsatz. [68] erweitert die Möglichkeiten der Reaktionsvorhersage im Rahmen des Programmsystems *EROS 7*. Allerdings bleibt die exakte Bestimmung von Reaktivitäten ein nur mit hohem Aufwand durch experimentelle Messungen oder quantenchemische Berechnungen lösbares Problem. Interessant ist dahingehend ein neuerer Ansatz [15, 16], der versucht, die notwendigen Energieberechnungen mit einem stark vereinfachten Modell zu erstellen.



## 1.6 Erweiterungen des Molekülmodells

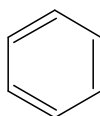
Es gibt in der Chemie Phänomene, die durch unser graphentheoretisches Modell zunächst nicht beschrieben werden können. Für die organische Chemie sind dabei Mesomerie und die geometrische Struktur von Molekülen von besonderem Interesse.

### 1.6.1 Mesomerie

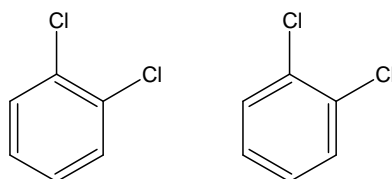
Bei aromatischen Strukturen ist es nicht mehr möglich, jede Bindung eindeutig durch eine ganzzahlige Kantenvielfachheit zu beschreiben. Stattdessen können mehrere *mesomere Grenzstrukturen* angegeben werden, um aromatische Verbindungen darzustellen. Allerdings sind diese generell nicht isomorph im Sinne von Definition 1.3.12.

#### 1.6.1 Beispiele:

Ein sehr einfaches Beispiel einer aromatischen Verbindung ist Benzol:



Bereits bei zweifach substituierten Benzol-Derivaten können nichtisomorphe, aber chemisch äquivalente Darstellungen angegeben werden, wie man am Beispiel von 1,2-Dichlorbenzol sieht:



#### 1.6.2 Bemerkung:

Im Allgemeinen besteht das Interesse, Strukturräume so irredundant wie möglich darzustellen. Dazu ist es insbesondere notwendig, mesomere Doppelten z.B. in der Ausgabe eines Strukturgenerators erkennen und filtern zu können. Des Weiteren ist es bei der Suche nach Struktur-Eigenschafts-Beziehungen (Abschnitt 4.4) wichtig, chemische Verbindungen so genau wie möglich zu beschreiben. Es gibt molekulare Deskriptoren, die Informationen über die Aromatizität von Bindungen berücksichtigen.

Die mathematischen Aspekte der Erkennung und Filterung mesomerer Doppelten wurden eingehend in [41] behandelt. Problematisch ist es aber immer

wieder, den in der Chemie rein phänomenologisch definierten Begriff der Aromatizität in graphische Regeln zu fassen.

Bereits in *MOLGEN 3.5* war ein Filter für aromatische Dubletten integriert. Dieser konnte jedoch nur aromatische Ringe bestehend aus sechs C-Atomen finden. Inzwischen steht ein Algorithmus zur Verfügung, welcher aromatische Systeme variabler Größe mit Heteroatomen und/oder Ladungen erkennt (vgl. auch [54], Abschnitt 9.8).

### 1.6.3 Definition:

Sei  $M \in \mathcal{M}$  ein molekularer Graph. Ein Kreis  $W$  mit Länge  $\text{len}_M(W) > 3$  heißt *aromatisch*, wenn

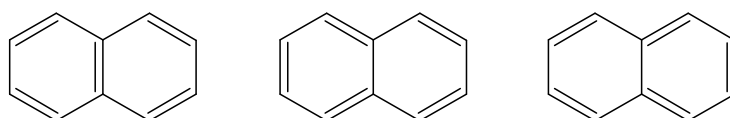
- (i) die Anzahl der *zyklisch delokalisierten  $\pi$ -Elektronen* auf  $W$   $4k + 2$  mit  $k \in \mathbb{N}^*$  beträgt. Dazu zählen
  - (a) pro Mehrfachbindung 2 Elektronen und
  - (b) pro Atom mit einem einsamen Elektronenpaaren 2 Elektronen,
- (ii) jede Mehrfachbindung mit genau zwei Einfachbindungen in  $W$  inzident ist, und
- (iii) jede Einfachbindung entweder
  - (a) mit genau zwei Mehrfachbindungen in  $W$  inzident ist oder
  - (b) mit genau einer Mehrfachbindung in  $W$  und am anderen Ende zu einem Atom mit einsamen Elektronenpaaren inzidiert oder
  - (c) mit genau einer Mehrfachbindung in  $W$  und am anderen Ende zu einem Atom mit Ladung inzidiert.

### 1.6.4 Bemerkung:

Kreise, die obige Bedingungen erfüllen, werden auch als *aromatische Ringe* bezeichnet. Bei dieser Nomenklatur ist jedoch zu beachten, dass es sich nicht um Ringe im Sinne von Definition 1.2.11 handeln muss. Bindungen auf solchen Kreisen nennt man *aromatische Bindungen*.

**1.6.5 Beispiel:**

Nach Definition 1.6.3 besitzt Naphthalin 11 aromatische Bindungen, wie man an den drei mesomeren Darstellungen überprüfen kann:



Die linke Struktur enthält zwei aromatische Ringe der Länge 6, die beiden rechten je einen der Länge 6 und einen der Länge 10. Die beiden rechten Strukturen sind isomorph im Sinne von Definition 1.3.12. Nach Markierung der aromatischen Bindungen werden alle drei Strukturen als isomorph erkannt.

**1.6.6 Algorithmus:**

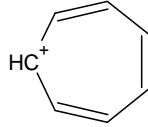
In einem molekularen Graphen  $M$  werden aromatische Bindungen folgendermaßen identifiziert:

- (i) Durchlaufe nach *depth-first* Methode alle Wege  $W$  in  $M$ , die Bedingungen (ii) und (iii) aus Definition 1.6.3 genügen.
- (ii) Ist  $W$  ein Kreis, der zusätzlich Bedingung (i) aus Definition 1.6.3 erfüllt, so markiere alle Kanten in  $W$  als aromatisch.
- (iii) Entferne Ladungen an Atomen, die nach Bedingung (iii, c) aus Definition 1.6.3 an dem aromatischen Ring beteiligt sind.

**1.6.7 Bemerkung:**

Da wir nun wissen, wie aromatische Bindungen zu finden sind, bleibt zu überlegen, wie diese kodiert werden sollen. Nahe liegend wäre zunächst, bei den Bindungsvielfachheiten auch rationale Zahlen zuzulassen. So könnte man beispielsweise im Benzol die aromatischen Bindungen mit Vielfachheit  $\frac{3}{2}$  kodieren. Jedoch stößt dieses Modell schon beim Naphthalin an seine Grenzen: Würde man hier alle aromatischen Bindungen mit Vielfachheit  $\frac{3}{2}$  versehen, so ergäbe dies für die beiden zentralen C-Atome eine Valenz von  $\frac{9}{2} \notin \mathbb{N}$ . Um dies zu vermeiden, könnte man Bindungsvielfachheiten von  $\frac{4}{3}$  und  $\frac{5}{3}$  zuordnen. Allerdings wäre man dann mit einem weiteren Problem konfrontiert, nämlich der Berechnung passender Bindungsvielfachheiten. In der gegenwärtigen Implementation werden aromatische Bindungen durch einen nicht numerisch interpretierbaren Bindungstyp „aromatisch“ dargestellt. Eine befriedigende Lösung gibt es auf Basis von Multigraphen nicht.

Ein weiteres Problem tritt auf, wenn Ladungen an dem aromatischen System beteiligt sind. So ist beispielsweise das Tropylium-Ion



aromatisch. Nach Markierung der aromatischen Bindungen muss konsequenterweise auch die positive Ladung an dem C-Atom entfernt werden. Insgesamt bleibt die Struktur jedoch einfach positiv geladen. Die positive Ladung ist auf das gesamte aromatische System verteilt. Auch diese Konstellation kann durch unser bisheriges Modell nicht dargestellt werden.

Ein alternativer Ansatz wäre ein Molekülmodell, welches die Zuordnung von Elektronen zu Atomen widerspiegelt und in Form von Multi-Hypergraphen darstellt:

Sei  $m \in \mathbb{N}^*$ ,  $m \geq 2$  eine obere Grenze für die Anzahl von Elektronen in einem aromatischen System. Eine chemische Verbindung mit  $n$  Atomen kann dann dargestellt werden durch einen Multi-Hypergraphen

$$\varphi \in m^{\mathcal{P}^*(n)}.$$

Für  $V \in \mathcal{P}^*(n)$  entspricht  $\varphi(V)$  der Anzahl von Elektronen, die sich die Atome aus  $V$  „teilen“. Wie in Abschnitt 1.3 müsste eine Verteilungsfunktion für die Elemente zur Beschreibung eines Moleküls hinzugezogen werden. Da in  $\varphi$  auch die freien Elektronenpaare und einsamen Elektronen kodiert sind, kann über die Anzahl der Valenzelektronen des chemischen Elements die Ladung am Atom berechnet, und somit auf eine Verteilungsfunktion für Atomzustände verzichtet werden.  $\varphi \in m^{\mathcal{P}^*(n)}$  würde sich dann wie folgt aus  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$  berechnen:

$$\begin{aligned} \varphi(\{i\}) &= 2p_{\zeta(i)} + r_{\zeta(i)}, \\ \varphi(\{i, j\}) &= \begin{cases} 2 & \text{falls } i \text{ und } j \text{ aromatisch gebunden,} \\ 2\gamma(\{i, j\}) & \text{sonst} \end{cases} \\ \varphi(\{i_1, \dots, i_k\}) &= \begin{cases} 0 & \text{falls } \{i_1, \dots, i_k\} \text{ kein aromatisches System ist,} \\ \text{sonst :} & \text{Anzahl von Elektronen, die } i_1, \dots, i_k \\ & \text{an das aromatische System abgeben,} \end{cases} \end{aligned}$$

wobei  $i \in n$ ,  $\{i, j\} \subseteq n$  und  $\{i_1, \dots, i_k\} \subseteq n$  mit  $k \geq 3$ . Um die graphentheoretischen Begriffe aus Abschnitt 1.2 auf dieses Modell anwenden zu können, muss man die Projektion  $\frac{1}{2}\varphi|_{\binom{n}{2}}$  auf die zweielementigen Mengen von  $\mathcal{P}^*(n)$  betrachten.

Ein ähnliches Modell zur Repräsentation von Molekülen wird in [80] und [81] beschrieben. In [11] wird zwischen  $\sigma$ - und  $\pi$ -Elektronensystemen unterschieden. Unter anderem ist dort auch die Darstellung von delokalisierten  $\pi$ -Elektronen vorgesehen, wodurch aromatische Systeme dargestellt werden. Eine sehr umfassende Repräsentation für Verbindungen mit nicht kovalenten Bindungen zeigt [49], wo zudem auch eine Darstellung konfigurativer Aspekte, d.h. der Stereochemie ermöglicht wird.

### 1.6.2 Geometrie

Bislang haben wir vernachlässigt, dass Moleküle auch als Objekte im dreidimensionalen Raum zu verstehen sind. Man kann dieser Tatsache Rechnung tragen, indem man  $M \in \mathcal{M}_n$  mit 3D-Koordinaten

$$\xi \in (\mathbb{R}^3)^n$$

versieht.  $\xi$  ordnet dabei jedem Atom einen Punkt aus  $\mathbb{R}^3$  zu. Eine solche *3D-Platzierung* von  $M$  kann beispielsweise mit Hilfe des MM2-Kraftfeld-Modells nach [1] berechnet werden. Dabei wird eine gegebene Startplatzierung bzgl. einer Energiefunktion, in die Abstände, Winkel und Torsionswinkel zwischen Atomen eingehen, optimiert. Die Optimierung erfolgt mittels einem modifizierten Gauß-Newton-Verfahren. Für die Minima der Energiefunktion erhält man in vielen Fällen eine plausible 3D-Platzierung des Moleküls. Mitunter terminiert der Optimierungsalgorithmus allerdings auch in lokalen Minima, die chemisch irrelevanten Platzierungen entsprechen. Das Ergebnis der Energieoptimierung ist abhängig von der Startplatzierung für das Verfahren. Die Wahl der Startplatzierung kann nach verschiedenen Strategien erfolgen:

- (i) Zufallsplatzierung,
- (ii) 2D-Koordinaten und Zufallswerte für die dritte Koordinate,
- (iii) auf eine Kugeloberfläche projizierte *planare Platzierung*.

Variante (i) ist vor allem dann sinnvoll, wenn in mehreren Optimierungsläufen verschiedene Platzierungen berechnet werden sollen. Dies kann z.B. im Rahmen einer Konformationsanalyse [12, 17] erwünscht sein. Allerdings ist im Vergleich zu den beiden anderen Methoden die Wahrscheinlichkeit, unrealistische 3D-Platzierungen zu finden, relativ hoch.

Variante (ii) wird in der aktuellen Version von *MOLGEN* benutzt. Die *2D-Platzierung* wird von einem Algorithmus [133] berechnet, der versucht chemische Strukturen ausgehend von ihren Ringsystemen bestmöglich verteilt auf einer rechteckigen Grundfläche darzustellen.

Wie [120] zeigt, sind die meisten organischen Verbindungen graphentheoretisch planar. Man kann für solche Strukturen eine planare Platzierung in der Ebene berechnen, und diese auf eine Kugeloberfläche projizieren. Inwiefern diese Strategie gute Startplatzierungen für die Energieoptimierung liefert, bleibt zu untersuchen.

## 1.7 Molekulare Graphen und existente chemische Verbindungen

Zwar ist die Vielfalt real existierender chemischer Verbindungen erstaunlich groß, jedoch macht sie nur einen Bruchteil der mathematisch möglichen molekularen Graphen aus. Dies liegt in erster Linie daran, dass viele mathematisch darstellbare molekulare Graphen als chemische Verbindungen energetisch zu instabil wären.

Man kann diesen Bruchteil unter bestimmten Einschränkungen quantitativ angeben. Es gibt eine Datenbank<sup>2</sup>, in der ein Großteil aller real existierenden Verbindungen der organischen Chemie, d.h. Naturstoffe und künstlich synthetisierte Kohlenstoffverbindungen erfasst ist. Diese Datenbank umfasst derzeit 8711107 Einträge. Daraus wurden alle zusammenhängenden Strukturen extrahiert, die sich nur aus Elementen in  $\mathcal{E}_4$  zusammensetzen und eine Molekülmasse von 150 amu nicht überschreiten: Derer fanden sich 174290. Weiterhin wurden davon solche entfernt, bei denen Atome auftraten, deren Masse vom häufigsten Isotop abwich. Ebenso wurden all diejenigen verworfen, die Atome mit Ladungen oder Radikalstellen aufwiesen, oder deren Valenz größer als die Standardvalenz  $v_X$  war. Zudem wurde berücksichtigt, dass in der Datenbank auch Stereoisomere abgelegt sind, d.h. chemische Verbindungen, die durch den gleichen molekularen Graphen dargestellt werden, aber sich durch ihre geometrische Gestalt unterscheiden. Nach kanonischer Nummerierung und Filterung von Dubletten verblieben 103040. Diese wurden schließlich nach ihren Bruttoformeln sortiert.

In Anhang E wird für jede Bruttoformel mit Elementen aus  $\mathcal{E}_4$  mit Molekülmasse  $\leq 150$  aufgelistet, wie viele Konstitutionsisomere theoretisch (mathematisch) und wie viele real (in der *Beilstein*-Datenbank) existieren. Es wird nicht verwundern, dass nur ein verschwindend geringer Anteil als real existent nachgewiesen ist. Beispielsweise sind von den 217 Konstitutionsisomeren zur Bruttoformel  $C_6H_6$  nur 29 in der *Beilstein*-Datenbank erfasst (Abbildung 1.1).

Eine mögliche Erklärung ist die mangelnde chemische Stabilität vieler theoretisch denkbarer Konstitutionsisomere. Dieses Argument kann durch die *sterische Energie* (kurz *SE*) in Zahlen gefasst werden. Die sterische Energie bildet die Zielgröße in dem oben angesprochenen Optimierungsproblem zur Berechnung einer dreidimensionalen Platzierung.

Wir führen dazu folgendes Experiment durch: Auf jedes der 217 Konstitutionsisomere von  $C_6H_6$  wurde der Optimierungsalgorithmus 10 mal mit ver-

---

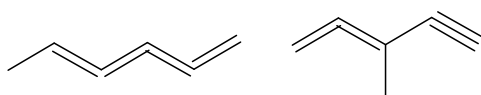
<sup>2</sup>Beilstein Datenbank BS0302PR mit MDL CrossFire Commander Server-Software, Version 6.0, MDL Information Systems GmbH

SE:43,298 1	SE:43,819 2	SE:44,289 3	SE:47,358 4	SE:48,222 5	SE:48,313 6
SE:48,383 7	SE:49,140 8	SE:50,916 9	SE:50,931 10	SE:51,539 11	SE:52,169 12
SE:53,003 13	SE:68,791 14	SE:79,335 15	SE:103,073 16	SE:116,170 17	SE:118,654 18
SE:149,494 19	SE:151,925 20	SE:189,271 21	SE:220,191 22	SE:222,097 23	SE:252,305 24
SE:253,675 25	SE:312,426 26	SE:328,932 27	SE:329,048 28	SE:332,724 29	

Abbildung 1.1: Konstitutionsisomere von  $C_6H_6$  in der *Beilstein*-Datenbank zusammen mit den berechneten Werten für die sterische Energie

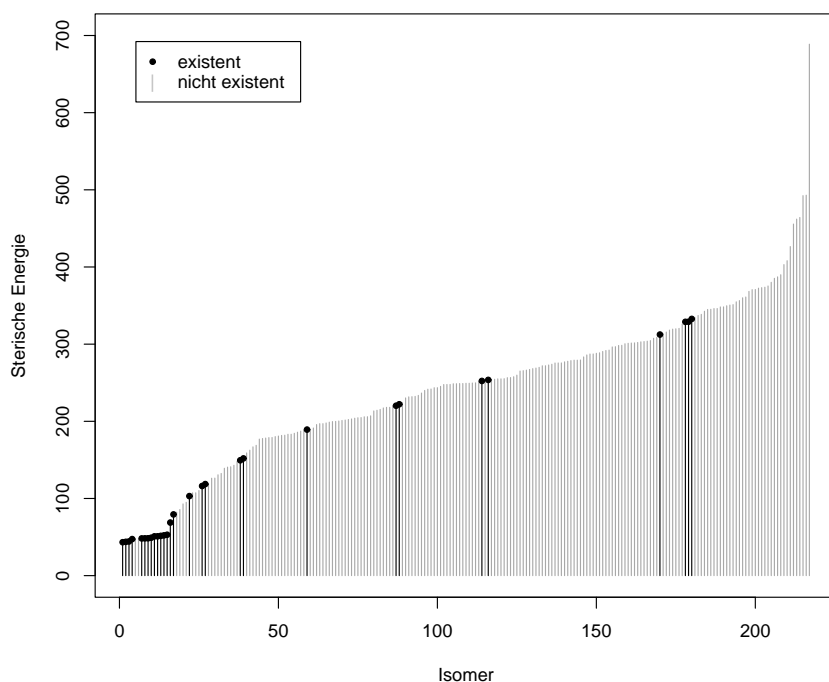
schiedenen Startplatzierungen angewendet. Die jeweils kleinsten Werte sind für die  $C_6H_6$ -Isomere der *Beilstein*-Datenbank in Abbildung 1.1 eingetragen und sind für alle 217  $C_6H_6$ -Isomere in Abbildung 1.2 visualisiert: Graue Linien stehen für Energiewerte nicht existenter Isomere, solche real existierender Verbindungen sind schwarz dargestellt.

Man sieht deutlich, dass real existenten Isomeren eher kleinere Energiewerte zugeordnet sind. Nicht in der *Beilstein*-Datenbank vertretene Strukturen mit kleinsten Energiewerten sind (von links) Hexa-1,2,3,4-tetraen (47,12 kcal/mol) und 3-Methylpenta-1,2-dien-4-in (48,15 kcal/mol):

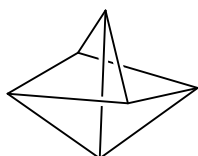


Die größten Energiewerte unter den in der *Beilstein*-Datenbank enthaltenen  $C_6H_6$ -Isomeren haben Bi(cycloprop-1-en-1-yl) (332,72 kcal/mol), Prisma (329,05 kcal/mol) und Bi(cycloprop-2-en-1-yl) (328,93 kcal/mol). Nicht überraschend ist, dass das  $C_6H_6$ -Isomer, dessen Kohlenstoffgerüst dem Gra-



Abbildung 1.2: Sterische Energie der Konstitutionsisomere von  $C_6H_6$ 

phen  $K_{3,3}$  entspricht, maximale sterische Energie (688,76 kcal/mol) besitzt:



Bei genauerer Betrachtung der Strukturen und ihrer Energiewerte erkennt man, dass die 15 azyklischen  $C_6H_6$ -Isomere gerade die 15 kleinsten Energiewerte annehmen. Den kleinsten Energiewert unter den zyklischen  $C_6H_6$ -Isomeren hat Benzol (86,79 kcal/mol).

Ausgehend von einer räumlichen Platzierung kann man auch einen Wert für das Volumen eines Moleküls, das so genannte *Van der Waals Volumen* berechnen. Dabei wird jedes Atom als Kugel mit einem so genannten Van der Waals Radius betrachtet. Die 3D-Koordinaten der Atome stellen die Kugelmittelpunkte dar. Details zur Volumenberechnung folgen in Beispiel 4.3.6. Je nachdem, wie groß die Überschneidungen der Van der Waals Kugeln sind, ergeben sich verschiedene Molekülvolumina für die 217  $C_6H_6$ -Isomere. Diese sind in Abbildung 1.3 visualisiert. Wiederum sind die Volumina nach *Beilstein* real existenter Isomere durch schwarze Linien hervorgehoben. Man kann sehen, dass diese eher im oberen Bereich der Volumina aller mathematisch

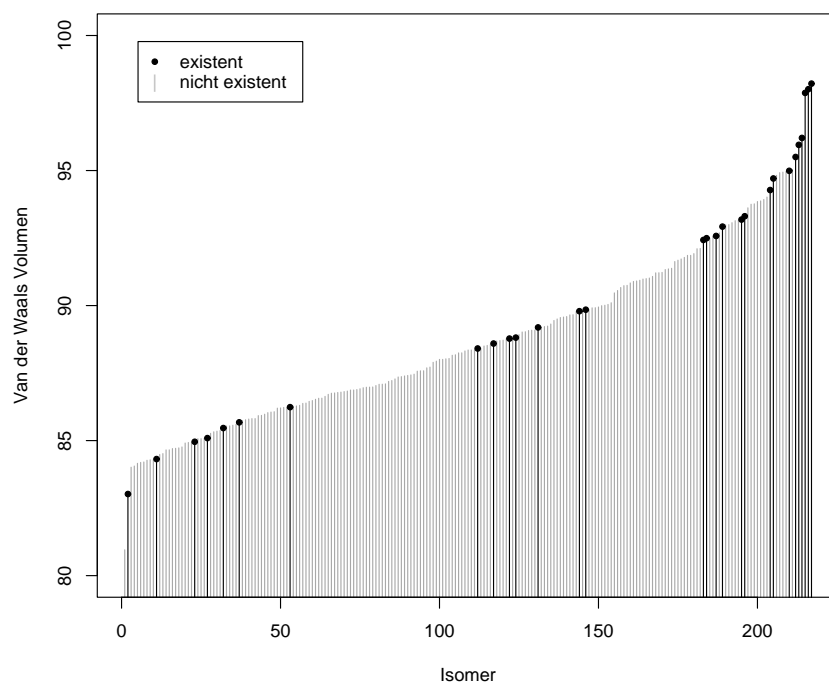
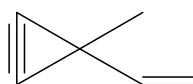


Abbildung 1.3: Van der Waals Volumen der Konstitutionsisomere von  $C_6H_6$

möglichen  $C_6H_6$ -Isomere liegen. Die nicht in der *Beilstein*-Datenbank vertretene Struktur mit größtem Volumen ist 3-Methyl-3-vinylcyclopropin ( $95,039 \text{ \AA}^3$ ):

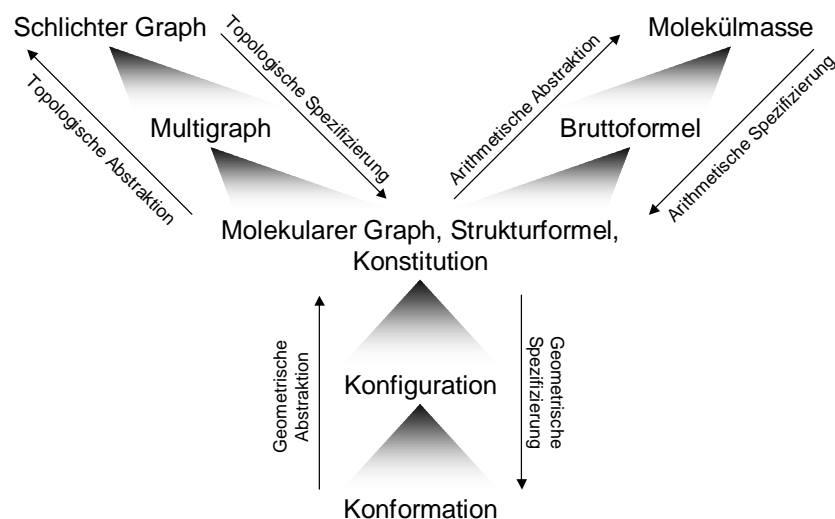


Kleinstes Volumen besitzt das Isomer mit dem Kohlenstoffgerüst des  $K_{3,3}$  ( $80,957 \text{ \AA}^3$ ), gefolgt von Prisman ( $83,022 \text{ \AA}^3$ ).

Existenz oder Nichtexistenz ist sicherlich eine der grundlegendsten Eigenschaften einer chemischen Verbindung. Allerdings kann sich diese Eigenschaft ändern, wenn eine zuvor nicht existent geglaubte Verbindung entweder synthetisiert, oder in der Natur nachgewiesen wird. In Kapiteln 4 und 5 werden wir uns mit anderen, weniger variablen Eigenschaften chemischer Verbindungen beschäftigen. Dabei werden wir versuchen, aus den chemischen Strukturen auf ihre Eigenschaften, oder umgekehrt von gegebenen Eigenschaften auf die Strukturen zu schließen. Eine wichtige Rolle bei diesen Betrachtung wird die Fähigkeit spielen, molekulare Strukturen nach bestimmten Vorschriften generieren, d.h. „virtuell synthetisieren“ zu können. Die wichtigsten Werkzeuge dazu werden wir in Kapitel 2 kennen lernen.

## 1.8 Zusammenfassung und Ausblick: Abstraktionsebenen des Molekülmodells

An dieser Stelle wollen wir kurz innehalten und auf einige Abstraktionsebenen unseres Molekülmodells zurück- und vorausblicken. Von zentraler Bedeutung in dieser Arbeit ist die Modellierung chemischer Verbindungen als *molekulare Graphen*. Ausgehend von den hier synonymen Bezeichnungen *Strukturformel* und *Konstitution*, wollen wir verschiedene Richtungen der Abstraktion bzw. Spezifizierung in der Modellierung chemischer Verbindungen betrachten:



In den Abschnitten 1.2 und 1.3 haben wir folgende *topologische Abstraktion* molekularer Graphen kennen gelernt. Vernachlässigt man in einem molekularen Graphen die Unterscheidung der Atome durch verschiedene Elemente und Atomzustände erhält man einen *Multigraphen*. Verzichtet man zudem auf die Kantenvielfachheiten des Multigraphen gelangt man schließlich zu einem *schlichten Graphen*.

Eine zweite Möglichkeit der Abstraktion eröffnet sich, wenn man in einer Strukturformel die Wechselwirkungen zwischen den Atomen außer Acht lässt und sich lediglich auf die Elemente der Atome und ihre Vielfachheiten beschränkt. Man erhält die *Bruttoformel*. Eine Möglichkeit diese *arithmetische Abstraktion* fortzuführen besteht darin, sich auf die Summe der Atommassen, die *Molekülmasse*, zurückzuziehen. Dieser Weg der Abstraktion wird in Kapitel 5 eine wichtige Rolle spielen: In der molekularen Strukturaufklärung beschreitet man diesen Weg oft in umgekehrter Richtung, indem man ausgehend von einer gemessenen Molekülmasse mögliche Bruttoformeln und anschließend plausible Strukturformeln sucht.

Die dritte in der obigen Abbildung dargestellte Richtung weist auf die *geometrische Spezifizierung* chemischer Verbindungen hin. Während die *Konstitution* noch keine Information über die geometrische Molekülgestalt enthält, gibt die *Konfiguration* bereits Auskunft über die relative räumliche Lage der Atome zueinander. Im Gegensatz zur *Konformation* unterscheidet die *Konfiguration* aber keine Anordnungen, die durch Drehungen um Einfachbindungen ineinander übergeführt werden können. Geometrische Aspekte spielen in dieser Arbeit nur eine untergeordnete Rolle, werden aber mit Gewissheit in der zukünftigen Forschung und Entwicklung von zentraler Bedeutung sein, insbesondere da viele biologisch–pharmazeutische Eigenschaften chemischer Verbindungen stark von ihrer geometrischen Struktur abhängen.

# Kapitel 2

## Molekulare Strukturgenerierung

In Abschnitt 1.8 haben wir verschiedene Abstraktionsebenen des Molekülmodells kennen gelernt. Während die Transformation einer Strukturrepräsentation in Richtung zunehmender Abstraktion meist trivial ist, können die Schritte zwischen den einzelnen Ebenen in Richtung zunehmender Spezifizierung enorme algorithmische Probleme aufwerfen. Mit dieser Problematik befasst sich u.a. die molekulare Strukturgenerierung. Die wichtigsten Anforderungen sind dabei *Vollständigkeit* unter den gegebenen Nebenbedingungen und *Redundanzfreiheit* hinsichtlich Isomorphie.

Beispielsweise können mit Hilfe des *Homomorphieprinzips* [84] ausgehend von einem schlichten Graphen alle zugehörigen Multigraphen und in einem weiteren Schritt alle molekularen Graphen generiert werden.

Im Rahmen dieser Arbeit wird die *bruttoformelbasierte* Strukturgenerierung eine wichtige Rolle spielen. Hierbei bildet die elementare Zusammensetzung der zu generierenden Strukturformeln den Ausgangspunkt der Strukturgenerierung. Im einfachsten Fall ist dies *eine* Bruttoformel. *Restriktionen* dienen zur Begrenzung des *Strukturraums*. Insbesondere können molekulare Substrukturen als Restriktionen herangezogen werden.

Eine weitere Problemstellung befasst sich mit der Transformation molekularer Graphen, wie sie durch chemische Reaktionen beschrieben werden. *Reaktionsbasierte* Strukturgenerierungsprobleme werden im Wesentlichen durch Reaktanden und Reaktionsschemata spezifiziert.

Eine dritte Möglichkeit zur Definition von Strukturräumen bilden *generische Strukturformeln* durch die Verwendung von variablen Teilstrukturen, so genannten *R-Gruppen*. Dabei sind verschiedene Formen der Variation erlaubt, beispielsweise Variation von Substituenten, Homologie, Position und Kettenlänge.

## 2.1 Bruttoformelbasierte Strukturgenerierung

Zunächst wollen wir kurz zusammenfassen, wie bruttoformelbasierte Strukturgenerierungsprobleme spezifiziert werden. Die Beschreibung reflektiert das Format der Eingabe für die derzeit jüngste, objektorientierte Implementierung des Strukturgenerators *MOLGEN 4.1* von T. Grüner. Notwendige Eingabe ist eine Bruttoformel. Eine Verallgemeinerung bilden weiche Bruttoformeln:

### 2.1.1 Definition:

Sei  $\mathcal{E}$  eine Menge chemischer Elemente und

$$\mathcal{I}(\mathbb{N}) := \{[a, b] \mid a, b \in \mathbb{N}, a \leq b\} \cup \{[a, \infty[ \mid a \in \mathbb{N}\} \subseteq \mathcal{P}(\mathbb{N})$$

die Menge der Intervalle natürlicher Zahlen. Eine *weiche Bruttoformel* ist eine Abbildung  $B \in \mathcal{I}(\mathbb{N})^{\mathcal{E}}$ . Die Menge der zu  $B$  kompatiblen Bruttoformeln ist

$$\mathcal{B}_B := \{\beta \in \mathbb{N}^{\mathcal{E}} \mid \forall X \in \mathcal{E} : \beta(X) \in B(X)\}.$$

Wir nennen  $B$  *endlich*, wenn  $\mathcal{B}_B$  endlich ist.

Um den Input möglichst variabel gestalten zu können, sind auch mehrere weiche Bruttoformeln als Eingabe möglich. Jede davon kann mit mehreren Bruttoformel–Restriktionen versehen werden.

### Bruttoformel–Restriktionen

Folgende Typen von Bruttoformel–Restriktionen sind derzeit verfügbar:

- Bruttoformel–Restriktion *Atom–Anzahl*: Hier kann ein Intervall natürlicher Zahlen angegeben werden, welches die Anzahl von Atomen einschränkt.
- Bruttoformel–Restriktion *Heteroatom–Anzahl*: Hierbei kann die Anzahl von *Heteroatomen* über ein Intervall natürlicher Zahlen festgelegt werden.

Diese Bruttoformel–Restriktionen können im Falle einer nicht endlichen weichen Bruttoformeln die Anzahl kompatibler Bruttoformeln dennoch auf eine endliche Menge reduzieren. In diesem Zusammenhang sollte man auch eine Erweiterung nennen, die zwar noch nicht in *MOLGEN 4.1* eingebunden, aber bereits in *MOLGEN–MS* implementiert ist und genutzt wird:

- Bruttoformel–Restriktion *Molekülmasse*: Ein Intervall für die Summe der Atommassen wird festgelegt. Dabei kann zwischen ganzzahligen oder exakten Atommassen gewählt werden. Näheres dazu folgt in den Abschnitten 5.4 und 5.8.

Die oben aufgeführten Bruttoformel–Restriktionen haben lediglich Einfluss auf die elementare Zusammensetzung der Bruttoformel(n). In der Spezifikation des Inputs für *MOLGEN 4.1* gibt eine Reihe weiterer Bruttoformel–Restriktionen, die die Verteilung der Atomzustände beeinflussen.

- Bruttoformel–Restriktion *Konnektivität*: Hier kann festgelegt werden, ob nur Bruttoformeln zulässig sind, für die zusammenhängende molekulare Graphen existieren. Die Überprüfung dieser Restriktion basiert auf Satz 1.2.15. Dieser verwendet die Valenzen der Atome.
- Bruttoformel–Restriktion *Bindungen*: Ein Intervall natürlicher Zahlen zur Einschränkung der Summe der Vielfachheiten aller Bindungen kann angegeben werden.
- Bruttoformel–Restriktion *Doppelbindungsäquivalente*: Hier kann ein Intervall für die Summe der DBE basierend auf den verwendeten Atomzuständen spezifiziert werden.
- Bruttoformel–Restriktion *Ladung*: Hier kann die gewünschte Gesamtladung auf ein Intervall ganzer Zahlen eingeschränkt werden.
- Bruttoformel–Restriktion *Radikale*: Für die Anzahl von Atomen mit einem ungepaarten Elektron kann ein Intervall natürlicher Zahlen angegeben werden.
- Bruttoformel–Restriktion *Atomzustände*: Über diese Restriktion können für die einzelnen Atomzustände explizit Intervalle natürlicher Zahlen angegeben werden, mit denen die Vielfachheit ihres Auftretens festgelegt wird.

In diesem Zusammenhang soll noch auf eine weitere Bruttoformel–Restriktion hingewiesen werden, die bereits starken strukturellen Charakter aufweist.

- Bruttoformel–Restriktion *Wasserstoff–Verteilung*: Hier kann pro Element und Anzahl von H–Atomen ein Intervall angegeben werden. Dieses legt fest, wie viele Atome des jeweiligen Elements mit der vorgegebenen Anzahl benachbarter H–Atome auftreten müssen.

### Strukturelle Restriktionen

In *MOLGEN 4.1* gibt vier Typen struktureller Restriktionen: Den Filter für aromatische Dubletten, einen Filter, der Symmetrie-Eigenschaften berücksichtigt sowie die Restriktionstypen Makro und Substruktur.

- Strukturelle Restriktion *Aromatizität*: Aromatische Bindungen werden identifiziert und aromatische Dubletten gegebenenfalls unterdrückt.
- Strukturelle Restriktion *Symmetrie*: Hier kann die Anzahl der Kohlenstoff-Signale aus einem  $^{13}\text{C}$  NMR-Spektrum angegeben werden. Dies bewirkt, dass nur solche Strukturen ausgegeben werden, die entsprechende Anzahl von Kohlenstoff-Bahnen unter ihrer topologischen Automorphismengruppe besitzen.
- Strukturelle Restriktion *Makro*: Makros sind mehrdeutige molekulare Graphen, die in den generierten Strukturen als molekulare Subgraphen enthalten sein müssen. Dabei dürfen mehrere Makros einander nicht überlappen.
- Strukturelle Restriktion *Substruktur*: Diese Restriktion setzt sich aus einem oder mehreren Substruktur-Einträgen zusammen sowie einem Intervall, welches die Anzahl der erfüllten Substruktur-Einträge festlegt. Die Substruktur-Einträge ihrerseits bestehen aus einer Substruktur und einem Intervall, welches die Vielfachheit des Auftretens der Substruktur in der untersuchten Struktur angibt. Ein Substruktur-Eintrag ist erfüllt, wenn die Substruktur mit einer Vielfachheit vorliegt, die im Vielfachheitsintervall des Substruktur-Eintrags enthalten ist. Soll also eine Substruktur verboten werden, so ist als Vielfachheitsintervall  $[0, 0]$  anzugeben.

Die Substruktur eines Substruktur-Eintrags ist entweder eine molekulare Substruktur, eine Summenformel-Substruktur oder eine Ring-Substruktur.

- *Molekulare Substrukturen* wurden in Abschnitt 1.4 gesondert behandelt.
- Eine *Summenformel-Substruktur* wird durch eine Bruttoformel spezifiziert. Ein molekularer Graph  $M$  enthält eine Summenformel-Substruktur mit einer bestimmten Vielfachheit, wenn eine entsprechende Anzahl zusammenhängender molekularer Teilgraphen mit der gegebenen Bruttoformel in  $M$  vorhanden sind.



- Eine *Ring-Substruktur* wird durch ein Intervall der gewünschten Ringlänge beschrieben. Ein molekularer Graph  $M$  enthält eine Ring-Substruktur mit einer bestimmten Vielfachheit, wenn eine entsprechende Anzahl von Ringen der geforderten Längen in  $M$  vorhanden ist.

Eine wichtige Methode zur Konstruktion diskreter Strukturen und insbesondere zur bruttoformelbasierten Strukturgenerierung wird im folgenden Abschnitt kurz beschrieben.

### 2.1.1 Ordnungstreue Erzeugung

Gegen Ende der 70er Jahre entwickelten R. C. Read [112] und I. A. Faradzev [38, 39] unabhängig voneinander eine Methode, die zur Konstruktion vollständiger Listen nicht-nummerierter diskreter Strukturen geeignet ist: *Ordnungstreue Erzeugung*.

Zunächst wollen wir Reads Methode allgemein formulieren: Sei  $(\Omega, \leq)$  eine total geordnete Menge und  $G$  eine Gruppe, die auf  $\Omega$  operiert. Dann ist

$$\text{rep}_{<}(\Omega//G) := \{\omega \in \Omega \mid \forall g \in G : \omega \leq \omega^g\}$$

eine kanonische Transversale der Bahnen von  $G$  auf  $\Omega$ . Die ordnungstreue Erzeugung nach Read beschreibt der folgende Satz.

#### 2.1.2 Satz:

Sei  $\Omega = \bigcup_{i \in n} \Omega_i$ ,  $\Omega_i^G \subseteq \Omega_i$ ,  $\Omega_i \cap \Omega_j = \emptyset$  ( $i \neq j$ ) und  $P$  ein Algorithmus, der  $\forall \omega \in \text{rep}_{<}(\Omega_i//G)$  eine Menge  $P(\omega) \subseteq \Omega$  erzeugt, so dass

- $\forall i : \text{rep}_{<}(\Omega_{i+1}//G) \subseteq \bigcup_{\omega \in \text{rep}_{<}(\Omega_i//G)} P(\omega)$ ,
- $\forall i : \forall \omega \in \text{rep}_{<}(\Omega_{i+1}//G) : \exists! \omega' \in \text{rep}_{<}(\Omega_i//G) : \omega \in P(\omega')$ .

Dann erhält man die gesuchte Transversale durch:

- i) Erzeuge  $\text{rep}_{<}(\Omega_0//G)$  und setze  $\text{rep}_{<}(\Omega//G) := \text{rep}_{<}(\Omega_0//G)$ .
- ii) Für  $i \in n$  durchlaufe  $\text{rep}_{<}(\Omega_i//G)$  mit  $\omega$ , und erzeuge dabei  $P(\omega)$ . Durchlaufe weiter  $P(\omega)$  mit  $\omega'$  und prüfe, ob  $\omega'$  minimal in seiner Bahn ist. Falls ja, so setze  $\text{rep}_{<}(\Omega//G) := \text{rep}_{<}(\Omega//G) \cup \{\omega'\}$ .

Im Falle schlichter Graphen  $\mathcal{G}_{n,2}$  kann man die *lexikographische Ordnung* auf der Menge der Kanten heranziehen.

**2.1.3 Definition:**

Die Menge  $\binom{n}{2}$  lässt sich wie folgt ordnen: Seien  $e = \{v, w\}$ ,  $e' = \{v', w'\} \in \binom{n}{2}$ ,  $v < w$  und  $v' < w'$ . Dann ist

$$e < e' \quad :\iff \quad v < v' \vee (v = v' \wedge w < w').$$

Dies induziert eine lexikographisch Ordnung auf  $\mathcal{G}_{n,2}$ : Für zwei Graphen  $\gamma, \gamma' \in \mathcal{G}_{n,2}$  mit  $E_\gamma = \{e_i \mid e_i \in t\}$  und  $E_{\gamma'} = \{e'_i \mid e'_i \in t'\}$  gelte  $e_0 < \dots < e_{t-1}$ ,  $e'_0 < \dots < e'_{t'-1}$ . Dann ist

$$\begin{aligned} \gamma < \gamma' \quad :\iff \quad & (\exists i < \min\{t, t'\} : e_i < e'_i \wedge \forall j < i : e_j = e'_j) \\ & \vee (t < t' \wedge \forall j < t : e_j = e'_j). \end{aligned}$$

Nun können wir Satz 2.1.2 auf Graphen anwenden [61]:

**2.1.4 Satz:**

Falls  $\gamma \in \text{rep}_<(\mathcal{G}_{n,2} // S_n)$  und  $\gamma_0 \in \mathcal{G}_{n,2}$  mit  $E_{\gamma_0} \subset E_\gamma$  und  $\gamma_0 < \gamma$ , so folgt  $\gamma_0 \in \text{rep}_<(\mathcal{G}_{n,2} // S_n)$ .

*Beweis:*

Sei  $E_\gamma = E_{\gamma_0} \cup E_{\gamma_1}$  und  $\gamma_0 \notin \text{rep}_<(\mathcal{G}_{n,2} // S_n)$  mit  $\gamma_0 < \gamma$ . Dann ist  $\gamma_0^\pi < \gamma_0$  für ein  $\pi \in S_n$ . Falls  $E_{\gamma_0^\pi} = \{e_0, \dots, e_{t-1}\}$  mit  $e_0 < \dots < e_{t-1}$ , dann existiert ein  $i$  mit  $E_{\gamma_0} = \{e_0, \dots, e_{i-1}, e'_i, \dots, e'_{t-1}\}$ ,  $e_{i-1} < e'_i < \dots < e'_{t-1}$  und  $e_i < e'_i$ . Aus  $E_{\gamma^\pi} = E_{\gamma_0^\pi} \cup E_{\gamma_1^\pi} \supseteq \{e_0, \dots, e_i\}$  folgt  $\gamma^\pi < \gamma_0 < \gamma$ , ein Widerspruch zu  $\gamma \in \text{rep}_<(\mathcal{G}_{n,2} // S_n)$ .  $\square$

**2.1.5 Bemerkung:**

Um Satz 2.1.2 anwenden zu können, müssen wir für einen Repräsentanten  $\gamma$  die Menge  $P(\gamma)$  definieren. Wir setzen

$$P(\gamma) := \{\gamma \cup \{e\} \mid e > \max\{e' \in \gamma\}\}.$$

Der folgende Algorithmus liefert dann (bei Aufruf mit dem leeren Graphen) die gesuchte Transversale von  $\mathcal{G}_{n,2}$ .

**2.1.6 Algorithmus:**  $OrdRek(\gamma)$ 

- (1) **if**  $\gamma \notin \text{rep}_<(\mathcal{G}_{n,2} // S_n)$
- (2)       **return**
- (3)      $Output(\gamma)$
- (4) **for each**  $\gamma' \in P(\gamma)$  (in aufsteigender Reihenfolge) **do**
- (5)        $OrdRek(Gr')$

Der zeitaufwendigste Schritt ist dabei der *Minimalitätstest* in Zeile 1. Vom naiven Standpunkt aus betrachtet müssen dabei alle  $n!$  Permutationen von  $S_n$  durchlaufen werden. Obwohl mit algebraisch-kombinatorischen Hilfsmitteln [50] der Aufwand deutlich gesenkt werden kann, bleibt dieser Test verhältnismäßig teuer. Allerdings erhält man im Fall eines positiv verlaufenen Minimalitätstests sogleich die Automorphismengruppe des getesteten Graphen in Form einer *Sims-Kette* [134].

Ebenfalls in R. Grunds Arbeit [50] wird ein wesentlich weniger aufwändig überprüfbares notwendiges Kriterium für Minimalität angegeben, die so genannte *Semikanonizität*. Des Weiteren erhält man im Falle eines negativ verlaufenen Minimalitätstests eine Permutation, die ein *Lernkriterium* nutzt, um weitere nachfolgende nicht minimale Kandidaten zu überspringen.

Diese Strategie wurde unter anderem erfolgreich angewendet zur Konstruktion von Katalogen diverser diskreter Strukturen, darunter Konfigurationen und Doppelnebenklassen [53], reguläre Graphen [95, 96] sowie molekulare Graphen [51]. Insbesondere für die Generierung *aller* Isomere zu einer gegebenen Bruttoformel ist ordnungstreue Erzeugung äußerst effizient, und kam auch in der vorliegenden Arbeit mehrfach zum Einsatz (Abschnitt 5.4.2 und Anhang E).

Oft spielen bei der Konstruktion strukturelle Einschränkungen eine Rolle. Man ist nur an solchen Strukturen interessiert, die bestimmte Restriktionen erfüllen.

**2.1.7 Definition:**

Eine *Restriktion* auf  $\mathcal{G}_{n,2}$  ist eine Invariante

$$R : \mathcal{G}_{n,2} \longrightarrow \mathbb{B}, \quad \gamma \longmapsto R(\gamma).$$

Man sagt,  $\gamma$  erfüllt die Restriktion  $R$ , falls  $R(\gamma) = \text{wahr}$ .  $R$  heißt *konsistent* (vgl. [28]), wenn  $\forall \gamma, \gamma' \in \mathcal{G}_{n,2}$  mit  $E_{\gamma'} \subseteq E_\gamma$  gilt:

$$R(\gamma') = \text{falsch} \implies R(\gamma) = \text{falsch}.$$

Anderenfalls nennt man  $R$  *inkonsistent*.

Beispiele für konsistente Restriktionen sind *graphentheoretische Planarität* oder eine feste untere Grenze für die Tailenweite. Die Existenz eines Kreises vorgegebener Länge ist hingegen inkonsistent.

Möchte man zu einer gegebenen Menge von Restriktionen alle Graphen konstruieren, die diese Einschränkungen erfüllen, könnte man natürlich zunächst den gesamten Katalog generieren und bezüglich der Restriktionen testen. Im Falle einer Menge konsistenter Restriktionen  $\mathcal{CR}$  kann man deren Überprüfung schon während der Generierung durchführen und Graphen, die mindestens eine Restriktion nicht erfüllen, verwerfen. Dazu ist in obigen Algorithmus die Zeile

```
(0)   if  $\exists R \in \mathcal{CR} : R(\gamma) = falsch$  return
```

einzuügen. Je nach Selektivität der Restriktionen kann man auf diese Weise eine enorme Verringerung des Zeitaufwands und damit eine Steigerung der Reichweite erzielen. Jedoch finden sich gerade in Strukturgenerierungsproblemen, die aus Fragestellungen der molekularen Strukturaufklärung resultieren, nicht selten eine Vielzahl inkonsistenter Restriktionen. Zudem ist man in diesem Zusammenhang bestrebt, möglichst kleine, im Idealfall ein-elementige Strukturräume zu definieren. Dabei tritt das Isomorphieproblem in den Hintergrund. Den weitaus schwierigeren Schritt stellt hier bereits die Konstruktion der nummerierten Strukturen dar. Ein Verfahren, das gut an diese Problemstellung angepasst ist, hat T. Grüner in seiner Arbeit [53, 54] als zielgerichtete Generierung vorgestellt.

### 2.1.2 Zielgerichtete Erzeugung

Zielgerichtete Erzeugung besitzt im Gegensatz zur ordnungstreuen Erzeugung keine feste Konstruktionsreihenfolge. Vielmehr wird die Konstruktionsreihenfolge durch die Restriktionen gesteuert. Dabei wird die Strategie zum Einsetzen der Kanten so gewählt, dass sich der Aufwand für die Backtrackprozedur reduziert. Heuristiken, die dieses Ziel verfolgen werden detailliert in [53] beschrieben.

Wegen dem Verzicht auf die ordnungstreue Konstruktionsreihenfolge müssen während der Strukturgenerierung alle bereits konstruierten Strukturen in ihrer kanonischen Form in einem Assoziativspeicher gehalten werden. In der Regel verwendet man dafür eine Hashtabelle.

Eine genaue Beschreibung des Kanonisierungsalgorithmus findet man in [20], eine Zusammenfassung der Methoden bietet [85]. Eine kurze Beschreibung des Softwarepakets *MOLGEN 4.0*, welches sowohl ordnungstreu als auch zielgerichtete Erzeugung umfasst, enthält [56].

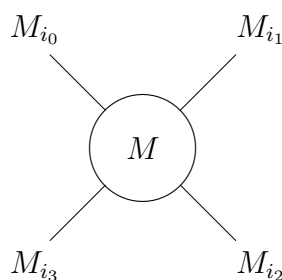
## 2.2 Reaktionsbasierte Strukturgenerierung

In der Chemie werden Veränderungen der Struktur chemischer Verbindungen durch Reaktionen beschrieben. In Abschnitt 1.5 haben wir kennen gelernt, wie chemische Reaktionen graphisch dargestellt und mit Hilfe von Reaktionsschemata virtuell nachvollzogen werden können. Es liegt nahe, dieses Modell zur Strukturgenerierung heranzuziehen. Ein Spezialfall, bei dem mehrere chemische Bausteine auf verschiedene Weisen einem Zentralkmolekül angelagert werden, ist in [164] ausführlich beschrieben. Wir wollen diese Problematik kurz rekapitulieren, bevor wir später zu Verallgemeinerungen übergehen.

### 2.2.1 Zentralkmolekül–Ligand–Anlagerungen

Gegeben sei ein *Zentralkmolekül*  $M$  und  $a$  *Liganden*  $M_i$ ,  $i \in a$  sowie ein Reaktionsschema  $R = (S, \Delta\zeta, \Delta\gamma)$ .  $R$  stelle eine Zweikomponentensynthese dar, d.h. der  $S$  zugrunde liegende MMG hat zwei Zusammenhangskomponenten  $A$  und  $B$ . Für  $A$  gebe es  $k$  verschiedene, einander nicht überlagernde Einbettungen  $\phi_j$ ,  $j \in k$  in  $M$ . Die durch die Bilder der  $\phi_j$  definierten Atome in  $M$  bilden die  $k$  *reaktiven Stellen* im Zentralkmolekül. Des Weiteren gebe es für  $B$  in jedem Liganden  $M_i$  genau eine Einbettung.

Bei Durchführung der Synthesereaktionen können die Liganden auf verschiedene Möglichkeiten den reaktiven Stellen des Zentralkmoleküls zugeordnet werden. Für  $k = 4$  lässt sich die Situation folgendermaßen skizzieren:



Die wesentlich verschiedenen Zuordnungen der Liganden auf die reaktiven Stellen des Zentralkmoleküls unter Berücksichtigung der Symmetrie von  $M$  können auf algebraisch–kombinatorische Weise ermittelt werden. Die Automorphismengruppe  $\text{Aut}(M)$  operiert auf den durch die reaktiven Stellen erklärten Atomen von  $M$  und induziert eine Untergruppe  $G \leq S_k$ , die auf den reaktiven Stellen operiert. Die wesentlich verschiedenen Belegungen der reaktiven Stellen mit Liganden sind gegeben durch die Bahnen  $a^k // G$ .

Konkrete Beispiele zu diesem Konstruktionsproblem folgen in Abschnitt 4.2. Wir wollen uns jetzt allgemeineren Fällen zuwenden, bei denen die oben geforderten Einschränkungen nicht zutreffen.

### 2.2.2 Konstruktion nach dem Netzwerkprinzip

Im Folgenden wird erarbeitet, wie durch sukzessive Anwendung von Reaktionsschemata Molekülbibliotheken generiert werden, die in der molekularen Strukturaufklärung und der kombinatorischen Chemie von besonderer Bedeutung sind.

Nahezu alle chemischen Prozesse können durch *chemische Reaktionsnetzwerke* dargestellt werden. Ein chemisches Reaktionsnetzwerk ist ein *bipartiter, gerichteter* Graph. Die Knotenmenge ist partitioniert in chemische Verbindungen und Reaktionen, die Knoten sind entweder beschriftet mit molekularen Graphen oder mit Reaktionsschemata. Die Kanten verlaufen von Edukten in Richtung der Reaktionen, an denen sie beteiligt sind bzw. von den Reaktionen in Richtung der Produkte. Während dasselbe Reaktionsschema mehrfach als Knotenbeschriftung auftreten kann, dürfen molekulare Graphen nur einmal in Form ihres kanonischen Bahnrepräsentanten als Knotenbeschriftung verwendet werden.

Wir werden nicht gesamte Reaktionsnetzwerke konstruieren, sondern sie als Grundlage verwenden, um die darin vorkommenden molekularen Graphen zu generieren. Dabei werden die Reaktionsnetzwerke beginnend bei gegebenen Ausgangssubstanzen nach *breadth-first* Strategie durchlaufen. Später werden Steuermechanismen eingeführt, um die Menge der generierten Strukturen bestmöglich an die konkreten Problemstellungen anzupassen.

Zunächst müssen dazu einige Begriffe aus Abschnitt 1.5 verallgemeinert werden. Dort hatten wir die Menge der Produktgraphen bei Anwendung eines Reaktionsschemas  $R = (S, \Delta\zeta, \Delta\gamma)$  auf einen molekularen Graphen  $M \in \mathcal{M}$  kennen gelernt:

$$\text{Prod}_R(M) = \{R \circ_\phi M \in \mathcal{M}_n \mid \phi \in \text{Emb}_{\subseteq i}(S, M)\}.$$

Wir erweitern diese Definition auf Mengen molekularer Graphen und Mengen von Reaktionsschemata. Wir gehen dabei aus von einer Menge

$$\mathcal{L} = \{M_i \mid i \in l\} \subset \mathcal{M}^C$$

zusammenhängender molekularer Graphen. Um die Menge der Produkte, die aus der Anwendung von  $R$  auf  $\mathcal{L}$  hervorgehen können zu bestimmen, müssen wir zunächst den Reaktionstyp von  $R$  untersuchen. Dieser manifestiert sich u.a. in den Zusammenhangskomponenten des  $S = (MMG, \{SR_i \mid i \in h\})$  zugrunde liegenden MMG

$$\text{Con}(R) := \text{Con}(S) := \text{Con}(MMG).$$

An der durch  $R$  dargestellten Reaktion können bis zu  $\text{Con}(R)$  Edukte beteiligt sein. Um dies zu berücksichtigen, müssen vor der Anwendung von  $R$

entsprechende Summen aus den *Kombinationen mit Wiederholung* der Reaktanden in  $\mathcal{L}$  konstruiert werden.

Kombinationen mit Wiederholung von  $n$  Elementen aus einer  $m$ -elementigen Grundmenge sind bijektiv zu den schwach monotonen Abbildungen

$$m_{\leq}^n := \{f \in m^n \mid \forall i \in n - 1 : f(i) \leq f(i + 1)\}.$$

Damit können wir die Produktgraphen bei der Anwendung von  $R$  auf  $\mathcal{L}$  erklären als

$$\text{Prod}_R(\mathcal{L}) := \bigcup_{k \in |\text{Con}(R)|} \bigcup_{f \in l_{\leq}^k} \text{Prod}_R \left( \bigoplus_{i \in k} M_{f(i)} \right),$$

und für eine Menge  $\mathcal{R}$  von Reaktionsschemata

$$\text{Prod}_{\mathcal{R}}(\mathcal{L}) := \bigcup_{R \in \mathcal{R}} \text{Prod}_R(\mathcal{L})$$

setzen. Schließlich müssen die Produktgraphen in Zusammenhangskomponenten zerlegt und unter diesen isomorphe Dubletten eliminiert werden. Dazu definieren wir für eine beliebige Menge molekularer Graphen  $\mathcal{L}$

$$\text{Con}(\mathcal{L}) := \bigcup_{M \in \mathcal{L}} \text{Con}(M)$$

und

$$\kappa(\mathcal{L}) := \{\kappa(M) \mid M \in \mathcal{L}\}.$$

Damit stehen alle Werkzeuge zur Verfügung, die für die Generierung von Molekülbibliotheken nach dem Netzwerkprinzip benötigt werden. Die Konstruktion aller aus einer Menge von Reaktanden  $\mathcal{L}$  und einer Menge von Reaktionsschemata  $\mathcal{R}$  resultierenden Bibliothek beschreibt folgender Algorithmus:

### 2.2.1 Algorithmus: *MolLib*( $\mathcal{L}, \mathcal{R}$ )

- (1)  $\mathcal{L}_0 \leftarrow \kappa(\mathcal{L}), k \leftarrow 0$
- (2) **while**  $\mathcal{L}_k \neq \emptyset$  **do**
- (3)      $k \leftarrow k + 1$
- (4)      $\mathcal{L}_k \leftarrow \kappa(\text{Con}(\text{Prod}_{\mathcal{R}}(\bigcup_{i \in k} \mathcal{L}_i))) \setminus \bigcup_{i \in k} \mathcal{L}_i$
- (5)     *Output*( $\mathcal{L}_k$ )
- (6) **end**

Zunächst werden in Zeile (0) die Reaktanden der Eingabe auf kanonische Form gebracht, etwaige Dubletten eliminiert und die kanonisch nummerierten Strukturen  $\mathcal{L}_0$  zugewiesen. Von zentraler Bedeutung in Algorithmus 2.2.1 ist Zeile (4). Dort werden aus den zuvor generierten Teilbibliotheken  $\mathcal{L}_i$ ,  $i \in k$  neue Strukturen  $\mathcal{L}_k$  gewonnen. Der Generierungsprozess ist dann abgeschlossen, wenn keine neuen Strukturen mehr produziert werden. Dies wird in Zeile (2) überprüft.

Im Folgenden werden wir den Algorithmus dahingehend modifizieren, dass er auf spezielle Probleme anwendbar wird, die im Rahmen dieser Arbeit eine Rolle spielen.

### Generierung von MS-Fragmenten

Ursprünglicher Anlass für die Entwicklung eines Strukturgenerators nach dem Netzwerkprinzip war die Anforderung, Fragmente zu generieren, die in einem Massenspektrometer auftreten können. Ohne Kapitel 5 allzu weit vorgehen zu wollen, seien hier kurz die Besonderheiten dieser Situation aufgezählt:

- i) Die Menge der Reaktanden in der Eingabe ist einelementig:  $\mathcal{L} = \{M\}$ .
- ii) Alle Reaktionsschemata sind unimolekular.
- iii) Die Menge der Reaktionsschemata ist partitioniert in zwei Teilmengen, Ionisierungsschemata und Fragmentierungsschemata:  $\mathcal{R} = \mathcal{R}_I \dot{\cup} \mathcal{R}_F$ .
- iv) In einem ersten Schritt wird auf  $M$  eine Ionisierung angewendet, es entsteht ein positiv geladenes Teilchen und optional ein Neutralteilchen.
- v) Für den weiteren Reaktionsverlauf sind nur die positiv geladenen Teilchen relevant.
- vi) Nach der Ionisierungsreaktion können beliebig viele Fragmentierungsreaktionen folgen.

Für i) und ii) sind keine Modifikationen notwendig. Um den übrigen Anforderungen gerecht zu werden, führen wir folgende Erweiterungen ein:

- Jedem Reaktionsschema wird eine *Tiefe* zugewiesen, in der es im Reaktionsverlauf verwendet werden darf. Um möglichst variabel zu bleiben, spezifizieren wir Intervalle:

$$\text{depth}_{\mathcal{R}} : \mathcal{R} \longrightarrow \mathcal{I}(\mathbb{N}^*), \text{ wobei } \text{depth}_{\mathcal{R}}(R) = \begin{cases} [1, 1] & \text{falls } R \in \mathcal{R}_I, \\ [2, \infty[ & \text{sonst.} \end{cases}$$

Hiermit können Anforderungen iii), iv) und vi) realisiert werden.



- Anstelle von  $\text{Con}()$  wird

$$\text{Con}^+(\mathcal{L}) := \{M \in \text{Con}(\mathcal{L}) \mid \text{cha}(M) = 1\},$$

für die Zerlegung und die Auswahl von Zusammenhangskomponenten der Produktgraphen eingeführt. Dabei bezeichnet  $\text{cha}(M)$  die Summe der Ladungen der Atome von  $M$ .

Da nur unimolekulare Reaktionen vorkommen, kann man sich in Zeile (4) von Algorithmus 2.2.1 bei der Produktbildung auf  $\mathcal{L}_{k-1}$  beschränken. Durch  $\text{Prod}_{\mathcal{R}}(\bigcup_{i \in k} \mathcal{L}_i)$  würden ansonsten nur Dubletten produziert. Der modifizierte Algorithmus erhält nun als zusätzliches Argument die Tiefen, in denen die Reaktionsschemata angewendet werden:

### 2.2.2 Algorithmus: $\text{MolLibMS}(\mathcal{L}, \mathcal{R}, \text{depth}_{\mathcal{R}}())$

- (1)  $\mathcal{L}_0 \leftarrow \kappa(\mathcal{L}), k \leftarrow 0$
- (2) **while**  $\mathcal{L}_k \neq \emptyset$  **do**
- (3)      $k \leftarrow k + 1$
- (4)      $\mathcal{R}' \leftarrow \{R \in \mathcal{R} \mid k \in \text{depth}_{\mathcal{R}}(R)\}$
- (5)      $\mathcal{L}_k \leftarrow \kappa(\text{Con}^+(\text{Prod}_{\mathcal{R}'}(\mathcal{L}_{k-1}))) \setminus \bigcup_{i \in k} \mathcal{L}_i$
- (6)      $\text{Output}(\mathcal{L}_k)$
- (7) **end**

### Generierung kombinatorischer Bibliotheken

Eine wichtige Rolle spielt im Rahmen dieser Arbeit auch die Generierung kombinatorischer Bibliotheken. In Abschnitt 2.2.1 haben wir bereits den speziellen Fall kennen gelernt, dass verschiedene Liganden *einem* Zentralmolekül angelagert werden sollen. Oft treten aber Situationen auf, in denen eine Generierung nach dem Netzwerkprinzip vorzuziehen ist, etwa wenn Ringschlüsse auftreten können oder wenn verschiedene Zentralmoleküle und Reaktionsschemata verwendet werden dürfen. Für unsere Zwecke können wir folgende Besonderheiten hinsichtlich der Generierung kombinatorischer Bibliotheken notieren:

- i) Die Menge der Reaktanden ist in zwei Teilmengen, Zentralmoleküle und Liganden, partitioniert:  $\mathcal{L} = \mathcal{L}_C \dot{\cup} \mathcal{L}_L$ .
- ii) Die Zentralmoleküle dürfen nur einmal während des Reaktionsverlaufs, und zwar zu dessen Beginn verwendet werden.
- iii) In jedem Reaktionsprodukt muss mindestens ein Zentralmolekül verwendet worden sein.

- iv) Alle Reaktionsschemata sind uni- oder bimolekular.
- v) Bimolekulare Reaktionen von zwei Zwischenprodukten sind zu vernachlässigen.
- vi) Während des Reaktionsverlaufs sind Nebenprodukte zu vernachlässigen.

Um diesen Anforderungen gerecht zu werden, nehmen wir folgende Erweiterungen vor:

- Jedem Reaktanden wird eine *Tiefe* zugewiesen, in der er während des Reaktionsverlaufs verwendet werden darf. Um möglichst variabel zu bleiben, geben wir Intervalle an:

$$\text{depth}_{\mathcal{L}} : \mathcal{L} \longrightarrow \mathcal{I}(\mathbb{N}), \text{ wobei } \text{depth}_{\mathcal{L}}(M) = \begin{cases} [0, 0] & \text{falls } M \in \mathcal{L}_C, \\ [1, \infty[ & \text{sonst.} \end{cases}$$

Somit können Anforderungen i) – iii) erfüllt werden.

- Um Anforderung iv) gerecht zu werden, verwenden wir zur Auswahl von Zusammenhangskomponenten in den Produktgraphen

$$\begin{aligned} \text{Con}^{\geq}(\mathcal{L}) &:= \bigcup_{M \in \mathcal{L}} \text{Con}^{\geq}(M), \quad \text{wobei} \\ \text{Con}^{\geq}(M) &:= \{M' \in \text{Con}(M) \mid \text{size}(M') \geq \frac{1}{2} \text{size}(M)\}. \end{aligned}$$

Dabei bezeichnet  $\text{size}(M)$  die Anzahl der Atome von  $M$ .

Als weiteres Argument wird nun  $\text{depth}_{\mathcal{L}}()$  an den Konstruktionsalgorithmus übergeben. Damit der Algorithmus für verschiedene Funktionen zur Auswahl der Zusammenhangskomponenten verwendbar bleibt, wird zudem  $\text{Con}^*(\cdot)$  als Argument aufgeführt. Für kombinatorische Bibliotheken wird man in der Regel  $\text{Con}^*(\cdot) = \text{Con}^{\geq}(\cdot)$  wählen. Anforderung v) fließt in Zeile (6) ein, wobei iv) verwendet wird.

### 2.2.3 Algorithmus: $\text{MolLibCC}(\mathcal{L}, \mathcal{R}, \text{depth}_{\mathcal{R}}(), \text{depth}_{\mathcal{L}}(), \text{Con}^*(\cdot))$

- (1)  $\mathcal{L}_0 \leftarrow \kappa(\{M \in \mathcal{L} \mid 0 \in \text{depth}_{\mathcal{L}}(M)\})$ ,  $k \leftarrow 0$
- (2) **while**  $\mathcal{L}_k \neq \emptyset$  **do**
- (3)      $k \leftarrow k + 1$
- (4)      $\mathcal{L}' \leftarrow \{M \in \mathcal{L} \mid k \in \text{depth}_{\mathcal{L}}(M)\}$
- (5)      $\mathcal{R}' \leftarrow \{R \in \mathcal{R} \mid k \in \text{depth}_{\mathcal{R}}(R)\}$
- (6)      $\mathcal{L}_k \leftarrow \kappa(\text{Con}^*(\text{Prod}_{\mathcal{R}'}(\mathcal{L}_{k-1} \cup \mathcal{L}')) \setminus \bigcup_{i \in k} \mathcal{L}_i)$
- (7)     Output( $\mathcal{L}_k$ )
- (8) **end**

**2.2.4 Bemerkung:**

In manchen Fällen ist es nützlich, wenn zur Generierung von Molekülbibliotheken weitere Steuermechanismen zur Verfügung stehen.

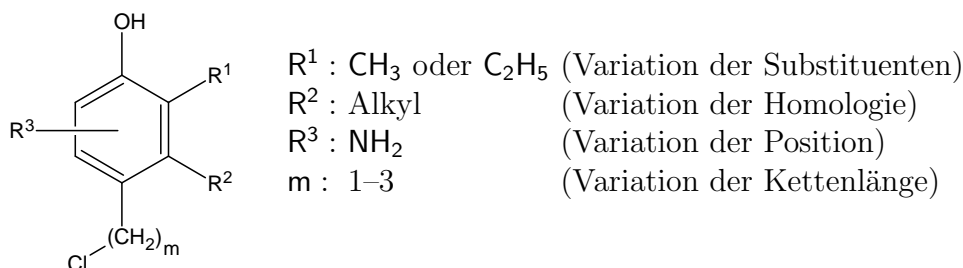
- Es sollen nur Endprodukte ausgegeben werden, Zwischenprodukte sind nicht relevant.
- Die einzelnen Reaktanden und Reaktionsschemata sollen mit vorgegebenen Vielfachheiten während des Reaktionsverlaufs auftreten.

Anforderungen letzteren Typs werden beispielsweise im folgenden Abschnitt verwendet. Für die Berücksichtigung von Vielfachheiten für Reaktanden und Reaktionsschemata ist zu beachten, dass im in manchen Reaktionsnetzwerken verschiedene Wege zu demselben Produkt führen können. Auf verschiedenen Wegen können Reaktanden und Reaktionsschemata unterschiedlich oft verwendet werden. Im Konstruktionsalgorithmus muss dies berücksichtigt werden können.

## 2.3 Generische Strukturformeln

Oft werden Strukturräume in der kombinatorischen Chemie [6] und vor allem im Patentwesen der Chemie [159] durch generische Strukturformeln beschrieben. Bislang existieren kaum Strukturgeneratoren, die generische Strukturformeln in ihrer gesamten Mächtigkeit verarbeiten können. Selbst ein einheitliches, umfassendes Format zur Repräsentation generischer Strukturformeln ist derzeit nicht bekannt.

Wir wollen ein einfaches Beispiele einer generischen Strukturformel aufzeigen und durch Kombination von bruttoformel- und reaktionsbasierter Strukturgenerierung den beschriebenen Strukturraum konstruieren. Folgende generische Strukturformel<sup>1</sup> *GS*



wurde in [7] aufgeführt, um die möglichen Variationsmöglichkeiten in generischen Strukturformeln zu demonstrieren:

- Variation der Substituenten: Für Substituent  $R^1$  stehen mehrere Alternativen zur Auswahl.
- Variation der Homologie:  $R^2$  kann ein beliebiger Vertreter der spezifizierten homologen Reihe sein. Alkyle sind alle Strukturen mit Summenformel  $\text{C}_n\text{H}_{2n+1}$ ,  $n \in \mathbb{N}^*$ .
- Variation der Position: Substituent  $R^3$  kann mehrere Positionen einnehmen. In der obigen Strukturformel ist dies graphisch dargestellt, indem die Bindung von  $R^3$  zwischen den beiden möglichen benachbarten C-Atomen verläuft. Die nicht gewählte Position wird mit einem H-Atom abgesättigt.
- Variation der Kettenlänge: Hier können ein bis drei Kettenglieder  $\text{CH}_2$  eingefügt werden.

<sup>1</sup>Bei generischen Strukturformeln erfolgt die Indizierung der verschiedenen Substituenten  $R^i$  typischerweise durch hochgestellte Nummern.

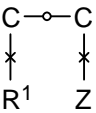
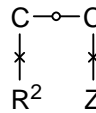
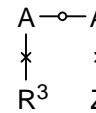
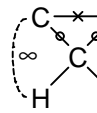
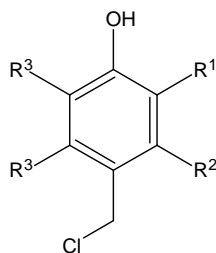
Variation der Substituenten	Variation der Homologie	Variation der Position	Variation der Kettenlänge
Reaktanden			
Z-CH <sub>3</sub> Z-C <sub>2</sub> H <sub>5</sub>	33 Isomere Z-C <sub>n</sub> H <sub>2n+1</sub> n ∈ <u>6</u>	Z-H Z-NH <sub>2</sub>	CH <sub>2</sub>
Depth: [1, 1] Mult: [0, ∞[	Depth: [2, 2] Mult: [0, ∞[	Depth: [3, 4] Mult: [1, 1]	Depth: [5, 6] Mult: [0, ∞[
Reaktionsschemata			
			
Depth: [1, 1] Mult: [1, 1]	Depth: [2, 2] Mult: [1, 1]	Depth: [3, 4] Mult: [2, 2]	Depth: [5, 6] Mult: [0, 2]

Tabelle 2.1: Reaktanden und Reaktionsschemata für die Generierung des durch *GS* beschriebenen Strukturraums

In der Realität sind generische Strukturformeln oft wesentlich komplexer und füllen nicht selten mehrseitige Patentschriften. Insbesondere kann die Komplexität durch Kombination und Rekursion der oben genannten Prinzipien erhöht werden. So kann beispielsweise die Variation von Substituenten und Position kombiniert, oder die einzelnen Substituenten R<sup>i</sup> ihrerseits als generische Strukturformeln angegeben werden.

Bei der Homologie-Variation ist zu beachten, dass die resultierenden Strukturräume unendliche Mächtigkeit annehmen. Für unser Rechenexempel wollen wir voraussetzen, dass nur Alkyle mit 1–6 C-Atomen zugelassen werden. Sicherlich sind derartige Einschränkungen auch in der Realität vertretbar oder sogar angebracht.

Zur Konstruktion des durch die generische Formel beschriebenen Strukturraums verwenden wir folgendes Zentralmolekül *M*:



Um es für die Generierung nach dem Netzwerkprinzip als Zentralmolekül

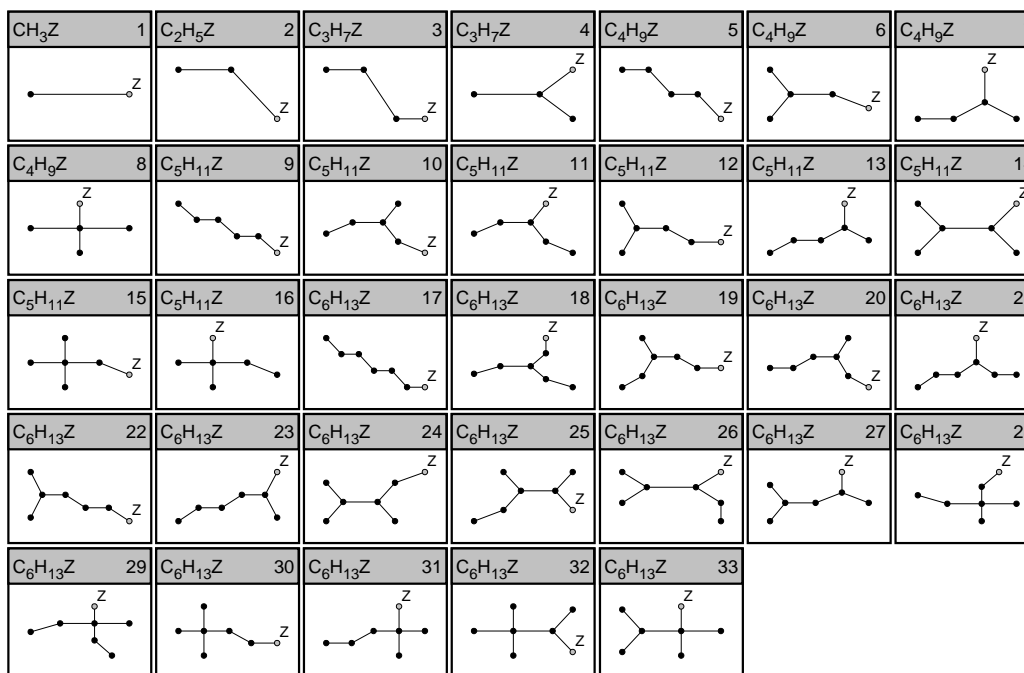


Abbildung 2.1: Alkylreste mit 1–6 C-Atomen

auszuzeichnen setzen wir  $\text{depth}_{\mathcal{L}}(M) = [0, 0]$ . Die Reste  $R^i$  werden wie Atome mit Valenz 1 und Elementsymbol  $R^1$ ,  $R^2$  bzw.  $R^3$  behandelt.

Die weiteren Reaktanden und die Reaktionsschemata sowie deren Tiefen und Vielfachheiten sind in Tabelle 2.1 angegeben. Z ist dabei auch als Atom aufzufassen und kennzeichnet die Stelle, an der das Zentralmolekül substituiert wird. Dabei werden die Reste  $R^i$  sukzessive mit Hilfe „künstlicher“ Reaktionsschemata der Form  $R^i \times X \ominus X \times Z$  eliminiert und durch die gewünschten Substituenten ersetzt.

Für der Substitution von  $R^2$  müssen zuvor alle Alkyle der Form  $Z-C_nH_{2n+1}$ ,  $n \in \underline{6}$  generiert werden, d.h. Konstitutionsisomere zu den Bruttoformeln  $C_nH_{2n+1}Z$ . Bruttoformelbasierte Strukturgenerierung liefert insgesamt 33 Isomere (Abbildung 2.1), die sich wie folgt verteilen:

$n$	1	2	3	4	5	6	Summe
Anzahl	1	1	2	4	8	17	33

Die Variation der Kettenlänge erfolgt schließlich, indem bis zu zwei  $CH_2$  durch Aufbrechen der C–Cl-Bindung eingesetzt werden. Um dabei intramolekulare Reaktionen auszuschließen, versehen wir die Reaktionssubstruktur dieses Reaktionsschemas mit einer Substruktur-Restriktion vom Typ Distanz (s. Tabelle 2.1). Auf diese Weise wird gewährleistet, dass bei der

Reaktionsdurchführung nur solche Einbettungen der Reaktionssubstruktur berücksichtigt werden, für die C–Cl und CH<sub>2</sub> in verschiedenen Edukten liegen.

Insgesamt erhält man einen Strukturraum mit 396 Verbindungen. In Kapitel 6 werden wir diesen Strukturraum hinsichtlich Überschneidungen mit einer zweiten, ebenfalls durch eine generische Strukturformel definierten Molekülbibliothek, untersuchen.





# Kapitel 3

## Überwachtes statistisches Lernen

Eine zentrale Problemstellung der Computerchemie besteht darin, Beziehungen zwischen der Struktur chemischer Verbindungen und ihren experimentell messbaren physiko-chemischen oder biologisch-pharmazeutischen Eigenschaften zu finden. Dies ist immer dann notwendig, wenn der funktionale Zusammenhang von Struktur und Eigenschaft entweder nicht bekannt oder nur mit extrem hohem Aufwand zu berechnen ist. Derartige Fragestellungen treten insbesondere bei folgenden, in gewisser Weise dualen Problemkreisen auf:

- Probleme, bei denen ausgehend von der Struktur messbare Eigenschaften vorhergesagt werden sollen.
- Probleme, bei denen ausgehend von gemessenen oder messbaren Eigenschaften auf die Struktur geschlossen werden soll.

Erstere werden als Suche nach *Struktur-Aktivitäts-Beziehungen* (engl. *Structure Activity Relationships*, kurz *SAR*) bezeichnet. Ist eine quantitativ erfassbare Eigenschaft von physiko-chemischer Natur, so spricht man auch von *quantitativen Struktur-Eigenschafts-Beziehungen* (engl. *Quantitative Structure Property Relationships*, kurz *QSPR*), im Falle einer quantitativ messbaren biologisch-pharmazeutischen Aktivität von *quantitativen Struktur-Aktivitäts-Beziehungen* (engl. *Quantitative Structure Activity Relationship*, kurz *QSAR*).

*Inverse QSAR/QSPR-Forschung* befasst sich mit Fragestellungen, bei denen zu einer messbaren quantitativen Eigenschaft Strukturen gesucht werden, die diese Eigenschaft besitzen. In der *molekularen Strukturaufklärung* werden zu

gemessenen physiko–chemischen Eigenschaften Strukturen gesucht, die diese Eigenschaften erfüllen.

Ziel dieser Bestrebungen ist zum einen, anhand bekannter Fälle Modelle zu finden, mit denen unbekannte Fälle vorhergesagt werden können. Zum anderen ist man auch daran interessiert, über die gefundenen Modelle ein tieferes Verständnis für die ursächlichen Zusammenhänge von Struktur und Eigenschaft zu gewinnen.

Wichtige mathematische Werkzeuge bei der Suche nach Zusammenhängen von Struktur und Eigenschaft bilden statistische Verfahren des *überwachten Lernens*. Voraussetzung dafür ist, dass ein gewisser Datenbestand von Paaren aus Struktur und Eigenschaft vorhanden ist.

## 3.1 Variablen und Vorhersagefunktionen

Ausgangspunkt für überwachtes Lernen ist eine Menge von  $m$  Beobachtungen,  $n$  unabhängige Variablen  $X_j$  und eine abhängige Variable  $Y$ . Anschaulicher sind die Bezeichnungen *Vorhersagevariable* für  $X_j$  und *Zielvariable* für  $Y$ . Zu jeder Beobachtung  $i \in m$  erhält man Werte  $x_{ij}$  für  $X_j$  und  $y_i$  für  $Y$ . Wir können diese Werte als Matrizen auffassen:  $\mathbf{X} = (\mathbf{x}_i) = (x_{ij})$  ist eine  $m \times n$ -Matrix<sup>1</sup>, wobei wir die Zeilen mit  $\mathbf{x}_i$  bezeichnen,  $\mathbf{Y} = (y_i)$  ist eine  $m \times 1$ -Matrix. Die Vorhersagevariablen können für unsere Zwecke durchwegs als *kontinuierlich*, d.h. reellwertig betrachtet werden. Die Zielvariable ist entweder kontinuierlich oder *diskret*. Ziel des überwachten Lernens ist es, eine *Vorhersagefunktion*  $f$  zu finden, deren Funktionswerte  $f(\mathbf{x}_i)$  mit den Werten  $y_i$  der Zielvariablen für die Beobachtungen  $i \in m$  möglichst genau übereinstimmen. Was dies für die verschiedenen Typen von Zielvariablen und Vorhersagefunktionen bedeutet, werden wir in den nächsten Abschnitten erläutern.

### 3.1.1 Beispiele:

- i) In der molekularen Strukturaufklärung sind die Beobachtungen Paare aus Spektren und chemischen Verbindungen. Als Vorhersagevariablen verwendet man *spektrale Deskriptoren*. Das sind Funktionen, die Spektren auf reelle Zahlen abbilden. Als Zielvariable wählt man beispielsweise einen *binären molekularen Deskriptor* zu einer strukturellen Eigenschaft  $S$ . Dieser liefert den Wert 1, falls eine chemische Verbindung die strukturelle Eigenschaft  $S$  besitzt, anderenfalls 0. Gesucht wird eine Funktion, die zu einem gegebenen Spektrum vorhersagen kann, ob bei der zugehörigen, unbekanntem chemischen Verbindung die fragliche strukturelle Eigenschaft  $S$  vorliegt oder nicht. Wir werden solche Vorhersagefunktionen in Abschnitt 5.5 berechnen.
- ii) Bei der QSAR-Suche sind die Beobachtungen Paare aus chemischen Verbindungen und Werten einer experimentell bestimmten Eigenschaft. Als Vorhersagevariablen dienen *molekulare Deskriptoren*. Diese bilden die topologische oder geometrische Struktur einer chemischen Verbindung auf reelle Zahlen ab. Die Zielvariable nimmt den Wert der *gemessenen Eigenschaft* an. Gesucht wird eine Funktion, die zu einer gegebenen chemischen Verbindung den fraglichen Eigenschaftswert vorher-

---

<sup>1</sup>Üblicherweise werden in der Statistik und der linearen Algebra Zeilen einer  $m \times n$ -Matrix mit  $i = 1, \dots, m$  und Spalten mit  $j = 1, \dots, n$  indiziert. Um mit der Notation aus den ersten Kapiteln konform zu bleiben wählen wir auch hier Indizes  $i \in m = \{0, \dots, m-1\}$  und  $j \in n = \{0, \dots, n-1\}$ .

sagen kann. Wir werden derartige Vorhersagefunktionen in Abschnitt 4.4 ermitteln.

### 3.1.1 Regression und Klassifikation

#### Regression

Nimmt die abhängige Variable kontinuierliche Werte an, so nennt man ein Verfahren, welches eine Vorhersagefunktion

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}, \quad \mathbf{x} \longmapsto f(\mathbf{x})$$

für  $Y$  liefert *Regression*. Steht dabei nur *eine* Vorhersagevariable zur Verfügung, spricht man von *einfacher* Regression, bei  $n > 1$  von *multipler* Regression. In jedem Fall wird angestrebt,  $f$  so zu wählen, dass die Differenzen zwischen den Werten der abhängigen Variable und den Funktionswerten der Vorhersagefunktion

$$y_i - f(\mathbf{x}_i),$$

die so genannten *Residuen*, betragsmäßig möglichst klein bleiben. Üblicherweise bestimmt man  $f$  derart, dass die Summe der Quadrate der Residuen (engl. *Residual Sum of Squares*, kurz *RSS*)

$$RSS := \sum_{i \in m} (y_i - f(\mathbf{x}_i))^2$$

minimal wird.

#### Klassifikation

Ist die abhängige Variable von diskreter Natur und nimmt Werte aus einer endlichen Menge  $\mathcal{C}$  an, so suchen wir nach einer Vorhersagefunktion

$$f : \mathbb{R}^n \longrightarrow \mathcal{C}.$$

Diese Vorgehensweise bezeichnet man als *Klassifikation*. Dabei sollen die vorhergesagten Werte  $f(\mathbf{x}_i)$  für möglichst viele Beobachtungen mit den Werten  $y_i$  der abhängigen Variable übereinstimmen. Um dies quantitativ ausdrücken zu können, bedienen wir uns einer *Kostenfunktion*

$$L : \mathcal{C} \times \mathcal{C} \longrightarrow \mathbb{R}_0^+$$

Der Funktionswert  $L(k, l)$  drückt aus, wie die Fehlklassifikation einer Beobachtung aus Klasse  $k$  als Klasse  $l$  zu werten ist. Sinnvollerweise sollten

die Kosten bei korrekter Klassifikation  $L(k, k) = 0$  sein. Falls nicht explizit angegeben, werden wir in dieser Arbeit die *Null-Eins-Kostenfunktion*  $L(k, l) = 1 - \delta(k, l)$  verwenden, wobei

$$\delta(k, l) = \begin{cases} 1 & \text{falls } k = l, \\ 0 & \text{sonst.} \end{cases}$$

die *Kroneckersche Deltafunktion* bezeichnet. In der Praxis kann es aber oft hilfreich oder sogar notwendig sein, die Kosten für Fehlklassifikationen an das konkrete Problem anzupassen. Nachdem die Kostenfunktion  $L$  festgelegt wurde, ist  $f$  so zu bestimmen, dass der *Gesamt-Klassifikationsfehler* (engl. *Total Classification Error*, kurz *TCE*)

$$TCE := \sum_{i \in m} L(y_i, f(\mathbf{x}_i))$$

minimal wird. Den Gesamt-Klassifikationsfehler kann man auch schreiben als die Summe

$$TCE = \sum_{k \in \mathcal{C}} CE^{(k)}$$

aus den Klassifikationsfehlern für die einzelnen Klassen:

$$CE^{(k)} := \sum_{i \in \Omega_k} L(k, f(\mathbf{x}_i)) \quad \text{mit } \Omega_k := \{i \in m \mid y_i = k\}, k \in \mathcal{C}.$$

Bei Null-Eins-Kostenfunktion ist TCE gerade die Anzahl der Fehlklassifikationen.

### Klassifikation durch Regression

Ein *binäres* Klassifikationsproblem d.h. ein Klassifikationsproblem mit  $|\mathcal{C}| = 2$  Klassen kann man als Regressionsproblem auffassen, indem man eine neue Zielvariable  $\tilde{Y}$  einführt, die Werte

$$\tilde{y}_i = \begin{cases} 1, & \text{falls } y_i = 1, \\ -1, & \text{sonst.} \end{cases}$$

für die Beobachtungen  $i \in m$  annimmt. Dann berechnet man eine Vorhersagefunktion  $\tilde{f}$  für  $\tilde{Y}$  durch Regression, die *Diskriminanzfunktion*. Bei Null-Eins-Kostenfunktion kann man die Vorhersagefunktion  $f$  für das Klassifikationsproblem aus  $\tilde{f}$  bestimmen durch

$$f(x) = \begin{cases} 1, & \text{falls } \tilde{f}(x) \geq 0, \\ 0, & \text{sonst.} \end{cases}$$

Im Fall von  $|\mathcal{C}| > 2$  Klassen führt man neue Zielvariablen  $\tilde{Y}_k$ ,  $k \in \mathcal{C}$  ein mit Werten

$$\tilde{y}_{ik} = \begin{cases} 1, & \text{falls } y_i = k, \\ -1, & \text{sonst,} \end{cases}$$

für  $i \in m$ ,  $k \in \mathcal{C}$ . Durch Regression erhält man Vorhersagefunktionen  $\tilde{f}_k$  für  $\tilde{Y}_k$ . Als Vorhersagefunktion  $f$  für das  $\mathcal{C}$ -Klassen-Problem wählt man

$$f(x) = \operatorname{argmax}_{k \in \mathcal{C}} \tilde{f}_k(x),$$

d.h. zu  $x \in \mathbb{R}^n$  wird diejenige Klasse  $k \in \mathcal{C}$  ausgegeben, für die  $\tilde{f}_k(x)$  maximal ist. Ist diese nicht eindeutig bestimmbar, so wird per Zufall entschieden.

### 3.1.2 Bewertung der Vorhersagefunktion

Oft gibt es verschiedene Alternativen, eine Vorhersagefunktion zu berechnen. Zum einen können Art und Anzahl der unabhängigen Variablen variiert werden, zum anderen gibt es aber auch unterschiedliche Typen von Vorhersagefunktionen sowie verschiedene Algorithmen und Parameter zu deren Bestimmung. Deshalb benötigt man Kriterien, die Aussagen über die Güte einer Vorhersagefunktion ermöglichen und zur Auswahl der besten Vorhersagefunktion herangezogen werden können.

#### Resubstitution

Zum einen ist es natürlich wichtig, wie gut die Vorhersagefunktion den beobachteten Werten der abhängigen Variable angepasst ist. Entsprechende Kennwerte bilden *RSS* im Falle einer Regression und *TCE* bei Klassifikationsproblemen. Man nennt diese Vorgehensweise *Resubstitution*, da zur Beurteilung der Vorhersagefunktion dieselben Werte eingesetzt werden, die schon zu ihrer Bestimmung verwendet wurden.

Ein typischer Kennwert zur Beurteilung der Vorhersagefunktion  $f$  im Falle einer Regression ist der *multiple Korrelationskoeffizient*

$$R := \sqrt{1 - \frac{RSS}{\sum_i (y_i - \bar{y})^2}},$$

wobei  $\bar{y} = \frac{1}{m} \sum_i y_i$  das *arithmetische Mittel* der  $y_i$  bezeichnet. Bei vollständiger Übereinstimmung der Vorhersagefunktion mit den Werten der Zielvariablen ist  $R = 1$ . Für die triviale Vorhersagefunktion  $f \equiv \bar{y}$  ist  $R = 0$ . In der Literatur wird oft das Quadrat des Korrelationskoeffizienten, das so genannte *Bestimmtheitsmaß*  $R^2$  angegeben.

$R^2$  hat den Nachteil, dass es kein geeignetes Maß ist, um die Güte von Vorhersagefunktionen mit unterschiedlicher Anzahl von Vorhersagevariablen zu vergleichen, denn  $R^2$  nimmt mit zunehmender Anzahl unabhängiger Variablen nicht ab. Dies findet Berücksichtigung im *Standardfehler* der Regression

$$S := \sqrt{\frac{RSS}{m-d}}.$$

Dabei fließt die Komplexität der Vorhersagefunktion über die Anzahl  $d$  ihrer *Freiheitsgrade* ein. Wir werden darauf in Abschnitt 3.2 für verschiedene Modelle von Vorhersagefunktionen eingehen. Bei guten Vorhersagefunktionen sollte  $S$  klein sein.

Schließlich wollen wir noch einen weiteren Kennwert erwähnen, der oft für Regressionsmodelle angegeben wird. Der *empirische  $F$ -Wert* ist definiert als

$$F := \frac{R^2}{1-R^2} \cdot \frac{m-d}{d-1}$$

und wird zum Test auf Signifikanz der Regression verwendet ([114], S. 598–599).  $F$  sollte für gute Modelle groß sein.

Im Fall von Klassifikationen ist der *mittlere Klassifikationsfehler* (engl. *Mean Classification Error*, kurz *MCE*)

$$MCE := \frac{1}{m} TCE$$

eine häufig genannte Größe zur Bewertung einer Vorhersagefunktion. Wir werden in dieser Arbeit ausschließlich  $\delta$  als Kostenfunktion verwenden. In diesem Fall bezeichnet man *MCE* auch als *Missklassifikationsrate*. Die Missklassifikationsrate ist 0, wenn alle vorhergesagten Werte mit den Werten der Zielvariable übereinstimmen, und 1, falls keine Übereinstimmungen vorliegen. Oft ist man auch daran interessiert, wie sich die richtigen und falschen Vorhersagen auf die verschiedenen beobachteten Klassen verteilen. Dazu definieren wir die Missklassifikationsrate für Klasse  $k$  als

$$MCE^{(k)} := |\Omega_k|^{-1} CE^{(k)}.$$

Dabei bezeichne  $\Omega_k$  wiederum die Indexmenge zur beobachteten Klasse  $k$ . Insbesondere darf  $\Omega_k$  nicht leer sein.

### Teststichprobe

Mindestens ebenso wichtig wie die Anpassung der Vorhersagefunktion an beobachtete Werte ist ihre *Vorhersagefähigkeit*. Um diese quantifizieren zu

können, partitioniert man die Gesamtheit der Beobachtungen zunächst per Zufall in einen *Lernsatz*  $LS$  und einen *Testsatz*  $TS$ , die *Teststichprobe*:

$$m = LS \dot{\cup} TS.$$

Es gibt keine Konventionen, die das Verhältnis von  $|LS|$  und  $|TS|$  festlegen. Wir werden in dieser Arbeit Lern- und Testsatz so wählen, dass sie gleiche Kardinalität besitzen. Beispiele dazu werden wir in den Abschnitten 4.4.2 und 5.5 kennen lernen.

Zur Bestimmung der Vorhersagefunktion  $f_{LS}$ , dem „Lernen“ verwendet man nur Beobachtungen aus  $LS$ . Eine Aussage über die Vorhersagefähigkeit ermöglicht dann im Falle einer Regression die Quadratsumme der Residuen für den Testsatz

$$RSS_{TS} := \sum_{i \in TS} (y_i - f_{LS}(\mathbf{x}_i))^2.$$

Entsprechend definieren wir

$$R_{TS}^2 := 1 - \frac{RSS_{TS}}{\sum_{i \in TS} (y_i - \bar{y}_{TS})^2}, \quad \text{wobei } \bar{y}_{TS} = \frac{1}{|TS|} \sum_{i \in TS} y_i.$$

Analog bestimmt man im diskreten Fall Gesamt- und mittleren Klassifikationsfehler für den Testsatz:

$$TCE_{TS} := \sum_{i \in TS} L(y_i, f_{LS}(\mathbf{x}_i)), \quad MCE_{TS} := \frac{1}{|TS|} TCE_{TS}$$

sowie Klassifikationsfehler und Missklassifikationsraten für einzelne Klassen im Testsatz:

$$CE_{TS}^{(k)} := \sum_{i \in TS_k} L(k, f(\mathbf{x}_i)), \quad MCE_{TS}^{(k)} := \frac{1}{|TS_k|} CE_{TS}^{(k)},$$

wobei  $TS_k := \{i \in TS \mid y_i = k\}$  die nichtleere Menge der Indizes im Testsatz mit Werten  $k$  für die Zielvariable bezeichne.

Voraussetzung für das Arbeiten mit einer Teststichprobe ist, dass genügend Beobachtungen zur Verfügung stehen. Gerade bei der Suche nach Struktur-Eigenschafts-Beziehungen in der Chemie ist oft nur eine geringe Anzahl von Beobachtungen vorhanden, so dass die Berechnung einer Vorhersagefunktion ohnehin schon an dem Mangel experimenteller Daten leidet. In dieser Situation ist es nicht vertretbar, die Anzahl der Datensätze zum Trainieren der Vorhersagefunktion durch die Aufspaltung in Lern- und Testsatz weiter zu reduzieren. Eine Lösung bietet in diesem Fall die mehrfache Verwendung der Beobachtungen sowohl zum Lernen als auch zum Testen.



### Kreuzvalidierung

Sei  $k \leq m$ . Bei der  $k$ -fachen *Kreuzvalidierung* (engl. *Crossvalidation*, kurz *CV*) wird  $m$  per Zufall in  $k$  etwa gleich mächtige Teilmengen partitioniert:

$$m = \bigcup_{l \in k} T_l.$$

Für jedes  $l \in k$  wird eine Vorhersagefunktion  $f_l$  basierend auf den Beobachtungen aus  $m \setminus T_l$  trainiert. Die beim Lernen ausgeschlossenen Beobachtungen aus  $T_l$  werden dann zur Vorhersage durch  $f_l$  herangezogen: Man erreicht damit das Ziel, ein Regressions- oder Klassifikationsverfahren auf seine Vorhersagefähigkeit zu untersuchen. Man kann CV sinnvoll zum Vergleich verschiedener Teilmengen von Vorhersagevariablen, Modelle von Vorhersagefunktionen und Parameter für Lernverfahren verwenden (s. Beispiele in Abschnitten 4.4.1 und 4.4.3). Im Falle von Regression berechnet man

$$RSS_{kCV} := \sum_{l \in k} \sum_{i \in T_l} (y_i - f_l(\mathbf{x}_i))^2,$$

bei Klassifikationsproblemen

$$TCE_{kCV} := \sum_{l \in k} \sum_{i \in T_l} L(y_i, f_l(\mathbf{x}_i)).$$

Für  $k < m$  sind diese Kennwerte abhängig von der zufälligen Zerlegung von  $m$ . Wählt man  $k = m$ , so spricht man auch von *leave one out* (kurz *LOO*-) *Kreuzvalidierung*. Dabei wird für jedes  $i \in m$  eine Vorhersagefunktion  $f_i$  berechnet, wobei jeweils nur Beobachtungen aus  $m \setminus \{i\}$  zum Lernen verwendet werden. Bei LOO-CV vereinfacht sich die Formel für die Quadratsumme der Residuen zu

$$RSS_{CV} := \sum_{i \in m} (y_i - f_i(\mathbf{x}_i))^2.$$

Entsprechend definiert man das Bestimmtheitsmaß<sup>2</sup> und den Standardfehler für LOO-CV:

$$R_{CV}^2 := 1 - \frac{RSS_{CV}}{\sum_i (y_i - \bar{y})^2}, \quad S_{CV} := \sqrt{\frac{RSS_{CV}}{m - d}}.$$

<sup>2</sup>Konsistenter wäre wohl die Definition  $R_{CV}^2 = 1 - \frac{RSS_{CV}}{\sum_i (y_i - \bar{y}_i)^2}$  mit  $\bar{y}_i = \frac{1}{m-1} \sum_{j \neq i} y_j$ , denn dann ist im Falle trivialer  $f_i \equiv \bar{y}_i$  der Korrelationskoeffizient  $R_{CV} = 0$ . Um nicht von der bestehenden Literatur abzuweichen, wurde von dieser Definition abgesehen.

Im diskreten Fall sind Gesamt- und mittlerer Klassifikationsfehler für LOO-CV gegeben durch

$$TCE_{CV} := \sum_{i \in m} L(y_i, f_i(\mathbf{x}_i)), \quad MCE_{CV} := \frac{1}{m} TCE_{CV}.$$

Es gibt weitere Möglichkeiten, das Prinzip der Kreuzvalidierung zu variieren. So kann man beispielsweise zu einem festen  $k > 1$  für alle  $k$ -Teilmengen  $T \subset m$  Vorhersagefunktionen  $f_T$  unter Ausschluss von  $T$  trainieren, und dann Vorhersagen über die Beobachtungen aus  $T$  heranziehen, um das Lernverfahren zu bewerten. Wir werden uns im Rahmen dieser Arbeit jedoch auf LOO-CV beschränken.

### 3.1.3 Datenvorverarbeitung

Manche Lernverfahren setzen voraus, dass Variablenwerte speziellen Voraussetzungen genügen. Andere Verfahren setzen zwar keine bestimmte Strukturierung der Daten voraus, liefern aber bessere Vorhersagefunktionen, wenn die Variablen einer Vorverarbeitung unterzogen wurden. Eine Auflistung solcher Methoden findet man beispielsweise in [104]. Vorausgesetzt wird dabei, dass die Werte einer Variablen nicht konstant sind. Variablen mit konstanten Werten sind für statistische Lernverfahren ohnehin nicht von Interesse, und werden im Zuge der Datenvorverarbeitung entfernt. Treten überdies zwei oder mehrere unabhängige Variablen auf, deren Werte für jede Beobachtung übereinstimmen, so genügt es, nur eine für das Lernverfahren zu berücksichtigen.

#### Lineare Transformationen

Es gibt mehrere vorverarbeitende lineare Transformationen, die auf eine abhängige oder unabhängige kontinuierliche Variable  $Z$  mit Werten  $z_i, i \in m$ , angewendet werden können:

- Als *Zentrierung* bezeichnet man die Verschiebung der Werte einer Variablen um deren arithmetisches Mittel:

$$z_i^* = z_i - \bar{z}.$$

Die zentrierten Daten  $z_i^*$  haben dann den Mittelwert 0.

- Mit *Bereichsskalierung* wird bezweckt, dass sich die Werte einer Variablen nach Vorverarbeitung über das Intervall  $[0, 1]$  erstrecken:

$$z_i^* = \frac{z_i - \check{z}}{\hat{z} - \check{z}}, \quad \text{wobei } \check{z} = \min_{i \in m} z_i \text{ und } \hat{z} = \max_{i \in m} z_i.$$

Für die so skalierten Werte  $z_i^*$  ist  $\min z_i^* = 0$  und  $\max z_i^* = 1$ .

- *Autoskalierung* bewirkt, dass die präprozessierten Daten den Mittelwert 0 und die Varianz 1 annehmen:

$$z_i^* = \frac{z_i - \bar{z}}{s}, \quad \text{wobei } s = \sqrt{\frac{\sum_{i \in m} (z_i - \bar{z})^2}{m - 1}}$$

die *Standardabweichung* der Variablen  $Z$  bezeichne. Für den des Vektor  $\mathbf{z}^* = (z_i^*)$  der autoskalierten Daten gilt dann

$$\|\mathbf{z}^*\|_2 = \sqrt{m - 1}, \quad \text{wobei } \|\mathbf{z}^*\|_2 := \sqrt{\sum_{i \in m} (z_i^*)^2}$$

für die *euklidische Norm* von  $\mathbf{z}^*$  steht.

### Nichtlineare Transformationen und Basiserweiterungen

Darüberhinaus kann es je nach Werteverteilung der Variablen auch sinnvoll sein, nichtlineare Transformationen wie  $n$ -te Wurzel oder Logarithmus auf die Variablen anzuwenden.

Man kann nichtlinear transformierte unabhängige Variablen auch als zusätzliche Vorhersagevariablen verwenden. Darüberhinaus kann man neue Variablen durch Anwendung arithmetischer Operationen auf Paare oder größere Teilmengen der Vorhersagevariablen  $X_j$ ,  $j \in n$  gewinnen. Man spricht dann von einer *Basiserweiterung*. Insbesondere finden oft quadratische Basiserweiterungen Anwendung. Dabei werden neben  $X_j$  auch die Quadrate  $X_j^2$ ,  $j \in m$  und Produkte  $X_k X_l$ ,  $\{k, l\} \in \binom{n}{2}$  als Vorhersagevariablen herangezogen.

#### 3.1.4 Variablen–Selektion

Aus verschiedenen Gründen ist es wichtig, die Komplexität der Vorhersagefunktion möglichst gering zu halten. Wird die Vorhersagefunktion zu komplex, besteht die Gefahr des *Overfittings* [65], d.h. die Vorhersagefunktion wird dem Lernsatz zu gut angepasst, zufällige Einflüsse werden in das Modell aufgenommen, und folglich wird das Modell für die Vorhersage unbrauchbar. Zum anderen können einfache Modelle besser zum Verständnis tatsächlicher Zusammenhänge beitragen.

Die Komplexität einer Vorhersagefunktion manifestiert sich unter anderem in der Anzahl der verwendeten Variablen. Allerdings ist a priori nur schwer zu entscheiden, welche Teilmenge von Variablen für das Lernverfahren gut geeignet sind.

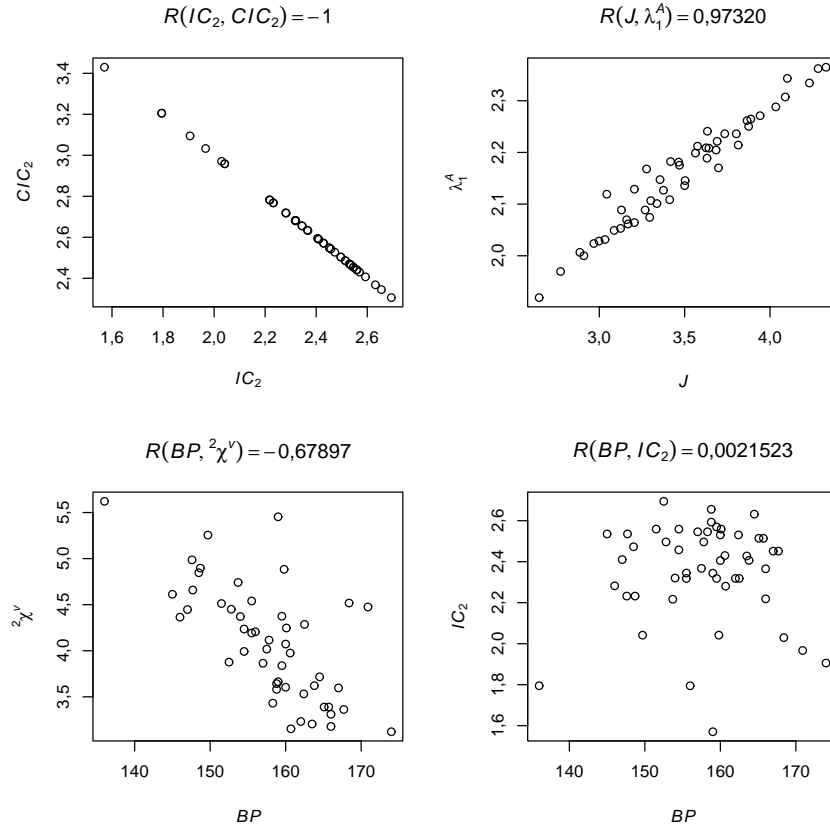


Abbildung 3.1: Beispiele starker und schwacher Korrelationen

### Korrelationsanalyse

Ein Maß für den linearen Zusammenhang zweier nicht konstanter Variablen  $X$  und  $Z$  bildet der Korrelationskoeffizient

$$\begin{aligned}
 R(X, Z) &:= \frac{\sum_i (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (z_i - \bar{z})^2}} \\
 &= \frac{\sum_i x_i z_i - m \bar{x} \bar{z}}{\sqrt{(\sum_i x_i^2 - m \bar{x}^2)(\sum_i z_i^2 - m \bar{z}^2)}} \\
 &= \frac{m \sum_i x_i z_i - \sum_i x_i \sum_i z_i}{\sqrt{(m \sum_i x_i^2 - (\sum_i x_i)^2)(m \sum_i z_i^2 - (\sum_i z_i)^2)}} \in [-1, 1].
 \end{aligned}$$

Ist  $|R(X, Z)| = 1$ , so werden  $X$  und  $Z$  *vollständig korreliert* genannt, falls  $|R(X, Z)| = 0$  *unkorreliert*. Bei der graphischen Darstellung sind die durch  $X$  und  $Z$  definierten Datenpunkte je nach Größe von  $R(X, Z) = 0$  mehr oder weniger stark um eine Gerade gestreut.

Abbildung 3.1 zeigt Beispiele verschieden starker Korrelationen mit ihren Korrelationskoeffizienten. Es handelt sich dabei um topologische Indizes und Siedepunkte einer realen Bibliothek von Decanen aus Abschnitt 4.4.1.

Sind  $X$  und  $Z$  vollständig korreliert, so gibt es eine Darstellung  $Z = aX + b$  mit  $a \neq 0$ . Die linear-algebraische Bezeichnung für diesen Sachverhalt ist *af-fine Abhängigkeit*. Vollständige Korreliertheit bildet eine Äquivalenzrelation auf der Menge der Variablen.

Soll für die Regression der Zielvariablen  $Y$  durch nur *eine* Vorhersagevariable  $X$  aus einer größeren Menge möglicher Vorhersagevariablen  $X_j$ ,  $j \in n$  ausgewählt werden, kann es durchaus sinnvoll sein, sich für diejenige Variable zu entscheiden, welche den betragsmäßig größten Korrelationskoeffizienten zu  $Y$  besitzt. Insbesondere bei linearer Regression (s. Abschnitt 3.2.1) trifft man auf diese Weise eine optimale Wahl da das Bestimmtheitsmaß  $R^2$  bei einfacher linearer Regression gerade gleich  $R(X, Y)^2$  ist.

Bei multipler und/oder nicht linearer Regression wird die Situation komplizierter. Vorhersagevariablen mit größten Korrelationskoeffizienten zur Zielvariablen bilden nicht notwendig beste Teilmengen für solche Lernverfahren. Dies ist insbesondere dann der Fall, wenn die Vorhersagevariablen untereinander starke Korrelationen aufweisen. Häufig beschränkt man sie auf solche Vorhersagevariablen, die untereinander keine vollständigen Korrelationen aufweisen.

### Fisher Ratios

Für binäre Klassifikationsprobleme werden gelegentlich in der Literatur [154, 148, 31] Fisher-Verhältnisse (engl. *Fisher Ratios*, kurz *FR*) zur Auswahl der besten Vorhersagevariablen für Klassifikation durch *eine* Vorhersagevariable angegeben. Sie sind definiert als

$$FR(X, Y) := \frac{\mu_0 - \mu_1}{\sigma_0 + \sigma_1},$$

wobei  $\mu_k$  die arithmetischen Mittel und  $\sigma_k$  die Varianzen der  $x_i$  innerhalb der durch  $Y$  erklärten Klassen angeben, d.h. mit  $\Omega_k = \{i \in m \mid y_i = k\}$ ,  $k \in \mathcal{C}$ , ist

$$\mu_k = |\Omega_k|^{-1} \sum_{i \in \Omega_k} x_i \quad \text{und} \quad \sigma_k = (|\Omega_k| - 1)^{-1} \sum_{i \in \Omega_k} (x_i - \mu_k)^2.$$

Allerdings können auch hierbei keine Abhängigkeiten zwischen mehreren Vorhersagevariablen berücksichtigt werden. Zur Auswahl der besten Vorhersagefunktionen mit mehr als einer Vorhersagevariablen bleibt oft keine andere Möglichkeit als die vollständige Suche unter allen Teilmengen verfügbarer Vorhersagevariablen.

### Beste Teilmengen–Selektion

Insbesondere sind wir daran interessiert, für gegebenes  $k \leq m$  diejenigen  $k$ -Teilmengen von Variablen zu finden, zu denen sich beste Vorhersagefunktionen bestimmen lassen (engl. *Best Subset Selection*, kurz *BSS*). Die nahe liegende Lösung ist, alle  $k$ -Teilmengen zu durchlaufen, jeweils eine Vorhersagefunktion zu bestimmen, und dann die beste(n) auszuwählen. Dies ist allerdings mit hohem Rechenaufwand verbunden. Für lineare Regression gibt es Verfahren [45], die diesen mit Hilfe von Techniken aus der linearen Algebra verringern. Trotzdem ist es oft nicht möglich, den Durchlauf aller  $k$ -Teilmengen in vertretbarer Zeit abzuarbeiten.

### Schrittweise Teilmengen–Selektion

Bei unseren Anwendungen in der Chemie steht prinzipiell eine unbeschränkte Anzahl von Variablen zur Verfügung. Wesentlich größere Reichweiten erlauben schrittweise Verfahren. Beim *einfachen* schrittweisen Verfahren (siehe z.B. [136], S. 174 ff.) wird zunächst diejenige Variable  $X_{i_1}$  bestimmt, welche die beste Vorhersagefunktion mit einer Variablen ermöglicht. Im zweiten Schritt werden alle 2-Teilmengen durchlaufen, die  $X_{i_1}$  enthalten und erneut Vorhersagefunktionen berechnet. Die beste darunter verwende die Variablen  $X_{i_1}$  und  $X_{i_2}$ . Im  $j$ -ten Schritt werden alle  $j$ -Teilmengen durchlaufen, die  $X_{i_1}, \dots, X_{i_{j-1}}$  enthalten. Insgesamt sind also im Vergleich zum Durchlauf aller  $\binom{n}{k}$   $k$ -Teilmengen nur  $\sum_{i \in k} n - i$  Vorhersagefunktionen zu bestimmen. Leider führt diese Vorgehensweise nicht notwendig zur besten Vorhersagefunktion mit  $k$  Variablen, wie sich anhand einfacher Beispiele zeigen lässt.

In manchen Fällen kann man Abhilfe schaffen, indem man bei jedem Schritt nicht nur die beste Teilmenge von Variablen weiter verwendet, sondern die  $l > 1$  besten Teilmengen im nächsten Schritt jeweils um eine Variable vergrößert. Zwar besteht bei diesem *l-fachen* schrittweisen Verfahren immer noch keine Gewissheit, die beste  $k$ -Teilmenge zu finden, meist führt diese Vorgehensweise mit größeren  $l$  jedoch zu besseren Modellen.

## 3.2 Modelle für Vorhersagefunktionen

### 3.2.1 Lineare Modelle

*Lineare Modelle* (kurz *LM*) basieren auf Vorhersagefunktionen  $f$ , bei denen die Vorhersagevariablen linear gewichtet werden:

$$f(\mathbf{x}) = \sum_{j \in n} a_j x_j + b.$$

Aus der Anzahl frei wählbarer Parameter  $a_j$  und  $b$  ergibt sich die Anzahl der Freiheitsgrade  $d = n + 1$ . Mit  $\mathbf{c} := (a_0, \dots, a_{n-1}, b)^T \in \mathbb{R}^{(n+1) \times 1}$  und  $\mathbf{z} := (x_0, \dots, x_{n-1} \mid 1)$  können wir  $f$  als Matrix-Multiplikation

$$f(x) = \mathbf{z}\mathbf{c}$$

auffassen. Mit Methoden der linearen Algebra und der mehrdimensionalen Analysis lässt sich  $\mathbf{c}$  derart bestimmen, so dass möglichst kleine Residuen auftreten. Diese Vorgehensweise wird als *multiple lineare Regression* (kurz *MLR*) bezeichnet.

#### Kleinste Quadrate Regression

Bei  $m$  Beobachtungen wollen wir die Parameter  $a_j$  und  $b$  so bestimmen, dass

$$RSS(\mathbf{c}) = \sum_{i \in m} (y_i - \mathbf{c}\mathbf{z}_i)^2$$

minimal wird. Dabei bezeichnen  $\mathbf{z}_i$  die Zeilenvektoren von

$$\mathbf{Z} = (x_{ij} \mid \mathbf{1}) \in \mathbb{R}^{m \times (n+1)}.$$

Man kann  $RSS(\mathbf{c})$  auch als Produkt

$$RSS(\mathbf{c}) = (\mathbf{y} - \mathbf{Z}\mathbf{c})^T (\mathbf{y} - \mathbf{Z}\mathbf{c}).$$

schreiben. Differentiation nach  $\mathbf{c}$  liefert:

$$\frac{\partial RSS}{\partial \mathbf{c}}(\mathbf{c}) = -2\mathbf{Z}^T (\mathbf{y} - \mathbf{Z}\mathbf{c}).$$

Hat  $\mathbf{Z}$  vollen Rang, so ist die Hessematrix

$$\frac{\partial^2 RSS}{\partial \mathbf{c} \partial \mathbf{c}^T}(\mathbf{c}) = -2\mathbf{Z}^T \mathbf{Z}$$

positiv definit, und  $RSS$  hat ein globales Minimum, wenn

$$\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}\mathbf{c}) = 0.$$

Nun kann man  $\mathbf{c}$  berechnen als

$$\mathbf{c} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y},$$

indem man die inverse Matrix  $(\mathbf{Z}^T\mathbf{Z})^{-1}$  durch die *Cholesky-Zerlegung* ermittelt. Wir verwenden im Folgenden die numerisch stabilere *QR-Zerlegung* von  $\mathbf{Z}$ . Dabei wird  $\mathbf{Z} = \mathbf{Q}\mathbf{R}$  als Produkt einer orthonormalen Matrix  $\mathbf{Q}$  und einer rechten oberen Dreiecksmatrix  $\mathbf{R}$  dargestellt. Man löst dann

$$\mathbf{y} - \mathbf{Z}\mathbf{c} = 0,$$

indem man zunächst  $\mathbf{Q}^{-1}\mathbf{y} = \mathbf{Q}^T\mathbf{y}$  berechnet und sich die Dreiecksstruktur von  $\mathbf{R}$  bei der Bestimmung von  $\mathbf{c}$  aus

$$\mathbf{R}\mathbf{c} = \mathbf{Q}^T\mathbf{y}$$

zunutze macht.

Man nennt diese Vorgehensweise,  $RSS$  zu minimieren, (*einfache*) *kleinste Quadrate Regression* (engl. *Ordinary Least Squares*, kurz *OLS*). Sie wird in den folgenden Kapiteln unser wichtigstes Werkzeug zur Bestimmung von Vorhersagefunktionen sein.

### Hauptkomponenten-Regression

Bei der *Hauptkomponenten-Regression* (engl. *Principal Component Regression*, kurz *PCR*) wird  $\mathbf{Z}$  durch *Singulärwert-Zerlegung* (engl. *Singular Value Decomposition*, kurz *SVD*) als Produkt

$$\mathbf{Z} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

zweier orthonormalen Matrizen  $\mathbf{U} \in \mathbb{R}^{m \times m}$  und  $\mathbf{V} \in \mathbb{R}^{(n+1) \times (n+1)}$  und einer Diagonalmatrix  $\mathbf{S} \in \mathbb{R}^{m \times (n+1)}$  dargestellt. Die Elemente auf der Diagonalen

$$s_0 \geq \dots \geq s_{r-1} > 0 = \dots = 0$$

sind dabei die *Singulärwerte*, wobei  $r$  den Rang von  $\mathbf{Z}$  bezeichnet. Hat  $\mathbf{Z}$  vollen Rang, dann können wir die Parameter der Vorhersagefunktion berechnen als

$$\mathbf{c} = \mathbf{V} \operatorname{diag}(1/s_i) \mathbf{U}^T\mathbf{y}$$

und sie sind identisch mit der Lösung der OLS-Regression. Im Gegensatz zur QR-Zerlegung ist SVD auch dann möglich, wenn  $\mathbf{Z}$  nicht vollen Rang hat. Hauptanliegen der PCR ist es, nur diejenigen Singulärwerte zu verwenden, für die die beste Vorhersagefähigkeit erzielt wird (s. Beispiel in Abschnitt 4.4.2).



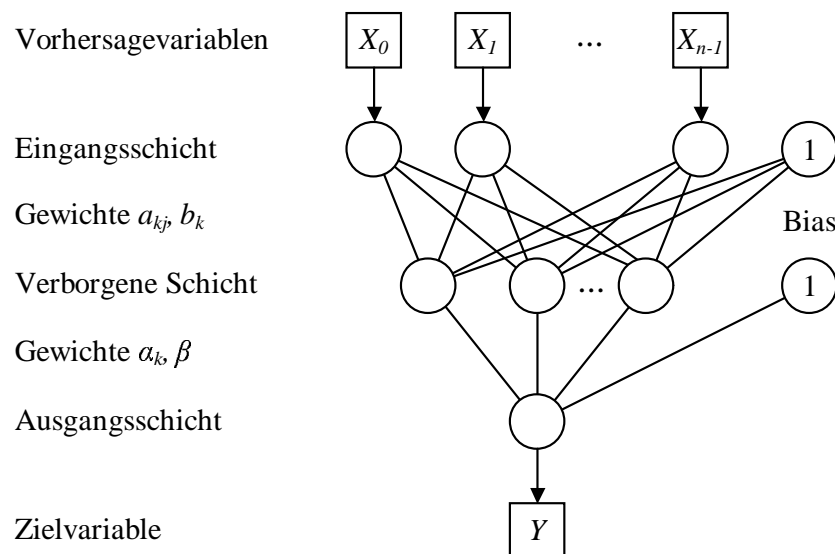


Abbildung 3.2: Schema eines neuronalen Netzes mit einer verborgenen Schicht und Bias-Neuronen

### Lineare Verfahren zur Klassifikation

Bei der Klassifikation durch lineare Verfahren werden Klassengrenzen als Hyperebenen im  $n$ -dimensionalen Datenraum beschrieben. Diese Hyperebenen gewinnt man entweder wie in Abschnitt 3.1.1 mittels *Klassifikation durch Regression* oder durch *lineare Diskriminanzanalyse (LDA)*. Beide Methoden werden eingehend in Kapitel 4 von [64] beschrieben. Wir werden in dieser Arbeit hauptsächlich binäre Klassifikation benötigen (s. Abschnitte 4.4.3 und 5.5.2).

#### 3.2.2 Neuronale Netze

Oft lassen sich die Zusammenhänge zwischen Vorhersage- und Zielvariablen nicht hinreichend gut durch lineare Vorhersagefunktionen darstellen. Beispiele für nichtlineare Vorhersagefunktionen liefern *künstliche neuronale Netze* (engl. *Artificial Neural Networks*, kurz *ANN*). Wir werden im Rahmen dieser Arbeit *Feedforward-Netze* mit einer *verborgenen Schicht* und linearen Ausgangsgewichten verwenden. Abbildung 3.2 zeigt die Architektur eines solchen Netzes. Zwischen *Eingangsschicht* und verborgener Schicht erfolgt für jedes *verborgene Neuron* (engl. *Hidden Neuron*, kurz *HN*)  $k \in h$  eine lineare Wichtung  $g_k(\mathbf{x}) = \mathbf{a}_k \mathbf{x}^T + b_k$ . Die  $b_k$  stellen dabei die Gewichte des so genannten *Bias-Neurons* zu den verborgenen Neuronen dar. In der verbor-

genen Schicht wird eine Aktivierungsfunktion  $a(z) = (1 + e^{-z})^{-1}$  angewandt. Zwischen verborgener Schicht und *Ausgangsschicht* erfolgt nochmals eine lineare Wichtung mit  $\alpha_k, k \in h$  und *Bias*  $\beta$ .

Ein Feedforward-Netz mit einer Schicht  $h$  verborgener Neuronen und linearen Ausgangs-Gewichten realisiert die folgende Modellfunktion:

$$f(\mathbf{x}) = \sum_{k \in h} \frac{\alpha_k}{1 + e^{-(\mathbf{a}_k \mathbf{x}^T + b_k)}} + \beta$$

Die Anzahl der Freiheitsgrade  $d = (n + 2)h + 1$  entspricht der Anzahl frei wählbarer Parameter  $a_{kj}, b_k, \alpha_k$  und  $\beta$ .

Zur Bestimmung dieser Parameter gibt es verschiedene Verfahren der nicht-linearen Optimierung (Levenberg–Marquardt, Gauss–Newton, Backpropagation, Steepest–descent). In der vorliegenden Arbeit wurde die im Statistik-Paket R enthaltene, auf Newton–Optimierung basierende Implementierung verwendet. Wichtig ist dabei, dass Ziel- und Vorhersagevariablen einer Bereichsskalierung unterzogen wurden. Üblicherweise startet man mit einer Zufallsbelegung der Parameter. Zwar hat dies den Nachteil, dass die Netze nicht reproduzierbar sind, jedoch findet man in der Regel durch zufällige Startparameter bessere Vorhersagefunktionen als bei einer festen Vorbelegung, etwa mit 0. Details zum Training neuronaler Netze findet man in Kapitel 11 von [64], eine Vielzahl weiterer Arten neuronaler Netze nebst Anwendungen in der Chemie stellt [168] vor.

Neuronale Netze liefern in vielen Situationen gute Vorhersagefunktionen. Allerdings sind sie mit zwei Nachteilen behaftet: Zum einen erlauben der hohe Vernetzungsgrad und die einhergehende Komplexität der Vorhersagefunktion kaum eine Interpretation. Zum anderen ist aus mathematischer Sicht unbefriedigend, dass im Allgemeinen keine Aussagen hinsichtlich der Optimalität der Vorhersagefunktion gemacht werden können. In der Regel terminiert der zum Training des ANN verwendete Optimierungsalgorithmus in einem lokalen Minimum. Einen mathematisch überzeugenderen Ansatz lernen wir im folgenden Abschnitt kennen.

### 3.2.3 Support–Vektor–Maschinen

Das Konzept der *Support–Vektor–Maschine* (kurz *SVM*) wurde Mitte der 90er Jahre von V. Vapnik [30, 146, 147] formuliert. Ursprünglich waren SVM zur Lösung binärer Klassifikationsprobleme vorgesehen. Dabei wird zunächst nach einer optimal separierenden Hyperebene zwischen den beiden Klassen gesucht (s. Abbildung 3.3). Diese wird so gewählt, dass sie ein *Rand* maximaler Größe zwischen den nächstgelegenen Punkten der beiden Klassen umgibt.

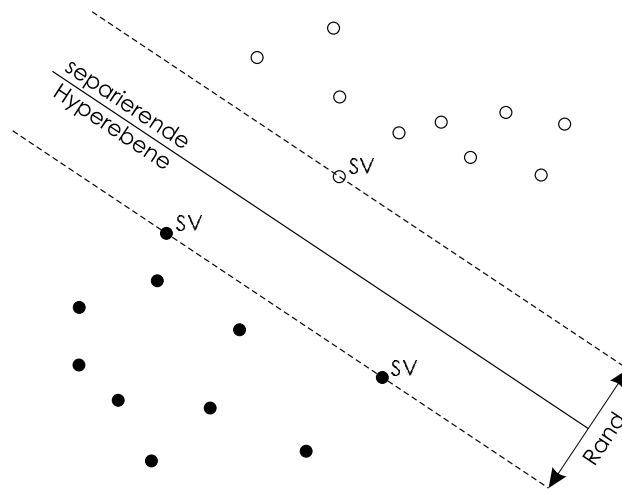


Abbildung 3.3: Support-Vektor-Klassifikator für den separablen Fall

Punkte, die auf den Grenzen des Randes liegen heißen *Support-Vektoren* und sind in Abbildung 3.3 mit „SV“ beschriftet.

Sind die beiden Klassen nicht linear separierbar, so kommen zwei weitere Strategien zum Tragen, die das Konzept der SVM umfasst. Zum einen lässt man auch Punkte auf der falschen Seite des Randes zu, versucht aber deren Einfluss möglichst gering zu halten. Zum anderen können die Datenpunkte mit Hilfe einer Basiserweiterung (vgl. Abschnitt 3.1.3) in einen höherdimensionalen Raum abgebildet werden, um dort eine bessere lineare Separierbarkeit zu erreichen.

Diese Vorgehensweisen kann man als quadratisches Optimierungsproblem mit linearen Ungleichungen als Nebenbedingungen formulieren und lösen (siehe z.B. [64]). Wir verwenden in dieser Arbeit die Implementation *libsvm* [26], die inzwischen auch über das *R*-Erweiterungspaket *e1071* [97] verfügbar ist. Bei binärer Klassifikation ist die Vorhersagefunktion von der Form

$$f(x) = \begin{cases} 1, & \text{falls } \tilde{f}(x) \geq 0, \\ 0, & \text{sonst,} \end{cases}$$

wobei

$$\tilde{f}(\mathbf{x}) = \sum_{i \in m} \tilde{y}_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b$$

und

$$\tilde{y}_i = \begin{cases} 1, & \text{falls } y_i = 1, \\ -1, & \text{sonst.} \end{cases}$$

$\alpha_i$  ist dabei nur für Support-Vektoren  $\mathbf{x}_i$  ungleich Null. Die so genannte *Kernel-Funktion*  $K$  realisiert die angesprochene Basiserweiterung. Bei der

in dieser Arbeit verwendeten Implementation [26] stehen folgende Kernel-Funktionen zur Verfügung:

$$\begin{aligned} \text{Linearer Kernel :} & \quad K(\mathbf{x}, \mathbf{x}') = \mathbf{x}'\mathbf{x}^T, \\ \text{Polynom vom Grad } d : & \quad K(\mathbf{x}, \mathbf{x}') = \gamma(\mathbf{x}'\mathbf{x}^T + c)^d, \\ \text{Radial-Basis-Funktion :} & \quad K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x}' - \mathbf{x}\|_2^2), \\ \text{Sigmoid-Funktion :} & \quad K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma\mathbf{x}'\mathbf{x}^T + c)^d. \end{aligned}$$

$\gamma$ ,  $c$  und  $d$  sind dabei frei wählbar. In der Regel setzen wir  $\gamma = 1$ , beste Werte für  $\gamma$  und  $d$  bestimmt man über CV oder mit Hilfe einer Teststichprobe.

SVM fanden in jüngster Vergangenheit immer häufiger Anwendung auf Fragestellungen der Computerchemie. Bei einem Vergleich von SVM mit ANN für die Klassifikation chemischer Verbindungen als pharmazeutische Wirkstoffe [23] lieferten SVM für verschiedene Testsätze durchwegs kleinere Klassifikationsfehler. Bei der Klassifikation von Massenspektren (s. Abschnitt 5.5.2) erwiesen sich SVM mit radialem Kernel als beste Vorhersagefunktionen.

Man kann das Konzept der SVM auch auf kontinuierliche Zielvariablen übertragen. Im Falle einer Regression ist die Vorhersagefunktion von der Form

$$f(\mathbf{x}) = \sum_{i \in m} a_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

SVM haben oft eine große Anzahl von Support-Vektoren. Die daraus resultierenden komplexen Vorhersagefunktionen können kaum zum Verständnis der ursächlichen Zusammenhänge zwischen Vorhersage- und Zielvariablen herangezogen werden. Weitaus besser interpretierbare Vorhersagefunktionen lernen wir im folgenden Abschnitt kennen.

### 3.2.4 Entscheidungsbäume

Ein weiteres Verfahren zur Bestimmung von Vorhersagefunktionen besteht in der *rekursiven Partitionierung* des  $\mathbb{R}^n$  in Hyperrechtecke. Die resultierenden Vorhersagefunktionen lassen sich als *Entscheidungsbäume* darstellen. Je nach Art der Zielvariablen spricht man von *Klassifikations-* oder *Regressionsbäumen*. Klassifikations- und Regressionsbäume (engl. *Classification and Regression Trees*, kurz *CART*) zeichnen sich besonders durch die ihre gute Interpretierbarkeit aus.

Ein *Entscheidungsbaum* ist ein *binärer Wurzelbaum*, d.h. er hat einen als Wurzel ausgezeichneten Knoten  $V_0$ , und jeder Knoten, der kein Blatt ist, hat genau zwei Nachfolger. Die Blätter werden auch *terminale* Knoten genannt, alle anderen Knoten sind *innere* Knoten. Innere Knoten  $V_k$  tragen

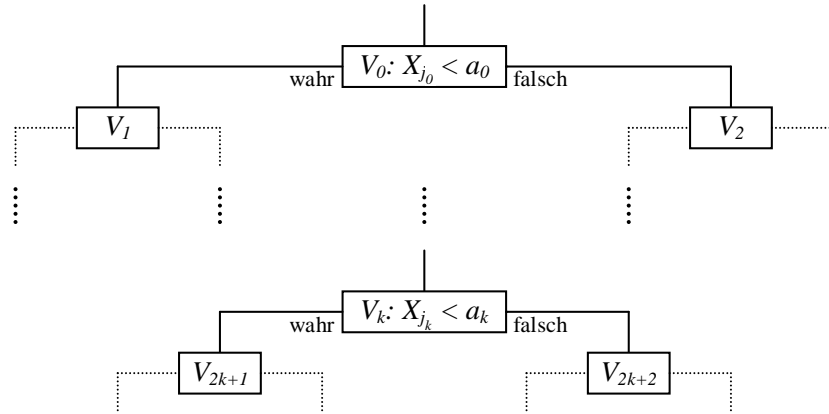


Abbildung 3.4: Schema eines Entscheidungsbaumes

*Entscheidungsregeln* der Form  $X_{j_k} < a_k$ , terminale Knoten Funktionswerte  $\hat{y}_k$ .

Die Nummerierung der Knoten ist dabei so gewählt, dass ein innerer Knoten  $V_k$  die Nachfolger  $V_{2k+1}$  und  $V_{2k+2}$  besitzt (vgl. Abbildung 3.4). Die Anwendung der durch einen Entscheidungsbaum dargestellten Vorhersagefunktion ist sehr einfach, so dass auch komplexe Bäume schnell ausgewertet werden können. Ausgehend von der Wurzel werden die inneren Knoten gemäß ihrer Entscheidungsregeln wie folgt durchlaufen: Ist die durch  $V_k$  dargestellte Entscheidungsregel erfüllt, so ist als nächster Knoten  $V_{2k+1}$  aufzusuchen, anderenfalls  $V_{2k+2}$ . Erreicht man dabei einen terminalen Knoten, so wird der durch ihn dargestellte Funktionswert ausgegeben, anderenfalls wird die nächste Entscheidungsregel abgearbeitet.

Bei der Konstruktion eines Entscheidungsbaumes wird der Lernsatz sukzessive in je zwei disjunkte Teilmengen partitioniert. Die Partitionierung erfolgt mit Hilfe einer binären Entscheidungsregel  $X_j < a$ . Sei  $\Omega_k$  die Indexmenge der durch  $V_k$  repräsentierten Beobachtungen. Des Weiteren seien

$$\Omega_k(j, a) := \{i \in \Omega_k \mid x_{ij} < a\} \quad \text{und} \quad \Omega'_k(j, a) := \{i \in \Omega_k \mid x_{ij} \geq a\}$$

die Indexmengen, die durch Aufspaltung von  $\Omega_k$  bzgl.  $X_j$  an der Stelle  $a$  resultieren. Nun werden  $j$  und  $a$  derart bestimmt, dass die Eigenschaftswerte in den nachfolgenden Knoten  $V_{2k+1}$  und  $V_{2k+2}$  möglichst homogen sind. Im Falle eines Regressionsbaumes bedeutet dies, dass

$$\sum_{i \in \Omega_k(j, a)} (y_i - \mu)^2 + \sum_{i \in \Omega'_k(j, a)} (y_i - \mu')^2$$

minimal werden soll. Dabei bezeichnen

$$\mu = |\Omega_k(j, a)|^{-1} \sum_{i \in \Omega_k(j, a)} y_i \quad \text{und} \quad \mu' = |\Omega'_k(j, a)|^{-1} \sum_{i \in \Omega'_k(j, a)} y_i$$

die Mittelwerte der Eigenschaftswerte in  $\Omega_k(j, a)$  und  $\Omega'_k(j, a)$ . Optimales  $j$  und  $a$  definieren schließlich die Entscheidungsregel für  $V_k$  und die Konstruktion kann bei  $V_{2k+1}$  und  $V_{2k+2}$  mit  $\Omega_{2k+1} = \Omega_k(j, a)$  und  $\Omega_{2k+2} = \Omega'_k(j, a)$  fortgeführt werden. Handelt es sich bei  $V_{2k+1}$  oder  $V_{2k+2}$  um terminale Knoten, so sind  $\hat{y}_{2k+1} = \mu$  bzw.  $\hat{y}_{2k+2} = \mu'$  zugleich die Funktionswerte.

Die Variablen-Selektion erfolgt also während der Konstruktion des Entscheidungsbaumes. So gesehen kann man die Konstruktion eines Entscheidungsbaumes auch als Methode zur Variablen-Selektion ansehen. Da die Auswahl jeweils nur aufgrund einer *lokalen* Bedingung getroffen wird, kann die Selektion äußerst schnell durchgeführt, und auch für sehr große Grundmengen unabhängiger Variablen in vertretbarer Zeit abgearbeitet werden.

Weitere Details wie Abbruchkriterien für das Wachstum von Entscheidungsbäumen, Strategien zum Zurückschneiden von Entscheidungsbäumen und zur Konstruktion von Klassifikationsbäumen werden in [21] beschrieben. Auf Seiten 205–215 wird dort ein Beispiel eines Klassifikationsbaums für Massenspektren gezeigt, mit dessen Hilfe das Vorhandensein von Brom im Analyten vorhergesagt werden kann. Wir werden ähnliche Problemstellungen in Kapitel 5 betrachten. Dabei verwenden wir eine Implementierung von B. D. Ripley ([115], Kapitel 7), auf die über eine Schnittstelle zur Statistik-Software R zugegriffen wird. Im Allgemeinen fanden CART im Zusammenhang mit Problemstellungen aus der Chemie bisher eher geringe Beachtung. Eine der wenigen Ausnahmen bildet [158], wo die Wirksamkeit chemischer Verbindungen als Arzneimittel untersucht wird.

Eine interessante Weiterentwicklung von Regressionsbäumen wurde in der Software *Cubist* [110] realisiert. Dabei wird das Verfahren der rekursiven Partitionierung kombiniert mit linearer Regression. Die Vorhersagefunktionen sind RT mit LM an den terminalen Knoten. In [22] wird mit Hilfe dieser Methode die Wasserlöslichkeit chemischer Verbindungen modelliert.

### 3.2.5 Nächste Nachbarn

In den vorangegangenen Abschnitten wurden jeweils Modellfunktionen durch Regressions- oder Klassifikationsverfahren algorithmisch bestimmt. Die Methode der *k* nächsten Nachbarn (kurz *KNN*) gehört zu den so genannten „modellfreien“ Lernverfahren. Zur Bestimmung der Vorhersagefunktion findet keine Strukturierung der Daten statt, es wird kein Fitting einer Modellfunktion vorgenommen. Vielmehr werden die Daten des Lernsatzes selbst

zur Vorhersage verwendet. KNN-Verfahren sind besonders dann geeignet, wenn die Daten des Lernsatzes nur in geringem Maße strukturierbar sind. Sie können aber kaum zum Verständnis tatsächlicher Zusammenhänge herangezogen werden.

Sei  $1 \leq k \leq n$ . Zur Bestimmung eines Funktionswertes  $f(\mathbf{x})$  für  $\mathbf{x} \in \mathbb{R}^n$  werden zunächst die Abstände der  $\mathbf{x}_i$ ,  $i \in m$  zu  $\mathbf{x}$  bestimmt und in aufsteigender Reihenfolge angeordnet:

$$\|\mathbf{x} - \mathbf{x}_{i_0}\| \leq \dots \leq \|\mathbf{x} - \mathbf{x}_{i_{k-1}}\| \leq \dots$$

Dabei ist es sinnvoll, dass die Vorhersagevariablen in autoskalierter Form vorliegen.  $\|\cdot\|$  kann eine beliebige Norm auf  $\mathbb{R}^n$  sein. Wir werden in dieser Arbeit die euklidische Norm  $\|\cdot\|_2$  verwenden. Als nächster Schritt wird die Menge der  $k$  nächsten Nachbarn zu  $\mathbf{x}$  bestimmt:

$$N_k(\mathbf{x}) = \{i_l \mid l \in k\}.$$

Ist dies nicht eindeutig möglich, so wird per Zufall entschieden. Bei KNN-Regression berechnet man den Funktionswert der Vorhersagefunktion als arithmetisches Mittel aus den Eigenschaftswerten der  $k$  nächsten Nachbarn zu  $\mathbf{x}$

$$f(\mathbf{x}) = \frac{1}{k} \sum_{i \in N_k(\mathbf{x})} y_i.$$

Bei KNN-Klassifikation wird die vorherzusagende Klasse per Mehrheitsentscheidung unter den  $k$  nächsten Nachbarn von  $\mathbf{x}$  ausgewählt:

$$f(\mathbf{x}) = \operatorname{argmax}_{c \in \mathcal{C}} \sum_{i \in N_k(\mathbf{x})} \delta(c, y_i).$$

Ist keine eindeutige Klassenbestimmung möglich, wird wiederum per Zufall entschieden. Alternativen bilden gewichtete Verfahren, die etwa in Pattsituationen zugunsten der Klasse des nächstgelegenen Nachbarn entscheiden. Wir werden in dieser Arbeit KNN-Klassifikation nur im binären Fall verwenden (s. Abschnitt 4.4.3). Dann können Zufallsentscheidungen vermieden werden, indem man ausschließlich ungerade  $k$  zulässt. Die Bestimmung des günstigsten Parameters  $k$  kann beispielsweise durch CV oder eine Teststichprobe erfolgen.





**Teil II**

**Anwendungen in der Chemie**



# Kapitel 4

## Kombinatorische Chemie

Der traditionelle Weg der chemischen Synthese bestand darin, jeweils einzelne chemische Verbindungen gezielt herzustellen. Motivation zur Synthese waren oft Vermutungen über bestimmte erwünschte physiko-chemische oder biologisch-pharmazeutische Eigenschaften, die man für diese Verbindungen annahm.

Im Zuge zunehmender Automatisierung und nicht zuletzt durch rasante Entwicklungen auf dem Gebiet der elektronischen Datenverarbeitung hat eine neue Methode der chemischen Synthese begonnen, die traditionelle Vorgehensweise abzulösen. Verfahren der *kombinatorischen Chemie* ermöglichen es, eine Vielzahl von Verbindungen in einem Arbeitsgang zu synthetisieren, und stellen damit ein mächtiges Werkzeug auf der Suche nach neuen Materialien oder Arzneimitteln dar.

Technisch wird dies beispielsweise durch Syntheseroboter realisiert, die Multititerplatten mit bis zu mehreren hundert miniaturisierten Reagenzkörpern innerhalb kürzester Zeit befüllen. Das Ergebnis sind so genannte *Bibliotheken* chemischer Verbindungen. Man spricht in diesem Zusammenhang auch von *kombinatorischen* Bibliotheken, da die Reaktanden auf verschiedene bzw. alle möglichen Weisen kombiniert werden.

Nach der Synthese erfolgt die Untersuchung der Produkte hinsichtlich bestimmter erwünschter Eigenschaften. Man bezeichnet diesen Arbeitsgang als *Screening*. Oft kann auch das Screening automatisiert ablaufen, so dass insgesamt hohe Durchsatzraten neu synthetisierter und „gescreenter“ Substanzen realisiert werden können. Entsprechend bezeichnet man solche Verfahren als *Hochdurchsatz-Screening* (engl. *High Throughput Screening*, kurz *HTS*).

## 4.1 Optimierung kombinatorisch–chemischer Experimente

Meist ist es dennoch nicht möglich, oder aber mit sehr hohen Kosten verbunden, kombinatorische Bibliotheken in ihrer gesamten Variabilität zu synthetisieren. Wir werden in Abschnitt 4.2 Beispiele betrachten, wo schon aus 10 bis 20 Reaktanden Bibliotheken mit mehreren zehntausend Verbindungen resultieren. Man wird also zunächst nur eine Teilmenge aller kombinatorisch möglichen Verbindungen, eine *reale Bibliothek* tatsächlich synthetisieren und screenen. Diese ist Teilmenge einer Gesamtheit der kombinatorisch möglichen Verbindungen, die wir als *virtuelle Bibliothek* bezeichnen.

Abbildung 4.1 beschreibt diese Situation und zeigt die Vorgehensweise, wie man in der virtuellen Bibliothek mit höherer Wahrscheinlichkeit Kandidaten findet, die eine gesuchte Eigenschaft besitzen. Mathematisch–algorithmische Teilprobleme sind hierbei grau unterlegt. Dazu zählt

- die Generierung der virtuellen Bibliothek (Abschnitt 4.2).

Je nach Situation muss

- die Bestimmung einer realen Teilbibliothek hoher Diversität (Abschnitt 4.5.1) oder
- die Überprüfung der Teilmengenrelation von realer und virtueller Bibliothek (Abschnitt 4.5.2)

erfolgen. Weitere Schritte umfassen

- die Berechnung molekularer Deskriptoren (Abschnitt 4.3) und
- die Bestimmung von Vorhersagefunktionen über statistische Lernverfahren (Kapitel 3) sowie
- die Anwendung der Vorhersagefunktion.

Zunächst werden Strukturen der realen Bibliothek mittels *molekularer Deskriptoren* auf reelle Zahlen abgebildet. Für jede Struktur erhält man einen Vektor reeller Zahlen gleicher Länge. Zudem hat man für jede Struktur einen Eigenschaftswert. Deskriptoren- und Eigenschaftswerte bilden die Eingabe für statistische Verfahren des überwachten Lernens. Diese liefern eine Vorhersagefunktion, die als Eingabe einen Vektor von Deskriptorenwerten erwartet, und als Ausgabe einen vorhergesagten Wert für die Eigenschaft liefert. Die Vorhersagefunktion wird dabei so bestimmt, dass gemessene und berechnete Eigenschaftswerte für die reale Bibliothek möglichst gut

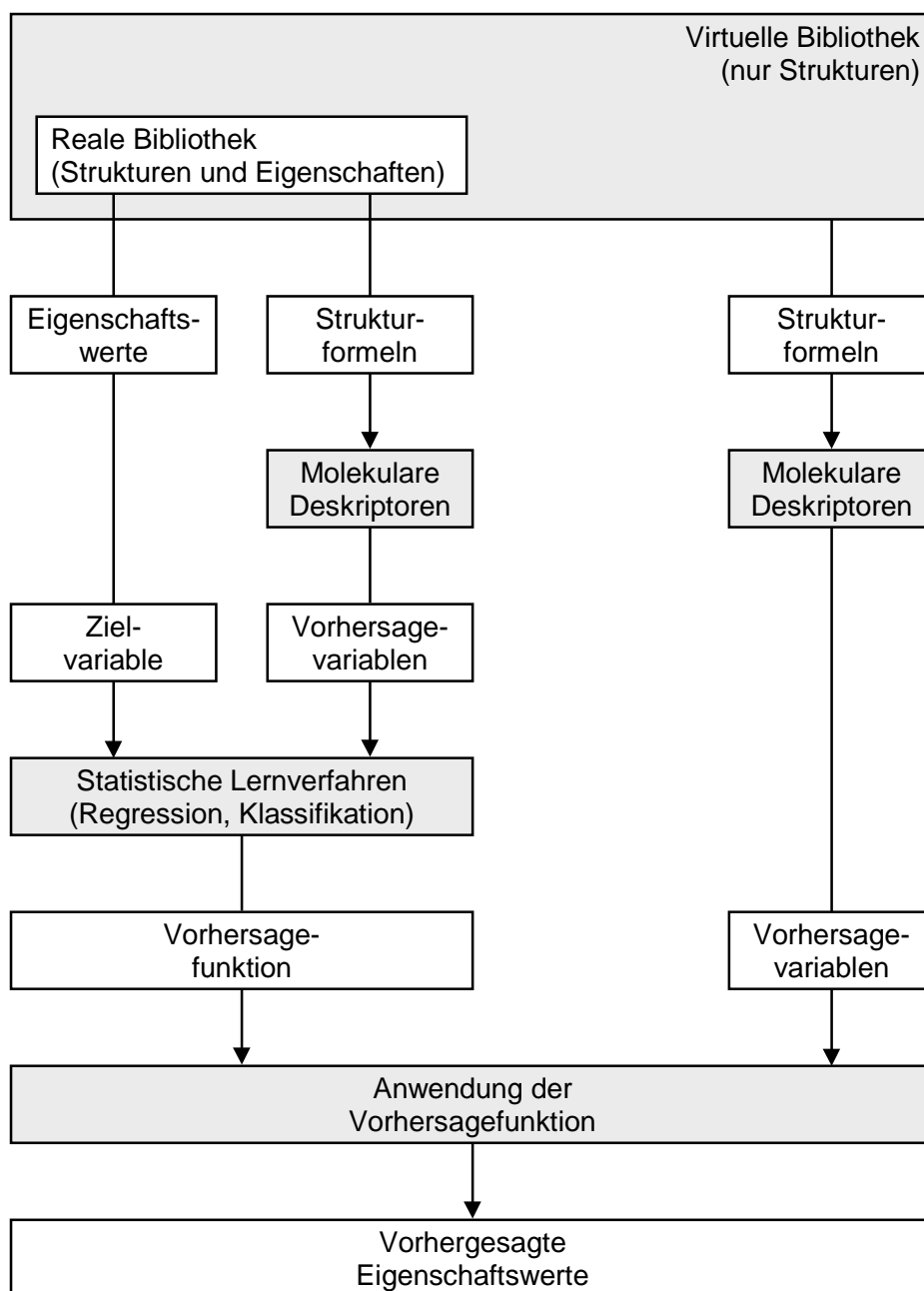


Abbildung 4.1: Vorgehensweise bei der Vorhersage von Eigenschaften für virtuelle kombinatorische Bibliotheken

übereinstimmen (vgl. Kapitel 3). Dieser gesamte Prozess der Deskriptorenberechnung und der Suche nach geeigneten Vorhersagefunktionen ist Gegenstand der SAR/QSAR/QSPR-Forschung.

Mit Hilfe eines Strukturgenerators wird die virtuelle Bibliothek konstruiert. Für deren Elemente kann man nun ebenfalls Deskriptorenwerte berechnen. Eingesetzt in die Vorhersagefunktion können somit Eigenschaftswerte für die virtuelle Bibliothek prognostiziert werden. Man spricht hierbei auch von *virtuellem Screening*. Strukturen, für die gute Eigenschaftswerte vorhergesagt werden, sind als Kandidaten für eine gezielte Synthese auszuwählen.

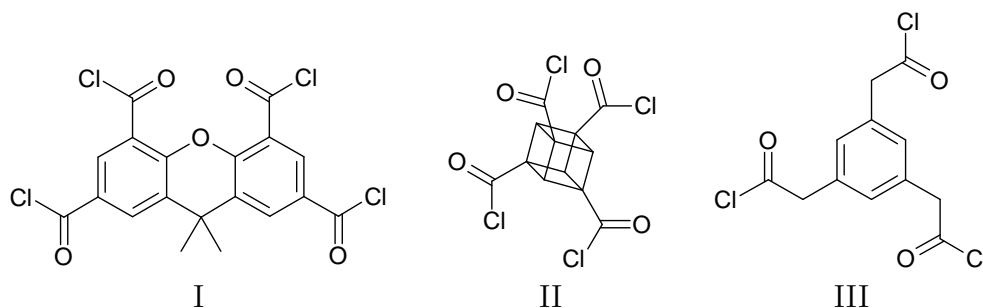
## 4.2 Generierung kombinatorischer Bibliotheken

In Abschnitt 2.2 wurden bereits Verfahren zur reaktionsbasierten Strukturgenerierung beschrieben. Im Folgenden wollen wir zwei typische Beispiele kombinatorischer Bibliotheken vorstellen.

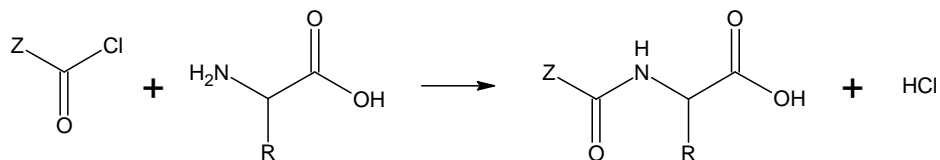
### 4.2.1 Beispiel: Amidierung von Säurechloriden mit Aminosäuren

Oft tritt bei kombinatorisch-chemischen Experimenten die Situation auf, dass an ein so genanntes *Zentralmolekül* mit mehreren *reaktiven Stellen* verschiedene *Liganden* angelagert werden (Abschnitt 2.2.1). Für diesen Fall gibt es ein sehr effizientes Verfahren, bei dem das Konstruktionsproblem auf die Generierung von Doppelnebenklassen zurückgeführt wird [57, 164].

Diese Vorgehensweise wollen wir anhand eines Beispiels von T. Carell [24, 25] demonstrieren. Als Zentralmoleküle werden verschiedene Säurechloride verwendet, Xanthentetracarbonsäurechlorid (I), Cubantetracarbonsäurechlorid (II) und Benzoltriessigsäurechlorid (III):



Diese wurden mit Aminosäuren (Abbildung 4.2) gekoppelt. Dabei reagiert der Säurechlorid-Rest mit der Aminogruppe in  $\alpha$ -Stellung zur Carboxylgruppe:



Bei einer genauen Betrachtung der Aminosäuren erkennen wir, dass bei Prolin das N-Atom nur mit einem Wasserstoff gebunden ist. Damit die Synthesereaktion auch bei Prolin durchgeführt werden kann, verzichten wir für das

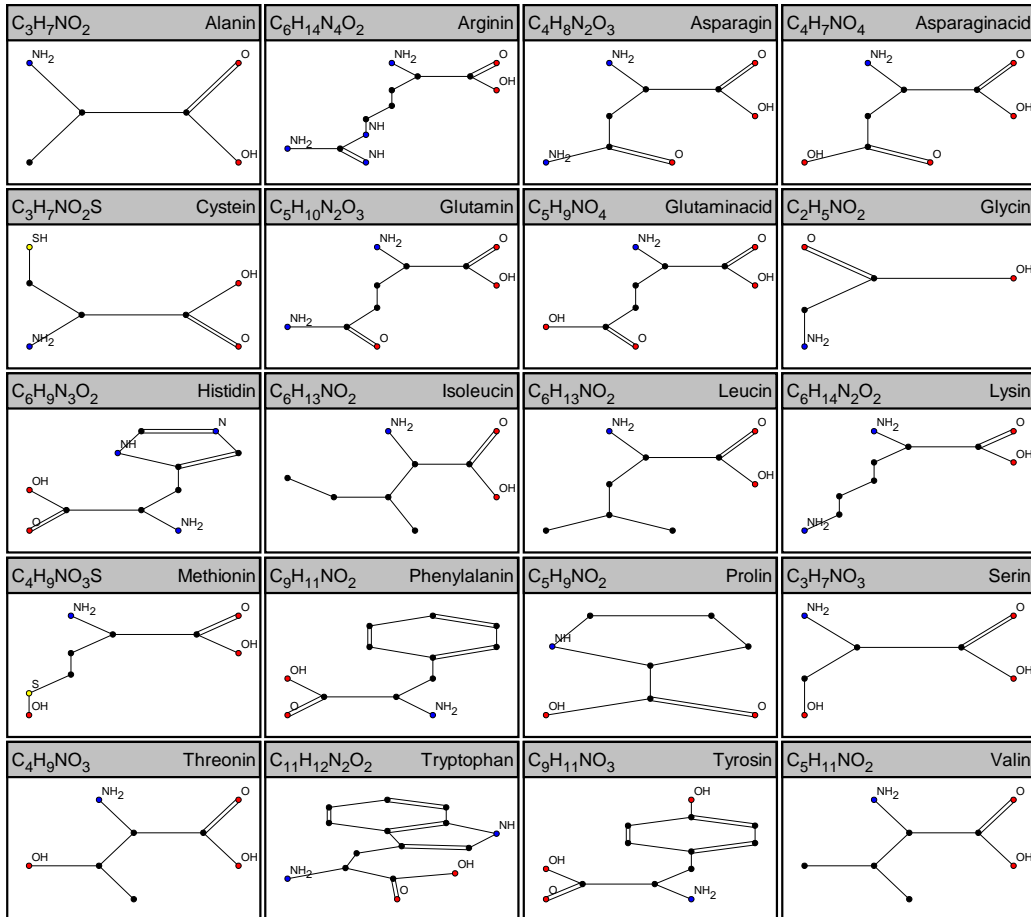
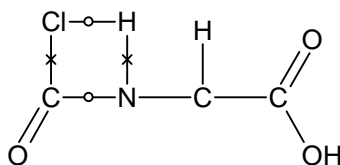


Abbildung 4.2: Strukturformeln von 20 natürlichen Aminosäuren

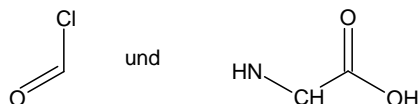
Reaktionsschema auf den zweiten Wasserstoff an dem N-Atom:



während der Reaktion neu geschlossene Bindungen sind dabei durch Kreise „o“, gebrochene Bindungen durch Kreuze „x“ markiert. Bei einer Synthesereaktion hat die Reaktionssubstruktur mindestens zwei Zusammenhänge-



komponenten. Im vorliegenden Beispiel sind dies



Sie werden zunächst in den Reaktanden gesucht. Der Säurechlorid-Rest wird in den Zentralmolekülen I und II  $k = 4$  mal, in III  $k = 3$  mal gefunden. Für die rechts abgebildete Zusammenhangskomponente der Reaktionssubstruktur gibt es in jeder Aminosäure genau eine Einbettung. Bei Berücksichtigung der Nummerierung des Zentralmoleküls  $M$  gibt es  $a^k$  Möglichkeiten, die  $k$  reaktiven Stellen mit  $a$  verschiedene Aminosäuren zu belegen. Unter Vernachlässigung der Nummerierung von  $M$  sind einige dieser Belegungen äquivalent. Die Automorphismengruppe  $\text{Aut}(M)$  operiert auch auf den durch die reaktiven Stellen von  $M$  definierten Atomen. Man erhält eine Untergruppe  $G \leq S_k$ , die auf den reaktiven Stellen operiert. Die unter Vernachlässigung der Nummerierung verschiedenen Belegungen sind gegeben durch die Bahnen

$$a^k // G.$$

Wir erhalten die gesuchte kombinatorische Bibliothek, indem wir ein Repräsentantensystem von  $a^k // G$  erzeugen, und dann für jeden Bahnrepräsentanten  $f \in \text{rep}(a^k // G)$  das Reaktionsschema der Amidierung an der durch  $f$  beschriebenen aktiven Stelle mit der ebenfalls durch  $f$  spezifizierten Aminosäure anwenden.

Für Zentralmolekül I besteht die auf die aktiven Stellen induzierte Untergruppe lediglich aus der Identität und einer Achsensymmetrie, bei Zentralmolekül II ist  $G = S_4$ , bei III die zyklische Gruppe  $C_3$ .

Tabelle 4.1 enthält die gesamte Anzahl  $a^4$  bzw.  $a^3$  möglicher Belegungen aller aktiven Stellen mit  $a$  verschiedenen Aminosäuren sowie die Anzahl nicht isomorpher Belegungen unter Berücksichtigung der Automorphismengruppen von Zentralmolekül I, II und III. Dabei sei bemerkt, dass zur Bestimmung der Anzahlen wesentlich verschiedener Einbettungen die Bibliotheken nicht *konstruiert* werden müssen, sondern mit Hilfe des Lemmas von Cauchy-Frobenius *abgezählt* werden können.

Besondere Aufmerksamkeit muss Zentralmolekül II gewidmet werden. Betrachtet man dieses als Objekt im dreidimensionalen Raum, so dürfen Spiegelungen nicht berücksichtigt werden. Für die aktiven Stellen bedeutet dies, dass anstelle von  $S_4$  nur die alternierende Gruppe  $A_4$  betrachtet werden darf. Die Berechnung der geometrischen Automorphismen wird durch [2] ermöglicht. Die Anzahlen wesentlich verschiedener Belegungen unter Berücksichtigung von  $A_4$  zeigt Tabelle 4.1 in Spalte II'.

$a$	$a^4$	I	II	II'	$a^3$	III
1	1	1	1	1	1	1
2	16	10	5	5	8	4
3	81	45	15	15	27	11
4	256	136	35	36	64	24
5	625	325	70	75	125	45
6	1296	666	126	141	216	76
7	2401	1225	210	245	343	119
8	4096	2080	330	400	512	176
9	6561	3321	495	621	729	249
10	10000	5050	715	925	1000	340
11	14641	7381	1001	1331	1331	451
12	20736	10440	1365	1860	1728	584
13	28561	14365	1820	2535	2197	741
14	38416	19306	2380	3381	2744	924
15	50625	25425	3060	4425	3375	1135
16	65536	32896	3876	5696	4096	1376
17	83521	41905	4845	7225	4913	1649
18	104976	52650	5985	9045	5832	1956
19	130321	65341	7315	11191	6859	2299
20	160000	80200	8855	13700	8000	2680

Tabelle 4.1: Amide aus  $a$  Aminosäuren und den verschiedenen Zentralmolekülen I, II und III

Eine weitere Problemstellung besteht darin, alle Strukturen zu generieren, in denen die verschiedenen Bausteine mit einer vorgegebenen Vielfachheit auftreten. Sei  $f \in a^k$  gegeben. Mit

$$a^k //_f G$$

bezeichnen wir die Bahnen mit gleichem *Gewicht* wie  $f$ , d.h. die Werte  $i \in k$  mit Vielfachheit  $|f^{-1}(i)|$  annehmen. Vermöge der Bijektion [117, 118]

$$a^k //_f G \longrightarrow G \backslash S_k / (S_k)_f, \quad G(f \circ \pi) \longmapsto G\pi(S_k)_f$$

können wir das Konstruktionsproblem lösen, indem wir eine Transversale von Doppelnebenklassen generieren. Der Stabilisator  $(S_k)_f$  von  $f$  in  $S_k$  ist gerade die direkte Summe der symmetrischen Gruppen  $S_{f^{-1}(i)}$ ,  $i \in k$ :

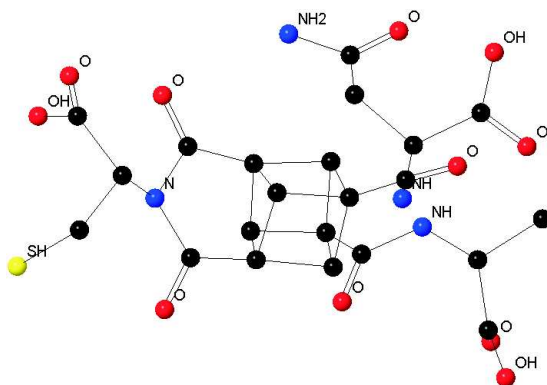
$$(S_k)_f = \bigoplus_{i \in k} S_{f^{-1}(i)}.$$

Zu Konstruktion von Doppelnebenklassen können vorhandene Lösungen herangezogen werden. So wurde diese Aufgabe bereits in [127] mit Hilfe von Untergruppenleitern und in [52] mit ordnungstreuer Erzeugung gelöst. In der Software *MOLGEN-COMB* [59] wird letzteres Prinzip verwendet, da es bei vergleichbarem Zeitaufwand mit deutlich geringerem Speicherbedarf auskommt.

Ist man wiederum nur an der Abzählung von  $a^k //_f G$  interessiert, so kann man dieses Problem mit Hilfe der Pólya-Theorie [106] lösen, wie es beispielsweise in [164] und [57] mit Hilfe des Computeralgebra-Systems SYMMETRICA [73] demonstriert wurde.

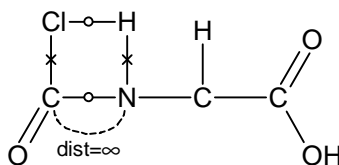
### Konstruktion nach dem Netzwerkprinzip

Schließlich wollen wir überprüfen, ob die Strukturgenerierung nach dem Netzwerkprinzip aus Abschnitt 2.2.2 dasselbe Ergebnis liefert. Erstaunlicherweise erhalten wir für 20 Aminosäuren (ohne Einschränkung der Vielfachheiten, unter Verwendung der topologischen Automorphismengruppe) 13035 Lösungen. Ein Blick auf eine der generierten Strukturen verdeutlicht, was geschehen ist:



Da wir bei unserem Reaktionsschema auf den zweiten Wasserstoff am N-Atom verzichten mussten, können bereits angelagerte Aminosäuren mit einer zweiten reaktiven Stelle an dem Zentralmolekül reagieren. Auf diese Weise kann es zu Ringschlüssen kommen, wie wir in der obigen Struktur auf der linken Seite am Cystein sehen. Wir wollen an dieser Stelle nicht erörtern, ob diese Reaktion chemisch sinnvoll ist. Mathematisch können wir diesen Reaktionsverlauf ausschließen, indem wir die Reaktionssubstruktur mit einer Substruktur-Restriktion vom Typ Distanz versehen, etwa zwischen dem C-

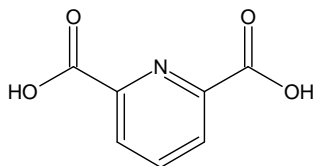
Atom auf Seiten des Säurechlorid-Rests und dem Stickstoff:



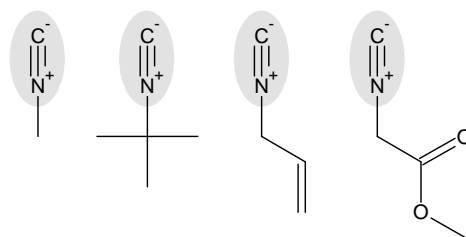
Auf diese Weise stellen wir sicher, dass der Abstand zwischen diesen beiden Atomen  $\infty$  ist, sie also vor der Reaktion nicht schon in der gleichen Zusammenhangskomponente liegen dürfen. Mit dem so modifizierten Reaktionsschema liefert auch der Netzwerk-Generator wie erwartet eine Bibliothek von 8855 Verbindungen.

#### 4.2.2 Beispiel: Ugis Siebenkomponentenreaktion

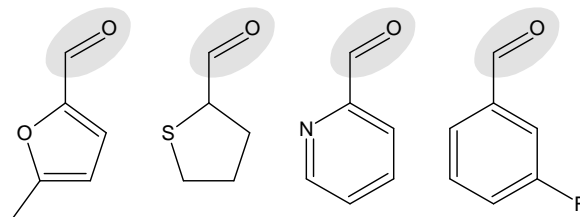
Nicht immer können virtuelle kombinatorische Bibliotheken so einfach wie im letzten Abschnitt beschrieben werden. Oft gehören die möglichen Reaktanden verschiedenen Substanzklassen an, die über komplizierte Mechanismen miteinander reagieren. Als Beispiel betrachten wir die Siebenkomponentenreaktion nach I. Ugi [145]. Zentralmolekül ist dabei Pyridin-2,6-dicarbonsäure:



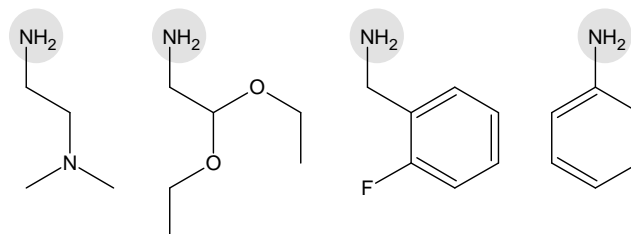
Als weitere Bausteine werden Isocyanide  $C^- \equiv N^+ - R^1$



Aldehyde  $O=CH-R^2$

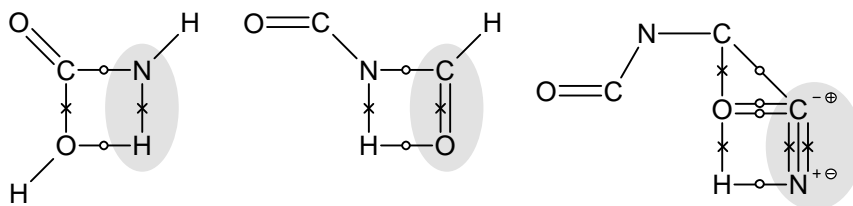


und Amine  $\text{H}_2\text{N}-\text{R}^3$



zugegeben. Die für die Substanzklassen typischen Substrukturen  $\text{C}^- \equiv \text{N}^+$ ,  $\text{O}=\text{CH}$  und  $\text{NH}_2$  sind dabei jeweils grau unterlegt.

In [164] wurde ein schrittweises Verfahren zur Konstruktion der durch Ugis Siebenkomponentenreaktion beschriebenen kombinatorischen Bibliothek vorgeschlagen. Abbildung 4.3 zeigt, wie dabei an das Pyridin-Grundgerüst sukzessive die Bausteine aus den drei Substanzklassen Amine, Aldehyde und Isocyanide angelagert werden. Die verwendeten Reaktionsschemata sind:



Wie gehabt sind dabei Bindungsbrüche durch Kreuze „ $\times$ “ und neu geschlossene Bindungen durch Kreise „ $\circ$ “ symbolisiert. Ladungsänderungen werden durch „ $\oplus$ “ und „ $\ominus$ “ dargestellt. Atome des Reaktionszentrums auf Seiten der Bausteine sind der besseren Überschaubarkeit wegen grau unterlegt.

Vernachlässigt man zunächst die Aromatizität des Pyridin-Grundgerüsts, so erhält man nach dem ersten Schritt  $4^2 = 16$ , nach dem zweiten Schritt  $16 \cdot 4^2 = 256$  und nach dem dritten Schritt  $256 \cdot 4^2 = 4096$  Strukturen. Anschließende Identifizierung der aromatischen Bindungen, kanonische Nummerierung und Entfernung von Dubletten liefern 2080 Verbindungen für die virtuelle Bibliothek.

Alternativ kann man auch von Beginn an die Aromatizität des Pyridin-Grundgerüsts berücksichtigen. Dann gibt es nach dem ersten Schritt 4 Strukturen mit symmetrischer Belegung der Substituenten  $\text{R}^1$  und  $\binom{4}{2} = 6$  mit asymmetrischer Belegung. Nach dem zweiten Schritt erhält man  $4 \cdot 4 = 16$  symmetrische und

$$6 \cdot 4 + 4 \cdot \binom{4}{2} + 6 \cdot \binom{4}{2} \cdot 2 = 120$$

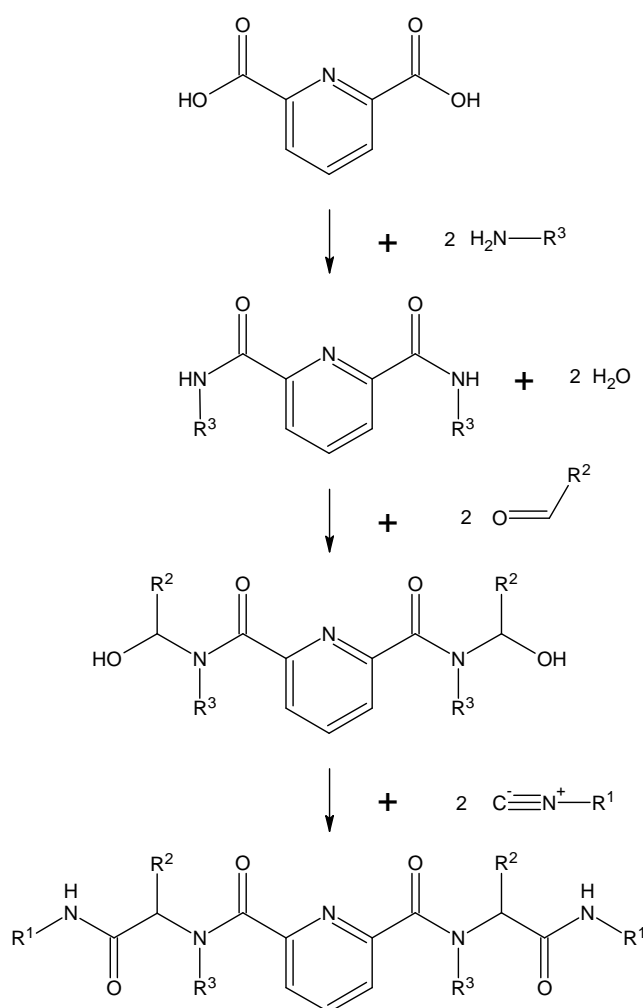


Abbildung 4.3: Schrittweise Darstellung der Siebenkomponentenreaktion

asymmetrische Belegungen von  $R^1$  und  $R^2$ . Letztere setzen sich zusammen aus asymmetrischen Belegungen von  $R^1$  und symmetrischen Belegungen von  $R^2$ , symmetrischen Belegungen von  $R^1$  und asymmetrischen Belegungen von  $R^2$  sowie asymmetrischen Belegungen von  $R^1$  und asymmetrischen Belegungen von  $R^2$ .

Nach dem dritten Schritt ergeben sich  $16 \cdot 4 = 64$  symmetrische Belegungen für  $R^1$ ,  $R^2$  und  $R^3$  sowie

$$120 \cdot 4 + 16 \cdot \binom{4}{2} + 120 \cdot \binom{4}{2} \cdot 2 = 2016$$

asymmetrische Belegungen. Insgesamt haben wir wie zuvor  $64 + 2016 = 2080$

Verbindungen in unserer virtuellen Bibliothek.

In Abschnitt 8.3 von [164] „Abzählen und Screenen einer Bibliothek“ wird gezeigt, wie auch diese Situation der „mehrstufigen Anlagerung an ein Zentralkmolekül“ als Konstruktionsproblem von Symmetrieklassen formuliert und auf beliebige Grundgerüste verallgemeinert werden kann. Wir wollen uns in den nächsten Abschnitten auf das Screening virtueller Bibliotheken konzentrieren.

### 4.3 Molekulare Deskriptoren

Wie bereits zu Beginn des Kapitels erwähnt, möchten wir statistische Lernverfahren zur Bestimmung von Vorhersagefunktionen für experimentell messbare Eigenschaften chemischer Verbindungen heranziehen. Nun können aber solche Verfahren molekulare Graphen nicht direkt verarbeiten. Vielmehr erwarten sie als Eingabe für jede Beobachtung einen Vektor reeller Zahlen. Vermöge *molekularer Deskriptoren* können molekulare Graphen auf reelle Zahlen abgebildet werden. In der vorliegenden Arbeit wurde dazu eine Implementierung von J. Braun [19] zur Berechnung molekularer Deskriptoren herangezogen. Diese orientiert sich im Wesentlichen an [144] und R. Todeschini Enzyklopädie [142] über molekulare Deskriptoren. Eine notwendige Bedingung an einen molekularen Deskriptor ist die Invarianz bzgl. der Nummerierung der Knoten eines molekularen Graphen:

#### 4.3.1 Definition:

Eine Abbildung

$$D : \mathcal{M} \longrightarrow \mathbb{R}, \quad M \longmapsto D(M),$$

heißt *molekularer Deskriptor*, wenn

$$M \in \mathcal{M}_n, \pi \in S_n \implies D(M^\pi) = D(M).$$

Es gibt verschiedene Möglichkeiten, molekulare Deskriptoren zu definieren. Aufgrund der Informationen, die zur Deskriptorenberechnung verwendet werden, kann man molekulare Deskriptoren verschiedenen Kategorien zuordnen: Man unterscheidet *arithmetische*, *topologische* und *geometrische Indizes*. Diese Bezeichnungen orientieren sich an den Abstraktionsebenen aus Abschnitt 1.8, jedoch hält sich die Nomenklatur in der Literatur<sup>1</sup> nicht streng an diese Einteilung.

Manche arithmetische Indizes verwenden neben der Bruttoformel beispielsweise auch die Vielfachheiten der Bindungstypen oder der Atomzustände. Viele topologische Indizes beschränken sich nicht nur auf den, einer chemischen Verbindung zugrunde liegenden Multigraphen. Um diesbezüglich mehr Klarheit zu schaffen, untergliedern wir zusätzlich in *rein* arithmetische und *rein* topologische Indizes.

Im zweiten Teil dieses Abschnitts werden wir eine besonders variabel einsetzbare Art molekularer Deskriptoren kennen lernen, sog. *Substruktur-Vielfachheiten*. Diese Deskriptoren werden mittels einer molekularen Substruktur definiert und geben als Deskriptorenwert ihre Vielfachheit in dem untersuchten molekularen Graphen zurück.

<sup>1</sup>Generelle Referenz: *Journal of Chemical Information and Computer Sciences*, American Chemical Society, seit 1974.



### 4.3.1 Arithmetische, topologische und geometrische Indizes

Viele Indizes werden unter Vernachlässigung der H-Atome berechnet. Zu einem molekularen Graphen  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$  sei

$$\Omega := \{i \in n \mid \varepsilon(i) \neq \text{H}\}$$

im Folgenden die Menge der Nicht-Wasserstoff-Atome. Wie gehabt bezeichnet dann  $\gamma|_{\Omega}$  die Einschränkung von  $\gamma$  auf  $\Omega$ .

Je nachdem, welche Informationen verarbeitet werden, wollen wir verschiedene Typen von Indizes unterscheiden.

#### Arithmetische Indizes

Ein *rein arithmetischer Index* verwendet nur die Elementverteilung eines molekularen Graphen. Für einen rein arithmetischen Index  $D_{\text{ra}}$  gilt

$$M = (\varepsilon, \zeta, \gamma), M' = (\varepsilon, \zeta', \gamma') \in \mathcal{M} \implies D_{\text{ra}}(M) = D_{\text{ra}}(M'),$$

Rein arithmetische Indizes sind also für Verbindungen mit gleichen Bruttoformeln konstant.

#### 4.3.2 Beispiele:

Sei  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}_n$  ein molekularer Graph. Häufig benutzte rein arithmetische Indizes sind die Anzahl von Nicht-Wasserstoff-Atomen

$$D_A(M) = |\Omega|,$$

die Anzahl von Atomen des Elements  $X$

$$D_{N_X}(M) = \beta_M(X),$$

und das Molekulargewicht unter Vernachlässigung der H-Atome

$$D_{MW}(M) = \sum_{X \neq \text{H}} \bar{m}_X \beta_M(X)$$

von  $M$ . Dabei ist  $\bar{m}_X$  die mittlere Atommasse von Element  $X$  (vgl. Definition 5.8.1). Tabelle 4.2 enthält mittlere Atommassen (in *Atomic Mass Units*<sup>2</sup> amu) für die Elemente aus  $\mathcal{E}_{11}$ .

<sup>2</sup>Ein amu ist definiert als der zwölfte Teil der Masse eines <sup>12</sup>C Atoms.

Viele Indizes vernachlässigen die Kantenvielfachheiten. Zu einem Multigraphen  $\gamma$  sei der zugrunde liegende schlichte Graph definiert durch

$$\gamma^s(\{i, j\}) := \begin{cases} 1 & \text{falls } \gamma(\{i, j\}) > 0, \\ 0 & \text{sonst.} \end{cases}$$

Die Knotengrade ohne Berücksichtigung von Kantenvielfachheiten bezeichnen wir mit

$$\deg_\gamma^s(i) := \deg_{\gamma^s}(i).$$

### 4.3.3 Beispiele:

Die Anzahl von Bindungen

$$D_B(M) = \frac{1}{2} \sum_i \deg_\gamma^s(i)$$

und die zyklomatische Zahl

$$D_C(M) = D_B(M) - D_A(M) + 1$$

sind keine rein arithmetischen Indizes. Sie können für Verbindungen mit gleicher Bruttoformel (und auch gleicher Zustandsverteilung) verschiedene Werte annehmen. Aufgrund der Tatsache, dass sie lediglich Anzahlen von Bindungen berücksichtigen, werden sie dennoch den arithmetischen Indizes zugeordnet.

### Topologische Indizes

Ein *topologischer Index* (kurz *TI*) verarbeitet neben Element- und Zustandsverteilung auch die Nachbarschaftsbeziehungen der Atome. Die bekanntesten topologischen Indizes kann man zudem als *rein topologisch* klassifizieren, d.h. sie verwenden nur den  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}$  zugrunde liegenden Multigraphen  $\gamma$ . Für einen rein topologischen Index  $D_{\text{rt}}$  gilt:

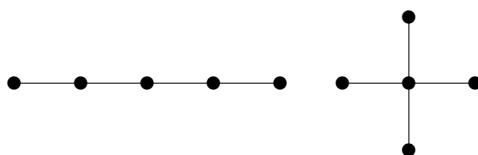
$$M = (\varepsilon, \zeta, \gamma), M' = (\varepsilon', \zeta', \gamma) \in \mathcal{M} \implies D_{\text{rt}}(M) = D_{\text{rt}}(M').$$

### 4.3.4 Beispiele:

Die erste Anwendung topologischer Indizes wurde von H. Wiener [165] entwickelt. Er verwendete den nach ihm benannten Index

$$D_W(M) = \frac{1}{2} \sum_{i \in \Omega} \sum_{j \in \Omega} \text{dist}_\gamma(i, j),$$

zur Modellierung der Siedepunkte von Alkanen (vgl. 4.4.1). Graphentheoretisch betrachtet kann der Wiener Index als ein Maß für die *Verzweigkeit* angesehen werden. Dies wird dadurch gerechtfertigt, dass  $D_W$  für Graphen mit derselben Anzahl von Knoten und Kanten maximale (minimale) Werte für minimal (maximal) verzweigte Graphen annimmt. Minimal verzweigte Graphen sind dabei Ketten (links), maximal verzweigt sind Sterne (rechts):



Zagreb Indizes [60] summieren Quadrate bzw. Produkte von Knotengraden:

$$D_{M_1}(M) = \sum_{i \in \Omega} (\deg_{\gamma|\Omega}^s(i))^2,$$

$$D_{M_2}(M) = \sum_{\{i,j\} \in E_{\gamma|\Omega}} \deg_{\gamma|\Omega}^s(i) \cdot \deg_{\gamma|\Omega}^s(j).$$

Randic Indizes [111, 79] der Ordnung  $m$  werden berechnet durch

$$D_{0\chi}(M) = \sum_{i \in \Omega} (\deg_{\gamma|\Omega}^s(i))^{-\frac{1}{2}}$$

falls  $m = 0$  und

$$D_{m\chi}(M) = \sum_{\substack{(v_0, \dots, v_m) \\ \text{Weg in } \gamma|\Omega}} \prod_{i=0}^m (\deg_{\gamma|\Omega}^s(v_i))^{-\frac{1}{2}}$$

für  $m > 0$ . Der *Knotendistanzgrad* von Knoten  $i$  in  $\gamma$  ist definiert als

$$\deg_{\gamma}^d(i) := \sum_{j \in n} \text{dist}_{\gamma}(i, j).$$

Er wird benötigt zur Berechnung des Balaban Index [4, 5]

$$D_J(M) = \frac{D_B(M)}{D_C(M) + 1} \sum_{(i,j) \in E_{\gamma|\Omega}} (\deg_{\gamma|\Omega}^d(i) \cdot \deg_{\gamma|\Omega}^d(j))^{-\frac{1}{2}}.$$

Der *Molecular Topological Index* nach Schultz [128, 129] ist definiert als

$$D_{MTI}(M) = \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{k \in \Omega} a_{ik} (a_{kj} + \text{dist}_{\gamma|\Omega}(k, j)),$$

wobei  $A_{\gamma^s|\Omega} = (a_{ij})$  die Adjazenzmatrix von  $\gamma^s|\Omega$  bezeichne. Der *Molecular Walk Count* der Länge  $k$  addiert alle Einträge der  $k$ -ten Potenz der Adjazenzmatrix von  $\gamma^s|\Omega$ :

$$D_{mwc(k)}(M) = \sum_{i \in \Omega} \sum_{j \in \Omega} a_{ij}^{(k)}, \text{ wobei } (a_{ij}^{(k)}) = (A_{\gamma^s|\Omega})^k.$$

Dabei ist zu bemerken, dass der Eintrag  $a_{ij}^{(k)}$  der  $k$ -ten Potenz der Adjazenzmatrix  $A_{\gamma^s|\Omega}$  gerade die Anzahl der Kantenzüge mit Länge  $k$  von  $i$  nach  $j$  in  $\gamma^s|\Omega$  angibt (Beweis s. [19], S. 19). Diese Indizes wurden erstmals von C. und G. Rücker [121] vorgestellt und beschreiben die Komplexität [122] eines (molekularen) Graphen. Der *Total Walk Count* summiert Molecular Walk Counts über alle Längen  $k$ :

$$D_{twc}(M) = \sum_{k \in |\Omega|} D_{mwc(k)}(M).$$

Der *maximale Eigenwert* der Adjazenzmatrix  $A_{\gamma^s|\Omega}$  kann ebenfalls als molekularer Deskriptor herangezogen werden. Wir wollen ihn mit  $D_{\lambda_1^A}$  bezeichnen.

Es gibt topologische Indizes, die nicht rein topologisch sind, sondern bei denen auch das chemische Element der Atome eine Rolle spielt.

#### 4.3.5 Beispiele:

Für einen molekularen Graphen  $M = (\varepsilon, \zeta, \gamma) \in \mathcal{M}$  ist der *Knotenvalenzgrad* von Atom  $i$  definiert als

$$\deg_M^v(i) = \frac{VE_{\varepsilon(i)} - HC_M(i)}{TE_{\varepsilon(i)} - VE_{\varepsilon(i)} - 1}.$$

$HC_M(i)$  bezeichnet dabei die Anzahl der zu  $i$  benachbarten H-Atome. Mit Hilfe des Knotenvalenzgrades werden Kier & Hall Indizes [77, 78, 79] berechnet. Ähnlich zu Randic Indizes der Ordnung  $m$  summieren Kier & Hall Indizes ebenfalls über alle Wege der Länge  $m$ , verwenden jedoch anstelle des Knotengrades den Knotenvalenzgrad:

$$D_{0\chi^v}(M) = \sum_{i \in \Omega} (\deg_M^v(i))^{-\frac{1}{2}},$$

$$D_{m\chi^v}(M) = \sum_{\substack{(v_0, \dots, v_m) \\ \text{Weg in } \gamma|\Omega}} \prod_{i=0}^m (\deg_M^v(v_i))^{-\frac{1}{2}}.$$

Eine weitere Klasse topologischer Indizes, die auch die Elemente der Atome berücksichtigt, sind Basaks informationstheoretische Indizes [9, 10]. Zur Berechnung müssen zunächst alle Atome nach ihren Elementen sowie den Bindungen und benachbarten Atomen bis zu einer Distanz  $r$  klassifiziert werden. Ergeben sich dabei  $k_r$  Klassen und seien  $n_{ri}$  die Anzahl der Atome in Klasse  $i$ , dann lassen sich folgende Indizes definieren:

$$\begin{aligned} D_{IC_r}(M) &= \sum_{i \in k_r} \frac{n_{ri}}{n} \log_2 \frac{n_{ri}}{n}, \\ D_{CIC_r}(M) &= \log_2 n - D_{IC_r}(M) \text{ und} \\ D_{SIC_r}(M) &= (\log_2 n)^{-1} D_{IC_r}(M), \end{aligned}$$

genannt Basaks Informationsgehalt, komplementärer und struktureller Informationsgehalt der Ordnung  $r$ .

Tabellen 4.5 und 4.6 enthalten berechnete Werte einiger hier beschriebener TI für die Strukturen aus Abbildung 4.6.

Im Prinzip sind dem Mathematiker beim „Erfinden“ topologischer Indizes keine Grenzen gesetzt. Eine gewisse Vielfalt verfügbarer Indizes kann sich beim Modellieren physiko-chemischer oder biologisch-pharmazeutischer Eigenschaften durchaus als hilfreich erweisen, wie wir später in Abschnitt 4.4 sehen werden.

### Geometrische Indizes

Für manche Anwendungen genügt es, nur Deskriptoren zu verwenden, die ausschließlich die Topologie der Strukturen beschreiben (siehe z.B. Abschnitt 4.4.1). Viele Eigenschaften hängen aber auch von der dreidimensionalen Gestalt der Moleküle einer chemischen Verbindung ab (vgl. Abschnitt 4.4.2). Deskriptoren, die diese Information berücksichtigen und widerspiegeln, heißen *geometrische Indizes*. Zur Berechnung geometrischer Indizes müssen die Atome der molekularen Graphen  $M \in \mathcal{M}_n$  zunächst mit dreidimensionalen Koordinaten  $\xi \in (\mathbb{R}^3)^n$  versehen werden (vgl. Abschnitt 1.6.2). Dafür stehen generell verschiedene algorithmische Lösungen zur Verfügung [123, 124]. Wir verwenden in dieser Arbeit eine Methode, die auf einem Kraftfeld-Modell nach [1] basiert.

Natürlich sollen die Werte geometrischer Deskriptoren invariant bzgl. Translationen und Drehungen sein. Deshalb wird bei geometrischen Indizes oft als erster Schritt zur Berechnung eine Zentrierung um den Koordinaten-Ursprung und eine Ausrichtung nach den Hauptachsen vorgenommen. Viele geometrische Deskriptoren beschreiben geometrische Größen wie Volumen und Oberfläche (siehe z.B. [29]) oder Durchmesser des Moleküls.

$X$	$\bar{m}_X$ [amu]	$r_X$ [Å]	$\rho_{vdw}$ [amu/Å <sup>3</sup> ]
H	1,0079	1,20	0,139
C	12,0107	1,70	0,584
N	14,0067	1,55	0,898
O	15,9994	1,52	1,088
F	18,9984	1,47	1,428
Si	28,0855	2,10	0,724
P	30,9738	1,80	1,268
S	32,0660	1,80	1,313
Cl	35,4527	1,75	1,579
Br	79,9040	1,85	3,013
I	126,9045	1,98	3,903

Tabelle 4.2: Mittlere Atommasse, Van der Waals Radius und Dichte für die Elemente aus  $\mathcal{E}_{11}$

Verbindung	$V_{vdw}$ [Å <sup>3</sup> ]	
	[119]	[71]
Methan	29,764	29,327
Ethan	47,645	47,210
Ethen	41,062	40,135
Acetylen	37,942	37,315
Benzol	84,174	87,182
Naphthalin	126,950	131,564
Cyclohexan	106,833	106,654
Chlorethan	56,402	55,690
Toluol	101,625	105,377
Brombenzol	95,538	98,771
Ethylbenzol	119,450	123,317

Tabelle 4.3: Berechnete Van der Waals Volumina einiger kleiner organischer Moleküle

#### 4.3.6 Beispiel:

Als Beispiel für einen geometrischen Index wollen wir das Van der Waals Volumen  $V_{vdw}$  genauer betrachten. Dazu wird zunächst jedem Element  $X$  ein Atomradius, der *Van der Waals Radius*  $r_X$  zugeordnet. Tabelle 4.2 enthält die Van der Waals Radien (in *Ångström* Å) für Elemente aus  $\mathcal{E}_{11}$ . In Abbildung 4.4 werden Atome der verschiedenen Elemente als Kugeln mit Van der Waals Radien gezeigt. Das Van der Waals Volumen eines Moleküls ist das Gesamtvolumen, das die Atomkugeln mit Radius  $r_{\varepsilon(i)}$  und Mittelpunkt  $\xi(i) \in \mathbb{R}^3$  einnehmen. Abbildung 4.5 zeigt ein Molekül, bei dem die Atome durch entsprechende Kugeln dargestellt sind. Prinzipiell erscheinen zwei Vorgehensweisen zur Berechnung des Van der Waals Volumens aussichtsreich:

- Der geometrische Ansatz beruht darauf, den Körper, der durch die einander überlagernden Atomkugeln bestimmt wird, exakt zu berechnen. Dabei kann man das Inklusions–Exklusions–Prinzip anwenden: Zunächst summiert man die Volumina aller Kugeln, subtrahiert dann die Durchschnitte von je zwei Kugeln, addiert wiederum die Überschneidungen von je drei Kugeln u.s.w. Allerdings bedarf bereits die Berechnung der Durchschnitte von drei Kugeln einer recht komplizierten Formel [116]. Für Durchschnitte von bis zu vier Kugeln ist eine exakte geometrische Berechnung möglich [48], für Durchschnitte höherer Ordnung ist laut [29] mathematisch bewiesen, dass sich diese im Allgemeinen nicht berechnen lassen. Natürlich könnte man die Berechnung

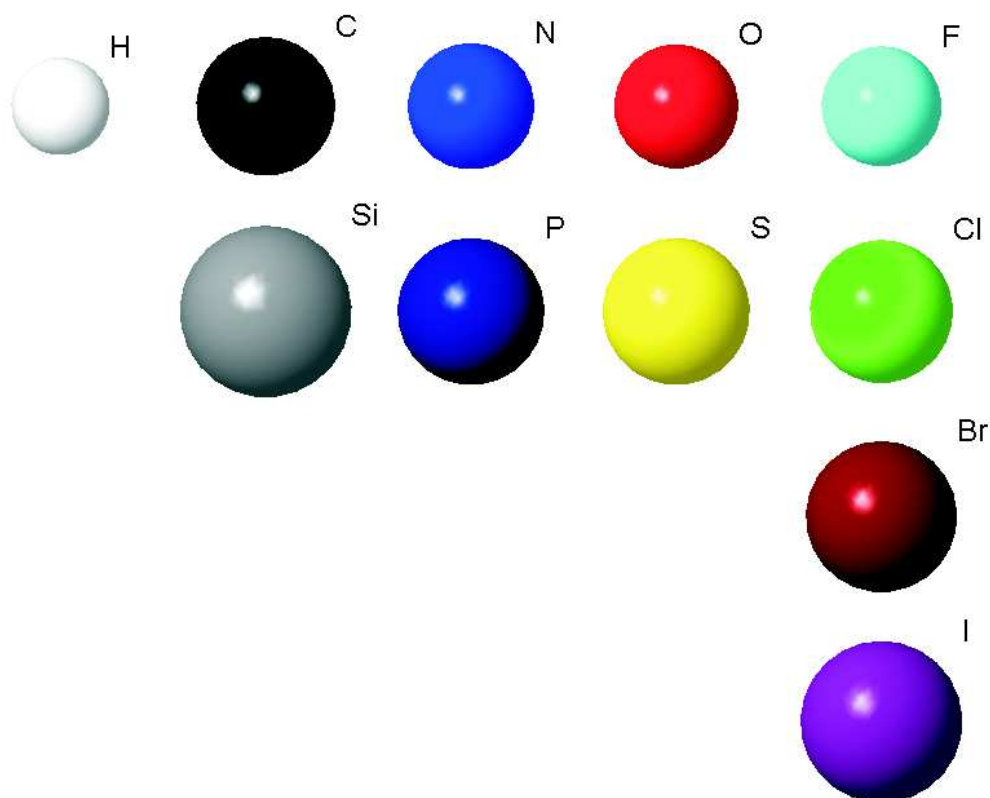


Abbildung 4.4: Atome aus  $\mathcal{E}_{11}$  dargestellt als Kugeln mit Van der Waals Radien

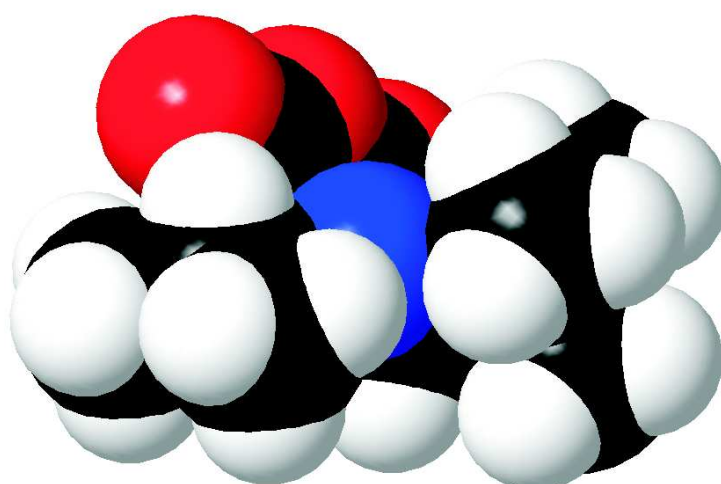


Abbildung 4.5: Darstellung eines hochsubstituiertes Bernsteinsäureanhydrids als Kalottenmodell mit Van der Waals Radien

etwa bei den Überschneidungen dritter Ordnung abbrechen und den bis dahin berechneten Wert als Näherung angeben. Ein anderer Ansatz, der von Methoden der numerischen Integration inspiriert ist, erscheint ebenfalls aussichtsreich.

- Als Ausgangspunkt eines Diskretisierungsverfahrens wird zunächst ein umschreibender Quader für das Molekül berechnet. Anschließend wird dieser in gleich große Würfel möglichst kleiner Kantenlänge zerlegt. Für jeden Würfelmittelpunkt kann man nun berechnen, ob er innerhalb oder außerhalb des durch die Atomkugeln beschriebenen Moleküls liegt. Die Summe der Volumina aller Würfel, deren Mittelpunkte innerhalb des Moleküls liegen wird als Näherung für das Van der Waals Volumen ausgegeben.

Das Problem bei dem Diskretisierungsverfahren besteht darin, dass der Rechenaufwand für kleinere Würfel beträchtlich ansteigt, insbesondere wenn das Volumen großer Moleküle berechnet werden soll. Eine wesentliche Verbesserung konnte durch J. Braun [119] erzielt werden. Dabei wird in dem umschreibenden Quader zunächst ein grobes äquidistantes Raster eingeführt, wobei die Kantenlänge eines Würfels dem Van der Waals Durchmesser eines Wasserstoff-Atoms entspricht. Auf diesem Raster wird zunächst für jeden Würfel festgestellt, welche Atomkugeln mit diesem überlappen. Würfel, die mit keinen Atomkugeln überlappen, scheiden für die folgenden Betrachtungen aus. Die verbleibenden Würfel werden jeweils nur von einer relativ kleinen Anzahl von Atomen überlappt. Für diese können wir fortfahren wie oben beschrieben: Wir führen ein feineres Raster ein (Kantenlänge  $0,01 \text{ \AA}$ ), und summieren die Volumina aller Würfel, deren Mittelpunkte innerhalb des Moleküls liegen.

Tabelle 4.3 enthält in der ersten Spalte auf diese Weise berechnete Van der Waals Volumen (in  $\text{\AA}^3$ ) für einige kleine organische Verbindungen. Zum Vergleich sind in der zweiten Spalte für die gleichen Verbindungen unter Verwendung derselben 3D-Platzierungen Van der Waals Volumen angegeben, die mit der Software *CODESSA* [71] berechnet wurden.

Die zweistufige Vorgehensweise brachte einen entscheidenden Fortschritt in Hinblick auf Reichweite und Genauigkeit. Der Volumen-Deskriptor findet insbesondere Anwendung in der Modellierung physikalischer Dichten chemischer Verbindungen, wie sie in Abschnitt 4.4.2 dargelegt wird.

Die oben vorgestellten Deskriptoren bilden nur einen verschwindend geringen Anteil der im Rahmen des *MOLGEN*-Projektes implementierten Indizes. Eine umfassendere Liste ist in Anhang B zu finden, eine detaillierte Spezifikation bietet [119].



### 4.3.2 Substruktur–Vielfachheiten

Eine weitere Möglichkeit, molekulare Deskriptoren zu definieren besteht darin, das Vorhandensein oder die Vielfachheit molekularer Substrukturen zu bestimmen. Zu einer molekularen Substruktur  $S$  wird dann der Deskriptorenwert für einen molekularen Graphen  $M$  als

$$D_S^b(M) = \begin{cases} 1 & \text{falls } S \subseteq M, \\ 0 & \text{sonst,} \end{cases} \quad \text{bzw.} \quad D_S(M) = |\text{Emb}_{\subseteq}(S, M)|$$

berechnet. Im ersten Fall spricht man von einem *binären molekularen Deskriptor*. K. Varmuzas Software *ToSim* [131] verwendet beispielsweise einen Vektor binärer molekularer Deskriptoren, um molekulare Graphen auf Ähnlichkeit zu untersuchen. Wir werden im Folgenden die an Information reichere *Substruktur–Vielfachheit* (engl. *Substructure Count*, kurz *SC*) verwenden.

#### Substruktur–Relationen und Abzählung der Einbettungen

Für die Definition auf Substrukturen basierter Deskriptoren gibt es zahlreiche Variationsmöglichkeiten: So kann man die Substrukturrelation „ $\subseteq$ “ durch die Teilstrukturrelation „ $\subseteq^i$ “ ersetzen (vgl. Definition 1.2.21). Bei der Berechnung der Vielfachheiten ist es üblich, Einbettungen, die sich nur in der Zuordnung von Wasserstoffatomen unterscheiden, lediglich einmal zu zählen. Allgemeiner könnte man natürlich alle verschiedenen Möglichkeiten aus Beispiel 1.2.23 heranziehen, um Substruktur–Vielfachheiten zu bestimmen. Chemisch am sinnvollsten ist es, nur die Symmetrie der Substruktur  $S$ , nicht aber die des molekularen Graphen  $M$  zu berücksichtigen. Für unsere weiteren Untersuchungen werden wir dieser Variante den Vorzug geben. Wasserstoffatome werden nicht berücksichtigt.

#### Auswahl der Substrukturen

Ein wichtiger Aspekt bei der Verwendung substrukturbasierter Deskriptoren betrifft die Auswahl der Substrukturen. Üblicherweise ist, wie etwa in [131], eine feste Menge von Substrukturen vorgegeben. Diese Vorgehensweise stößt dann an ihre Grenzen, wenn sich die zu untersuchenden chemischen Verbindungen nicht deutlich genug in diesen Substrukturen unterscheiden, oder es Substrukturen gibt, mit deren Hilfe sich die strukturellen Unterschiede deutlicher ausdrücken lassen.

Abhilfe schafft die Möglichkeit, benutzerdefinierte Substrukturen zuzulassen. Diese Option steht bei der im Rahmen dieser Arbeit entstandenen Software *MOLGEN–QSPR* zur Verfügung. Allerdings muss dafür der Benutzer die

zu untersuchende Bibliothek zunächst gut kennen lernen, was mit einem beträchtlichen Zeitaufwand verbunden ist. Dieser kann in erheblichem Maße durch gezielten Rechnereinsatz reduziert werden.

Algorithmus 4.3.7 findet zu einer gegebenen Bibliothek  $\mathcal{L}$  alle Substrukturen (nebst Vielfachheiten), die in der Bibliothek auftreten. Dabei wird ein assoziativer Speicher  $Map$  verwendet, der die Substrukturen auf Vektoren natürlicher Zahlen abbildet, welche zum Zählen der Vielfachheiten vorgesehen sind.

#### 4.3.7 Algorithmus: *SubstrCounts*( $\mathcal{L}$ )

```
(1)  for each  $M_i \in \mathcal{L}$ 
(2)      for each  $S \subseteq M_i$ 
(3)           $S \leftarrow \kappa(S)$ 
(4)           $Map[S][i] \leftarrow Map[S][i] + 1$ 
(5)      end
(6)  end
```

Zeile (1) durchläuft die gesamte Bibliothek mit  $M_i$ . In der Praxis wird man in Zeile (2) die Substruktur  $S$  bezüglich ihrer Größe beschränken, beispielsweise durch minimale und maximale Anzahl von Kanten. In Zeile (3) wird  $S$  kanonisch nummeriert und in Zeile (4) die Vielfachheit von  $S$  für  $M_i$  inkrementiert. Tritt dabei eine Substruktur erstmals auf, wird sie in  $Map$  eingefügt, und bekommt zunächst einen Vektor mit Einträgen Null zugeordnet. Am Ende enthält  $Map[S][i]$  die Vielfachheit von  $S$  in  $M_i$ .

#### 4.3.8 Beispiel:

Abbildung 4.6 zeigt eine Bibliothek von 50 *Decanen*, die wir in Abschnitt 4.4.1 zur Suche eines QSPR-Modells für *Siedepunkte* (engl. *Boiling Point*, kurz *BP*) verwenden werden. Decane sind Isomere mit der Bruttoformel  $C_{10}H_{22}$ . Insgesamt gibt es 75 Konstitutionsisomere zu dieser Summenformel. Substrukturen mit keiner oder einer Kante haben jeweils konstante Vielfachheiten (10 bzw. 9). In Abbildung 4.7 sind alle Substrukturen mit 2 bis 6 Kanten, die in der Bibliothek auftreten, zusammengestellt. Tabelle 4.4 gibt die entsprechenden Vielfachheiten an. Dabei sind die Substrukturen in den Spalten und die Verbindungen in den Zeilen gemäß ihren Nummerierungen in den Abbildungen eingetragen.

Im Gegensatz zu statischen Substruktur-Vektoren bietet die oben vorgestellte dynamische Variante eine natürliche Vorgehensweise zur Auswahl der Substrukturen. Für homogene Bibliotheken, wie den hier gezeigten Decanen,

BP:136,0	1	BP:145,0	2	BP:146,0	3	BP:147,0	4	BP:147,6	5
BP:147,7	6	BP:148,5	7	BP:148,7	8	BP:149,7	9	BP:151,5	10
BP:152,5	11	BP:152,8	12	BP:153,7	13	BP:154,0	14	BP:154,5	15
BP:154,5	16	BP:155,5	17	BP:155,5	18	BP:156,0	19	BP:157,0	20
BP:157,5	21	BP:157,8	22	BP:158,3	23	BP:158,8	24	BP:158,8	25
BP:159,0	26	BP:159,0	27	BP:159,5	28	BP:159,5	29	BP:159,8	30
BP:160,0	31	BP:160,0	32	BP:160,1	33	BP:160,6	34	BP:160,7	35
BP:162,0	36	BP:162,4	37	BP:162,5	38	BP:163,5	39	BP:163,8	40
BP:164,5	41	BP:165,1	42	BP:165,7	43	BP:166,0	44	BP:166,0	45
BP:167,0	46	BP:167,7	47	BP:168,4	48	BP:170,9	49	BP:174,0	50

Abbildung 4.6: Reale Bibliothek von Decanen mit ihren Siedepunkten

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	14	7	8	2	6	6	0	9	2	6	0	0	0	18	0	0	0	0	2	0
2	12	8	5	1	5	6	1	7	1	4	0	0	0	9	0	0	1	0	1	3
3	11	8	3	0	4	9	1	4	0	8	0	0	0	2	0	2	4	0	0	4
4	12	9	5	1	6	9	3	6	1	9	0	0	0	6	0	3	6	1	2	0
5	13	9	6	1	8	8	2	6	1	10	1	0	0	9	1	3	6	0	2	0
6	12	8	5	1	5	9	1	4	1	10	0	0	0	3	0	3	4	0	2	3
7	12	7	5	1	4	6	0	5	1	4	0	0	0	4	0	0	0	0	1	6
8	13	10	6	1	11	7	3	6	2	7	3	0	1	9	3	1	6	0	1	0
9	14	9	8	2	9	10	2	3	3	18	0	1	0	3	0	9	6	0	6	0
10	12	9	5	1	7	10	2	3	2	12	0	1	0	1	0	3	6	0	3	2
11	10	8	2	0	3	8	1	5	0	5	0	0	0	2	0	1	3	0	0	3
12	12	9	5	1	7	6	2	7	2	4	0	1	0	9	0	0	2	0	1	2
13	13	11	6	1	13	8	4	4	3	10	3	2	1	4	2	3	8	0	2	0
14	11	7	4	1	3	6	0	5	1	3	0	0	0	3	0	0	0	0	1	4
15	11	10	3	0	7	8	4	5	0	6	1	0	0	4	2	1	6	0	0	2
16	12	10	5	1	8	8	3	5	2	9	0	1	0	4	0	3	7	0	2	1
17	11	9	3	0	6	6	2	6	0	3	1	0	0	5	1	0	2	0	0	4
18	13	12	6	1	14	8	8	3	2	10	4	0	1	3	10	3	6	0	2	0
19	12	12	4	0	12	8	8	4	0	8	3	0	0	4	8	2	8	0	0	0
20	10	8	2	0	3	7	1	6	0	3	0	0	0	4	0	0	2	0	0	3
21	11	9	4	1	6	8	2	5	2	6	0	1	0	3	0	0	4	0	2	2
22	12	11	5	1	10	9	6	4	3	9	0	3	0	2	0	3	8	1	2	0
23	10	10	2	0	5	8	4	6	0	4	0	0	0	3	0	1	6	1	0	2
24	10	10	2	0	6	9	5	6	0	4	1	0	0	2	2	0	6	2	0	1
25	10	9	2	0	4	8	2	5	0	5	0	0	0	2	0	1	5	0	0	3
26	11	11	4	1	9	9	6	4	3	6	0	3	0	0	0	0	8	1	2	1
27	15	12	9	2	18	9	9	0	4	18	6	0	2	0	18	9	0	0	6	0
28	10	8	2	0	3	6	1	5	0	2	0	0	0	2	0	0	1	0	0	5
29	13	13	6	1	16	8	10	2	3	10	4	2	1	1	11	3	6	0	2	0
30	14	13	8	2	21	6	6	3	7	6	9	3	6	3	6	0	6	0	2	0
31	10	9	2	0	4	6	2	6	0	2	0	0	0	4	0	0	2	0	0	4
32	10	7	2	0	2	6	0	5	0	2	0	0	0	2	0	0	0	0	0	4
33	12	11	5	1	12	6	4	5	3	4	3	2	1	4	2	0	4	0	1	2
34	11	9	4	1	6	6	2	5	2	3	0	1	0	3	0	0	2	0	1	4
35	10	12	2	0	8	10	10	4	0	4	1	0	0	0	4	0	8	4	0	0
36	9	9	1	0	3	9	3	6	0	3	0	0	0	0	0	0	6	1	0	3
37	10	10	2	0	6	8	4	5	0	4	1	0	0	2	2	0	6	0	0	2
38	13	14	6	1	19	7	11	2	4	7	6	4	2	1	11	1	6	0	1	0
39	9	9	1	0	3	8	3	6	0	2	0	0	0	1	0	0	4	1	0	3
40	11	11	4	1	9	7	6	5	3	3	0	3	0	3	0	0	4	1	1	2
41	10	9	2	0	5	6	2	5	0	2	1	0	0	2	1	0	2	0	0	4
42	9	8	1	0	2	7	1	6	0	2	0	0	0	2	0	0	2	0	0	3
43	9	8	1	0	2	7	1	5	0	2	0	0	0	1	0	0	2	0	0	4
44	9	9	1	0	3	7	3	5	0	1	0	0	0	1	0	0	2	1	0	4
45	11	13	4	1	12	9	12	3	4	3	0	6	0	0	0	0	6	4	1	0
46	9	7	1	0	1	6	0	5	0	1	0	0	0	1	0	0	0	0	0	4
47	9	8	1	0	2	6	1	5	0	1	0	0	0	1	0	0	1	0	0	4
48	14	15	8	2	24	7	14	0	8	6	9	7	6	0	12	0	0	3	2	0
49	14	15	8	2	24	6	12	1	8	6	9	6	6	0	12	0	4	0	2	0
50	8	7	0	0	0	6	0	5	0	0	0	0	0	0	0	0	0	0	0	4

Tabelle 4.4: Substruktur–Vielfachheiten für die Bibliothek von Decanen aus Abbildung 4.6 bzgl. der Substrukturen aus Abbildung 4.7

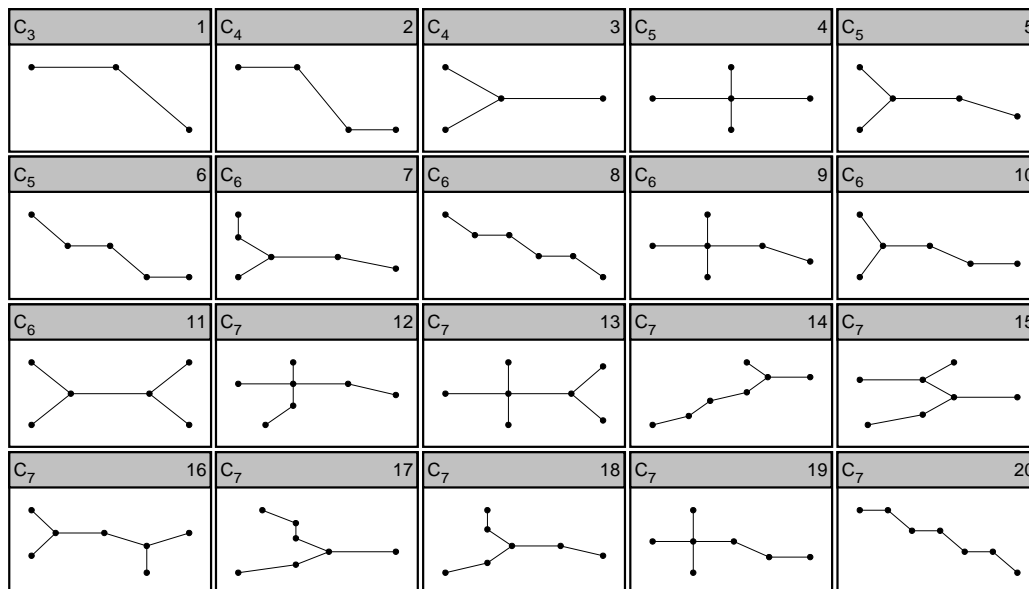


Abbildung 4.7: Substrukturen mit 2–6 Kanten für die Bibliothek von Decanen aus Abbildung 4.6

würden sich vordefinierte, eher auf allgemeine Chemie ausgelegte Substruktur-Vektoren nicht als Eingabe für statistische Lernverfahren eignen, da sich ihre Werte — wenn überhaupt — nur in sehr wenigen Komponenten unterscheiden. Dennoch haben auch statische Substruktur-Vektoren ihre Vorzüge, denn man kann genau diejenigen Substrukturen einbringen, für die eine chemische Relevanz bereits bekannt ist.

## 4.4 Struktur–Eigenschafts–Beziehungen

Gegeben sei eine reale Bibliothek mit  $m$  Verbindungen sowie einer gemessenen Eigenschaft  $Y$ . Für die folgenden Betrachtung wollen wir zunächst annehmen, dass  $Y$  reellwertig ist. Die Verbindungen der realen Bibliothek sind durch ihre molekularen Graphen dargestellt. Ausgangspunkt für unsere QSPR–Untersuchung sind also Tupel

$$(M_i, y_i) \in \mathcal{M} \times \mathbb{R}, \quad i \in m.$$

Gesucht wird nach einer Funktion

$$\Psi : \mathcal{M} \longrightarrow \mathbb{R},$$

die unsere QSPR mathematisch beschreibt. Wie man  $\Psi$  bestimmt, wurde bereits in Abschnitt 4.1 und Kapitel 3 erläutert. Typischerweise ist dann  $\Psi$  aus mehreren nacheinander auszuführenden Abbildungen zusammengesetzt:

- Zuerst werden die molekularen Graphen vermöge molekularer Deskriptoren  $\mathcal{D} = (D_i)_{i \in n}$  auf reelle Zahlen abgebildet:

$$\mathcal{D} : \mathcal{M} \longrightarrow \mathbb{R}^n, \quad M \longmapsto (D_i(M))_{i \in n}.$$

- Transformationen der Deskriptorenwerte

$$\tau = (\tau_i)_{i \in n} : \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

die zum Trainieren der Vorhersagefunktion notwendig oder hilfreich waren, müssen durchgeführt werden.

- Die eigentliche Vorhersagefunktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , die aus einem statistischen Lernverfahren gewonnen wurde, wird angewendet.
- Wurde zur Durchführung des Lernverfahrens auch die Zielvariable einer Transformation  $\sigma$  unterzogen, so muss schließlich die Rücktransformation des Eigenschaftswertes mittels  $\sigma^{-1}$  erfolgen.

Zusammenfassend können wir das QSPR–Modell schreiben als Komposition

$$\Psi = \sigma^{-1} \circ f \circ \tau \circ \mathcal{D}.$$

Ist  $Y$  diskret und nimmt Werte aus einer endlichen Menge  $\mathcal{C}$  an, dann haben auch  $f$  und  $\Psi$  den Wertebereich  $\mathcal{C}$  und  $\Psi = f \circ \tau \circ \mathcal{D}$ .

### 4.4.1 Beispiel: Siedepunkte von Alkanen

Abbildung 4.6 zeigt eine reale Bibliothek von 50 Decanen mit ihren Siedepunkten. Strukturen und BP wurden der *Beilstein*-Datenbank entnommen (vgl. Abschnitt 1.7), BP sind in °C angegeben. Wir wollen QSPR-Modelle für diese physikalische Eigenschaft finden.

#### Lineare Modellierung durch topologische Indizes

Dazu starten wir unsere Untersuchungen mit 30 topologischen Indizes<sup>3</sup>, die in Abschnitt 4.3.1 vorgestellt wurden und in *MOLGEN-QSPR* (vgl. Anhang B) enthalten sind:

$$W, M_1, M_2, {}^0\chi, {}^1\chi, {}^2\chi, {}^0\chi^v, {}^1\chi^v, {}^2\chi^v, {}^3\chi^v, J, MTI, twc, mwc^{(2)}, mwc^{(3)}, mwc^{(4)}, mwc^{(5)}, mwc^{(6)}, mwc^{(7)}, mwc^{(8)}, \lambda_1^A, IC_0, CIC_0, SIC_0, IC_1, CIC_1, SIC_1, IC_2, CIC_2, SIC_2.$$

Da Decane nur Einfachbindungen und keine Heteroatome enthalten, haben  ${}^k\chi$  und  ${}^k\chi^v$  identische Werte. Wir vernachlässigen deshalb  ${}^0\chi$ ,  ${}^1\chi$  und  ${}^2\chi$ . Per Definition haben Decane dieselbe Summenformel  $C_{10}H_{22}$ . Deshalb sind  $IC_0$ ,  $CIC_0$ ,  $SIC_0$  konstant und können auf die Modellierung ebenfalls keinen Einfluss nehmen. Tabellen 4.5 und 4.6 enthalten die Werte für die verbleibenden 24 Indizes angewendet auf die 50 Decane aus Abbildung 4.6. Wir sehen, dass  $M_1$  und  $mwc^{(2)}$  gleiche Werte annehmen. Diese Identität ist allgemein gültig.

#### 4.4.1 Satz:

Sei  $M \in \mathcal{M}$ . Dann ist

$$D_{mwc^{(2)}}(M) = D_{M_1}(M).$$

*Beweis:* Mit den Bezeichnungen aus Beispiel 4.3.4 gilt:

$$\begin{aligned} D_{mwc^{(2)}}(M) &= \sum_{i \in \Omega} \sum_{j \in \Omega} a_{ij}^{(2)} = \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{k \in \Omega} a_{ik} a_{kj} \\ &= \sum_{k \in \Omega} \sum_{i \in \Omega} \sum_{j \in \Omega} a_{ik} a_{kj} = \sum_{k \in \Omega} \sum_{i \in \Omega} a_{ik} \sum_{j \in \Omega} a_{kj} \\ &= \sum_{k \in \Omega} \sum_{i \in \Omega} a_{ik} \deg_{\mathcal{G}^s_{|\Omega}}(k) = \sum_{k \in \Omega} \deg_{\mathcal{G}^s_{|\Omega}}(k) \sum_{i \in \Omega} a_{ik} \\ &= \sum_{k \in \Omega} \deg_{\mathcal{G}^s_{|\Omega}}(k) \deg_{\mathcal{G}^s_{|\Omega}}(k) = D_{M_1}(M) \end{aligned}$$

□

	$W$	$M_1$	$M_2$	${}_0\chi^v$	${}_1\chi^v$	${}_2\chi^v$	${}_3\chi^v$	$J$	$MTI$	$twc$	$mwcc(2)$	$mwcc(3)$
1	127	46	44	8,4142	4,2071	5,6213	1,6250	3,5630	464	19248	46	88
2	134	42	41	8,1987	4,4545	4,6128	2,0841	3,3555	488	15138	42	82
3	135	40	39	8,1463	4,5197	4,3643	1,7475	3,3374	490	12930	40	78
4	126	42	42	8,1987	4,4925	4,4473	2,0557	3,6308	456	17334	42	84
5	124	44	44	8,3618	4,3272	4,9861	2,0724	3,6842	450	19018	44	88
6	131	42	41	8,1987	4,4545	4,6586	1,7423	3,4695	476	16146	42	82
7	139	42	40	8,1987	4,4165	4,8467	1,7083	3,2055	508	13874	42	80
8	123	44	45	8,3618	4,3372	4,8966	2,3034	3,7348	446	20498	44	90
9	119	46	46	8,4142	4,2678	5,2552	1,9660	3,8876	432	23048	46	92
10	127	42	42	8,1987	4,4772	4,5122	1,8876	3,6256	460	17946	42	84
11	142	38	37	7,9831	4,6639	3,8769	1,9243	3,1600	516	11114	38	74
12	131	42	42	8,1987	4,4772	4,4503	2,3556	3,4647	476	16602	42	84
13	120	44	46	8,3618	4,3599	4,7413	2,4973	3,8656	434	22234	44	92
14	146	40	38	8,0355	4,5607	4,3713	1,7803	3,0438	534	12390	40	76
15	130	40	41	8,1463	4,5746	3,9924	2,4585	3,5027	470	14984	40	82
16	126	42	43	8,1987	4,5152	4,2353	2,5551	3,6419	456	18280	42	86
17	136	40	40	8,1463	4,5366	4,1925	2,3374	3,3014	494	13242	40	80
18	118	44	47	8,3618	4,3921	4,5402	2,8635	3,9418	426	23206	44	94
19	121	42	45	8,3094	4,4641	4,2063	2,9325	3,8140	436	19426	42	90
20	143	38	37	7,9831	4,6639	3,8650	2,0183	3,1244	520	10786	38	74
21	134	40	40	8,0355	4,6213	4,0178	2,1339	3,4175	486	15664	40	80
22	122	42	44	8,1987	4,5378	4,1157	2,6082	3,8026	440	20028	42	88
23	133	38	39	7,9831	4,7399	3,4316	2,5873	3,4123	480	13028	38	78
24	131	38	39	7,9831	4,7187	3,5814	2,2617	3,4999	472	13848	38	78
25	138	38	38	7,9831	4,7019	3,6430	2,2831	3,2686	500	12020	38	76
26	126	40	42	8,0355	4,6820	3,6642	2,5607	3,6903	454	18298	40	84
27	111	48	51	8,5774	4,1547	5,4537	2,5981	4,2311	402	29658	48	102
28	146	38	37	7,9831	4,6639	3,8382	2,1753	3,0333	532	10236	38	74
29	116	44	48	8,3618	4,4147	4,3748	3,1439	4,0341	418	24610	44	96
30	115	46	50	8,4142	4,3107	4,8839	2,9053	4,1018	416	29160	46	100
31	141	38	38	7,9831	4,7019	3,6042	2,5461	3,1682	512	11298	38	76
32	151	38	36	7,9831	4,6259	4,0722	1,8129	2,9095	552	9316	38	72
33	127	42	44	8,1987	4,5040	4,2468	2,7376	3,6334	460	19738	42	88
34	138	40	40	8,0355	4,6213	3,9749	2,4142	3,2770	502	14774	40	80
35	125	38	41	7,9831	4,7948	3,1532	2,7642	3,6982	448	15866	38	82
36	138	36	36	7,8200	4,8461	3,2321	2,0908	3,2951	498	10950	36	72
37	135	38	39	7,9831	4,7187	3,5319	2,4594	3,3759	488	13386	38	78
38	115	44	49	8,3618	4,4248	4,2854	3,3705	4,0893	414	26106	44	98
39	141	36	36	7,8200	4,8461	3,2052	2,2402	3,2055	510	10570	36	72
40	129	40	42	8,0355	4,6820	3,6213	2,8410	3,5755	466	17588	40	84
41	143	38	38	7,9831	4,6807	3,7171	2,4011	3,1296	520	11616	38	76
42	149	36	35	7,8200	4,8081	3,3896	2,1010	2,9984	542	9330	36	70
43	150	36	35	7,8200	4,8081	3,3896	2,0820	2,9680	546	9194	36	70
44	145	36	36	7,8200	4,8461	3,1783	2,3706	3,0869	526	10052	36	72
45	121	40	44	8,0355	4,7426	3,3107	3,0303	3,8748	434	20526	40	88
46	158	36	34	7,8200	4,7701	3,5967	1,8850	2,7732	578	7896	36	68
47	153	36	35	7,8200	4,8081	3,3628	2,2474	2,8862	558	8788	36	70
48	110	46	52	8,4142	4,3713	4,5178	3,3713	4,3283	396	31916	46	104
49	111	46	52	8,4142	4,3713	4,4749	3,5999	4,2818	400	31632	46	104
50	165	34	32	7,6569	4,9142	3,1213	1,9571	2,6476	604	6500	34	64

Tabelle 4.5: Werte topologischer Indizes für die reale Bibliothek von Decanen aus Abbildung 4.6



	$muc(4)$	$muc(5)$	$muc(6)$	$muc(7)$	$muc(8)$	$\chi_A$	$IC_1$	$CIC_1$	$SIC_1$	$IC_2$	$CIC_2$	$SIC_2$
1	218	432	1040	2114	4978	2,1987	1,3245	3,6755	0,26489	1,7947	3,2053	0,35895
2	188	376	854	1728	3900	2,1474	1,4227	3,5773	0,28455	2,5354	2,4646	0,50707
3	174	342	764	1506	3366	2,1010	1,3602	3,6398	0,27205	2,2823	2,7177	0,45645
4	198	402	942	1926	4494	2,1889	1,4227	3,5773	0,28455	2,4104	2,5896	0,48207
5	210	430	1012	2098	4894	2,2047	1,3870	3,6130	0,27739	2,2322	2,7678	0,44645
6	194	382	908	1794	4272	2,1753	1,4227	3,5773	0,28455	2,5354	2,4646	0,50707
7	184	356	818	1590	3660	2,1289	1,4227	3,5773	0,28455	2,4729	2,5271	0,49457
8	212	450	1040	2250	5144	2,2361	1,3870	3,6130	0,27739	2,2322	2,7678	0,44645
9	234	472	1198	2422	6140	2,2646	1,3245	3,6755	0,26489	2,0416	2,9584	0,40832
10	200	404	968	1962	4710	2,2089	1,4227	3,5773	0,28455	2,5590	2,4410	0,51179
11	158	312	668	1328	2844	2,0698	1,3716	3,6284	0,27433	2,6945	2,3055	0,53891
12	192	396	896	1874	4214	2,1813	1,4227	3,5773	0,28455	2,4965	2,5035	0,49929
13	218	470	1102	2402	5608	2,2616	1,3870	3,6130	0,27739	2,2169	2,7831	0,44338
14	170	328	738	1436	3242	2,1192	1,3213	3,6787	0,26427	2,3204	2,6796	0,46407
15	180	376	822	1730	3770	2,1455	1,3602	3,6398	0,27205	2,4576	2,5424	0,49151
16	200	416	968	2020	4704	2,2082	1,4227	3,5773	0,28455	2,5590	2,4410	0,51179
17	172	354	754	1566	3326	2,1067	1,3602	3,6398	0,27205	2,3448	2,6552	0,46895
18	222	484	1138	2494	5854	2,2711	1,3870	3,6130	0,27739	2,3183	2,6817	0,46367
19	202	442	986	2170	4826	2,2143	1,1995	3,8005	0,23989	1,7947	3,2053	0,35895
20	156	310	650	1306	2724	2,0529	1,3716	3,6284	0,27433	2,5460	2,4540	0,50919
21	182	372	852	1756	4030	2,1823	1,3213	3,6787	0,26427	2,3675	2,6325	0,47351
22	206	438	1024	2186	5106	2,2361	1,4227	3,5773	0,28455	2,4965	2,5035	0,49929
23	166	346	736	1538	3270	2,1085	1,3716	3,6284	0,27433	2,5460	2,4540	0,50919
24	168	354	760	1614	3456	2,1358	1,3716	3,6284	0,27433	2,5931	2,4069	0,51863
25	162	328	702	1426	3056	2,0886	1,3716	3,6284	0,27433	2,6556	2,3444	0,53113
26	192	410	942	2018	4642	2,2216	1,3213	3,6787	0,26427	2,3439	2,6561	0,46879
27	258	558	1404	3042	7650	2,3344	1,2575	3,7425	0,25151	1,5704	3,4296	0,31407
28	154	304	632	1252	2602	2,0314	1,3716	3,6284	0,27433	2,5695	2,4305	0,51391
29	226	502	1180	2626	6174	2,2882	1,3870	3,6130	0,27739	2,3183	2,6817	0,46367
30	242	552	1310	3038	7156	2,3433	1,3245	3,6755	0,26489	2,0416	2,9584	0,40832
31	158	322	668	1368	2834	2,0615	1,3716	3,6284	0,27433	2,5306	2,4694	0,50613
32	150	288	596	1154	2374	2,0000	1,3716	3,6284	0,27433	2,4056	2,5944	0,48113
33	200	436	986	2174	4916	2,2410	1,4227	3,5773	0,28455	2,5590	2,4410	0,51179
34	178	364	816	1680	3784	2,1679	1,3213	3,6787	0,26427	2,4300	2,5700	0,48601
35	178	386	838	1818	3946	2,1701	1,3716	3,6284	0,27433	2,2806	2,7194	0,45613
36	150	306	642	1314	2760	2,0743	1,3009	3,6991	0,26017	2,3183	2,6817	0,46367
37	166	348	742	1568	3342	2,1268	1,3716	3,6284	0,27433	2,5306	2,4694	0,50613
38	228	522	1208	2778	6424	2,3073	1,3870	3,6130	0,27739	2,3183	2,6817	0,46367
39	148	302	624	1280	2648	2,0642	1,3009	3,6991	0,26017	2,4280	2,5720	0,48560
40	188	404	908	1962	4418	2,2120	1,3213	3,6787	0,26427	2,4064	2,5936	0,48129
41	158	324	674	1394	2904	2,0886	1,3716	3,6284	0,27433	2,6320	2,3680	0,52641
42	142	282	574	1150	2344	2,0285	1,3009	3,6991	0,26017	2,5141	2,4859	0,50282
43	142	280	572	1134	2324	2,0237	1,3009	3,6991	0,26017	2,5141	2,4859	0,50282
44	146	296	604	1230	2516	2,0491	1,3009	3,6991	0,26017	2,3655	2,6345	0,47310
45	200	444	1010	2246	5110	2,2504	1,3213	3,6787	0,26427	2,2189	2,7811	0,44379
46	136	260	520	1000	2000	1,9696	1,3009	3,6991	0,26017	2,4516	2,5484	0,49032
47	140	276	554	1098	2208	2,0066	1,3009	3,6991	0,26017	2,4516	2,5484	0,49032
48	252	586	1402	3286	7826	2,3649	1,3245	3,6755	0,26489	2,0294	2,9706	0,40588
49	250	584	1388	3266	7734	2,3623	1,3245	3,6755	0,26489	1,9669	3,0331	0,39338
50	122	232	444	848	1626	1,9190	1,1216	3,8784	0,22433	1,9056	3,0944	0,38113

Tabelle 4.6: Werte topologischer Indizes für die reale Bibliothek von Decanen aus Abbildung 4.6 (fortgesetzt)

Offenbar wenig bekannt ist die Beziehung zwischen dem zweiten Zagreb Index und dem Molecular Walk Count für Länge 3. So wird in einem jüngst erschienenen Übersichtsartikel [103] zu den Zagreb Indizes nicht über die Abhängigkeit zwischen  $M_2$  und  $mw_c^{(3)}$  berichtet. Wir formulieren

#### 4.4.2 Satz:

Sei  $M \in \mathcal{M}$ . Dann ist

$$D_{mw_c^{(3)}}(M) = D_{M_2}(M).$$

*Beweis:* Mit den Bezeichnungen aus Beispiel 4.3.4 gilt:

$$\begin{aligned} D_{mw_c^{(3)}}(M) &= \sum_{i \in \Omega} \sum_{j \in \Omega} a_{ij}^{(3)} = \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{k \in \Omega} a_{ik} a_{kj}^{(2)} \\ &= \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{k \in \Omega} a_{ik} \sum_{l \in \Omega} a_{kl} a_{lj} = \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{k \in \Omega} \sum_{l \in \Omega} a_{ik} a_{kl} a_{lj} \\ &= \sum_{k \in \Omega} \sum_{l \in \Omega} \sum_{i \in \Omega} \sum_{j \in \Omega} a_{kl} a_{ik} a_{lj} = \sum_{k \in \Omega} \sum_{l \in \Omega} a_{kl} \sum_{i \in \Omega} a_{ik} \sum_{j \in \Omega} a_{lj} \\ &= \sum_{k \in \Omega} \sum_{l \in \Omega} a_{kl} \deg_{\gamma|\Omega}^s(k) \deg_{\gamma|\Omega}^s(l) \\ &= 2 \sum_{\{k,l\} \in E_{\gamma|\Omega}} \deg_{\gamma|\Omega}^s(k) \deg_{\gamma|\Omega}^s(l) = 2D_{M_2}(M) \end{aligned}$$

□

$M_1$  und  $M_2$  können demnach ebenfalls aus der Liste der Indizes entfernt werden.

Ein Blick auf die Korrelationsmatrix verrät weitere paarweise affine Abhängigkeiten zwischen den betrachteten Deskriptoren. Tabelle 4.7 zeigt die ersten Spalten der Korrelationsmatrix der verbleibenden Indizes. Die Vorzeichen der Korrelationskoeffizienten wurden dabei vernachlässigt. Die erste Spalte enthält Absolutbeträge der Korrelationskoeffizienten zwischen BP und den Deskriptoren. Die Deskriptoren wurden nach abfallenden absoluten Korrelationskoeffizienten bzgl. BP angeordnet. Die weiteren Spalten enthalten Absolutbeträge von Korrelationskoeffizienten zwischen je zwei Deskriptoren. Für lineare Regression ist es nicht sinnvoll, vollständig korrelierte Vorhersagevariablen zu verwenden. In unserem Beispiel sind die Paare aus  $\{IC_1, CIC_1, SIC_1\}$  vollständig korreliert. Dies resultiert daraus, dass die Verbindungen in der Decan-Bibliothek gleiche Anzahl von Atomen haben.

<sup>3</sup>Der kürzeren Schreibweise wegen werden wir in diesem und den folgenden Abschnitten, falls Verwechslungen ausgeschlossen sind, nur die Symbole der Indizes angeben.

	<i>BP</i>	${}^2\chi^v$	${}^1\chi^v$	<i>IC</i> <sub>1</sub>	<i>CIC</i> <sub>1</sub>	<i>SIC</i> <sub>1</sub>	${}^0\chi^v$	${}^3\chi^v$	<i>max</i> <sup>(2)</sup>	<i>max</i> <sup>(4)</sup>	<i>W</i>	<i>MTI</i>
<i>BP</i>	1,000	0,679	0,587	0,513	0,513	0,513	0,485	0,478	0,447	0,290	0,254	0,237
${}^2\chi^v$	0,679	1,000	0,975	0,297	0,297	0,297	0,892	0,054	0,896	0,768	0,586	0,558
${}^1\chi^v$	0,587	0,975	1,000	0,302	0,302	0,302	0,970	0,163	0,964	0,876	0,732	0,708
<i>IC</i> <sub>1</sub>	0,513	0,297	0,302	1,000	1,000	1,000	0,310	0,042	0,272	0,222	0,283	0,281
<i>CIC</i> <sub>1</sub>	0,513	0,297	0,302	1,000	1,000	1,000	0,310	0,042	0,272	0,222	0,283	0,281
<i>SIC</i> <sub>1</sub>	0,513	0,297	0,302	1,000	1,000	1,000	0,310	0,042	0,272	0,222	0,283	0,281
${}^0\chi^v$	0,485	0,892	0,970	0,310	0,310	0,310	1,000	0,371	0,986	0,951	0,867	0,850
${}^3\chi^v$	0,478	0,054	0,163	0,042	0,042	0,042	0,371	1,000	0,368	0,539	0,641	0,654
<i>max</i> <sup>(2)</sup>	0,447	0,896	0,964	0,272	0,272	0,272	0,986	0,368	1,000	0,970	0,862	0,844
<i>max</i> <sup>(4)</sup>	0,290	0,768	0,876	0,222	0,222	0,222	0,951	0,539	0,970	1,000	0,943	0,931
<i>W</i>	0,254	0,586	0,732	0,283	0,283	0,283	0,867	0,641	0,862	0,943	1,000	0,999
<i>MTI</i>	0,237	0,558	0,708	0,281	0,281	0,281	0,850	0,654	0,844	0,931	0,999	1,000
<i>max</i> <sup>(6)</sup>	0,202	0,710	0,831	0,180	0,180	0,180	0,921	0,602	0,945	0,995	0,948	0,939
$\lambda_1^A$	0,196	0,628	0,762	0,245	0,245	0,245	0,875	0,644	0,898	0,969	0,969	0,964
<i>max</i> <sup>(3)</sup>	0,175	0,680	0,818	0,195	0,195	0,195	0,922	0,665	0,932	0,986	0,954	0,945
<i>max</i> <sup>(5)</sup>	0,142	0,655	0,794	0,172	0,172	0,172	0,902	0,675	0,919	0,983	0,955	0,947
<i>J</i>	0,141	0,553	0,707	0,195	0,195	0,195	0,848	0,696	0,853	0,949	0,990	0,989
<i>max</i> <sup>(8)</sup>	0,140	0,675	0,803	0,146	0,146	0,146	0,899	0,635	0,926	0,987	0,940	0,932
<i>max</i> <sup>(7)</sup>	0,105	0,637	0,777	0,144	0,144	0,144	0,885	0,684	0,908	0,977	0,945	0,938
<i>twc</i>	0,097	0,642	0,779	0,131	0,131	0,131	0,883	0,674	0,909	0,978	0,937	0,930
<i>IC</i> <sub>2</sub>	0,002	0,459	0,500	0,594	0,594	0,594	0,511	0,260	0,540	0,551	0,435	0,423
<i>CIC</i> <sub>2</sub>	0,002	0,459	0,500	0,594	0,594	0,594	0,511	0,260	0,540	0,551	0,435	0,423
<i>SIC</i> <sub>2</sub>	0,002	0,459	0,500	0,594	0,594	0,594	0,511	0,260	0,540	0,551	0,435	0,423

Tabelle 4.7: Ausschnitt der Korrelationsmatrix für Siedepunkte und topologische Indizes der realen Bibliothek von Decanen

Für molekulare Graphen  $M$  von Decanen ist  $D_{CIC_1}(M) = 5 - D_{IC_1}(M)$  und  $D_{SIC_1}(M) = \frac{1}{5}D_{IC_1}(M)$ . Gleiches gilt für  $\{IC_2, CIC_2, SIC_2\}$ . Deshalb können wir  $CIC_1, SIC_1, CIC_2$  und  $SIC_2$  von unseren weiteren Betrachtungen ausschließen.

Verwendet man alle 18 verbleibenden Indizes, erhält man ein lineares Modell mit  $R^2 = 0,97439$  und  $R_{CV}^2 = 0,94191$ . Um Overfitting zu vermeiden<sup>4</sup>, suchen wir nach Modellen, die weniger Deskriptoren verwenden. Wir durchlaufen dazu für  $n = 1, \dots, 5$  alle  $n$ -Teilmengen der 18 topologischen Indizes und geben jeweils die Modelle mit größten  $R^2$ -Werten an. Dabei werden zunächst die als Vorhersagevariablen verwendeten Deskriptoren  $X_j, j \in n$  genannt, gefolgt von der mittels OLS-Regression ermittelten Vorhersagefunktion  $f$ . Des Weiteren sind die Vorhersagefunktionen auch für die autoskalierten Vorhersagevariablen  $X_j^*$  angegeben. Anhand dieser Darstellung kann man den Einfluss der verschiedenen Deskriptoren auf das Modell besser erkennen.

$n = 1$  Deskriptor:  ${}^2\chi^v$ ,

$$\begin{aligned} f &= -8,0356X_0 + 190,74 \\ &= -5,0362X_0^* + 157,85. \end{aligned}$$

<sup>4</sup>Eine Faustregel besagt, dass pro 5 Beobachtungen etwa ein zusätzlicher Freiheitsgrad gerechtfertigt werden kann. Arbeitet man ohne Testsatz, so ist es in der Tat ein schwieriges Problem, die Anzahl der Freiheitsgrade sinnvoll zu beschränken (siehe z.B. [143]).

$n$	$R^2$	$R_{CV}^2$	$R^2 - R_{CV}^2$	$S$	$S_{CV}$	$S_{CV} - S$	$F$
1	0,46101	0,40131	0,059698	5,5019	5,7986	0,29669	41,06
2	0,89336	0,87999	0,013366	2,4732	2,6236	0,15042	196,87
3	0,93721	0,92689	0,010325	1,9183	2,0700	0,15172	228,87
4	0,95011	0,94126	0,008856	1,7287	1,8759	0,14718	214,27
5	0,95814	0,94709	0,011048	1,6015	1,8005	0,19896	201,42
6	0,96339	0,95022	0,013176	1,5149	1,7666	0,25173	188,62
7	0,96450	0,95043	0,014074	1,5095	1,7838	0,27431	163,02
8	0,96520	0,94761	0,017590	1,5127	1,8561	0,34331	142,13
9	0,96686	0,94794	0,018922	1,4944	1,8731	0,37868	129,67
10	0,97045	0,95468	0,015764	1,4292	1,7699	0,34062	128,07
11	0,97151	0,95542	0,016090	1,4216	1,7783	0,35671	117,81
12	0,97275	0,95591	0,016840	1,4092	1,7924	0,38323	110,05
13	0,97304	0,95294	0,020097	1,4209	1,8772	0,45629	99,94
14	0,97424	0,95061	0,023625	1,4087	1,9504	0,54171	94,53
15	0,97426	0,94917	0,025088	1,4287	2,0075	0,57889	85,79
16	0,97438	0,94563	0,028750	1,4468	2,1075	0,66078	78,43
17	0,97439	0,94191	0,032484	1,4688	2,2122	0,74343	71,63
18	0,97439	0,94191	0,032484	1,4923	2,2476	0,75532	65,53

Tabelle 4.8: Charakteristika der besten LM mit  $n$  Deskriptoren (aus 18 TI) für Siedepunkte von Decanen

$$\begin{aligned}
 n = 2 \text{ Deskriptoren: } & mwc^{(4)}, mwc^{(8)}, \\
 & f = -1,2961X_0 + 0,026540X_1 + 287,83 \\
 & = -42,917X_0^* + 41,312X_1^* + 157,85.
 \end{aligned}$$

$$\begin{aligned}
 n = 3 \text{ Deskriptoren: } & {}^3\chi^v, twc, mwc^{(5)}, \\
 & f = 16,793X_0 + 0,0085894X_1 - 0,69764X_2 + 246,86 \\
 & = 7,7409X_0^* + 53,768X_1^* - 59,883X_2^* + 157,85.
 \end{aligned}$$

$$\begin{aligned}
 n = 4 \text{ Deskriptoren: } & {}^3\chi^v, mwc^{(6)}, mwc^{(7)}, mwc^{(8)}, \\
 & f = 10,930X_0 - 0,32884X_1 - 0,042581X_2 + 0,064274X_3 + 229,69 \\
 & = 5,0382X_0^* - 79,236X_1^* - 25,319X_2^* + 100,05X_3^* + 157,85.
 \end{aligned}$$

$$\begin{aligned}
 n = 5 \text{ Deskriptoren: } & W, {}^3\chi^v, twc, mwc^{(4)}, mwc^{(8)}, \\
 & f = 0,44512X_0 + 9,7937X_1 - 0,0038957X_2 - 0,95038X_3 + 0,03649X_4 + 164,25 \\
 & = 5,6464X_0^* + 4,5145X_1^* - 24,386X_2^* - 31,468X_3^* + 56,794X_4^* + 157,85.
 \end{aligned}$$

Tabelle 4.8 gibt die statistischen Kennwerte  $R^2$ ,  $R_{CV}^2$ ,  $S$ ,  $S_{CV}$  und  $F$  sowie die Differenzen zwischen durch Resubstitution und LOO-CV ermittelten Charakteristika der besten LM mit  $n = 1, \dots, 18$  topologischen Indizes an. Für

$R^2$ ,  $R_{CV}^2$  und  $F$  sind jeweils die maximalen, in den übrigen Spalten die minimalen Werte unterstrichen. In Abbildungen 4.8 bis 4.10 sind die Werte durch Dreiecke ( $\triangle$ ) graphisch dargestellt, grau unterlegte Dreiecke markieren durch CV ermittelte Werte. Man beachte, dass in Abbildungen 4.8 und 4.9 die Y-Achsen logarithmisch skaliert sind. In beiden Abbildungen ist deutlich zu erkennen, dass die CV-Werte mit zunehmender Komplexität der Modelle wieder schlechter werden, was als Hinweis auf Overfitting verstanden werden kann.

$R^2$  muss notwendigerweise mit wachsendem  $n$  ansteigen. Für die Auswahl eines Modells durch  $R^2$  könnte man allenfalls die Zuwachsraten dieses Wertes heranziehen.  $R_{CV}^2$  nimmt sein Maximum für  $n = 12$  an. 12 Deskriptoren sind bei 50 Beobachtungen jedoch zu viele Vorhersagevariablen. Es würden Effekte des Overfittings bemerkbar. Aus dem gleichen Grund ist auch eine Auswahl des Modells bei kleinstem  $S$  mit  $n = 14$  Deskriptoren nicht empfehlenswert. Eine sinnvolle Wahl wäre das Modell mit  $n = 6$  Deskriptoren, gestützt durch das Argument, dass hier  $S_{CV}$  minimal wird. In [103] wird die Differenz  $S_{CV} - S$  als ein Maß für die Stabilität einer QSPR herangezogen. Diese Schlussweise würde für das Modell mit 4 Deskriptoren sprechen. Unterstützt würde diese Wahl durch die kleinste Abweichung von  $R^2$  und  $R_{CV}^2$ . Aber auch das Modell mit nur 3 Deskriptoren erreicht bereits recht gute Kennwerte. Insbesondere ist hier  $F$  maximal. In Abbildung 4.11 sind für dieses Modell vorhergesagte gegen experimentelle Siedepunkte abgetragen. Die schwarzen Markierungen stehen für Werte, die durch LOO-Kreuzvalidierung ermittelt wurden. Die gute Übereinstimmung von Vorhersagen durch Re-substitution und Kreuzvalidierung spricht für die Konsistenz des Modells. Überdies sieht man bereits bei diesem, mit nur 3 Deskriptoren sehr einfachen Modell eine gute Übereinstimmung vorhergesagter und experimenteller Werte.

Die Wahl des Auswahlkriteriums bleibt in letzter Instanz dem Anwender überlassen. In Kapitel 6 von [98] werden weitere Kriterien diskutiert. Zudem könnte man auch noch für jedes  $n$  die nächstbesten Modelle in die engere Auswahl einbeziehen, oder — mit höherem Rechenaufwand — gezielt nach Modellen mit maximalen  $R_{CV}^2$  bzw. minimalem  $S_{CV}$  suchen.

### Lineare Modellierung durch Substruktur-Vielfachheiten

An dieser Stelle wollen wir einen weiteren Aspekt der QSPR-Suche genauer beleuchten, nämlich die Art der verwendeten Deskriptoren. Dazu wurden die in den Verbindungen der realen Bibliothek enthaltenen Substrukturen mit 2 bis 6 Bindungen sowie deren Vielfachheiten ermittelt (vgl. Beispiel 4.3.8). Unter den resultierenden 20 SC gibt es keine vollständigen Korrelationen

$n$	$R^2$	$R_{CV}^2$	$R^2 - R_{CV}^2$	$S$	$S_{CV}$	$S_{CV} - S$	$F$
1	0,55334	0,51266	0,040681	5,0085	5,2316	0,22312	59,47
2	0,78507	0,74739	0,037677	3,5111	3,8064	0,29533	85,84
3	0,88372	0,86106	0,022654	2,6105	2,8535	0,24298	116,53
4	0,95621	0,94538	0,010825	1,6197	1,8089	0,18915	245,64
5	0,96185	0,94958	0,012277	1,5288	1,7577	0,22887	221,88
6	0,96669	0,95539	0,011298	1,4450	1,6723	0,22722	208,00
7	0,96869	0,95637	0,012318	1,4176	1,6733	0,25579	185,65
8	0,97133	0,96009	0,011242	1,3729	1,6199	0,24699	173,66
9	0,97167	0,95772	0,013947	1,3819	1,6880	0,30619	152,42
10	0,97202	0,95444	0,017577	1,3907	1,7746	0,38383	135,48
11	0,97566	0,95558	0,020084	1,3140	1,7752	0,46121	138,48
12	0,97587	0,95267	0,023197	1,3259	1,8569	0,53100	124,69
13	0,97627	0,95329	0,022978	1,3330	1,8701	0,53718	113,94
14	0,97676	0,94543	0,031325	1,3380	2,0501	0,71210	105,05
15	0,97711	0,94390	0,033215	1,3471	2,1090	0,76199	96,78
16	0,97790	0,93361	0,044297	1,3435	2,3289	0,98536	91,28
17	0,98026	0,94166	0,038607	1,2894	2,2170	0,92754	93,49
18	0,98028	0,93778	0,042497	1,3095	2,3260	1,01650	85,61
19	0,98030	0,93399	0,046304	1,3306	2,4354	1,10480	78,56
20	0,98030	0,93399	0,046304	1,3534	2,4771	1,12370	72,14

Tabelle 4.9: Charakteristika der besten LM mit  $n$  Deskriptoren (aus 20 SC) für Siedepunkte von Decanen

bezüglich der realen Bibliothek. Wie zuvor für topologische Indizes berechnen wir lineare Modelle mit größten  $R^2$ . Nachfolgend sind diese für  $n = 1, \dots, 5$  aufgelistet. Die Nummerierung der SC folgt dabei Abbildung 4.7. Die  $X_j^*$  bezeichnen wieder die autoskalierten Deskriptorenwerte.

$$\begin{aligned}
 n = 1 \text{ Deskriptor: } SC_{14}, \\
 f &= -1,7205X_0 + 163,08 \\
 &= -5,5175X_0^* + 157,85.
 \end{aligned}$$

$$\begin{aligned}
 n = 2 \text{ Deskriptoren: } SC_1, SC_5, \\
 f &= -6,3403X_0 + 1,6051X_1 + 216,85 \\
 &= -10,703X_0^* + 9,3142X_1^* + 157,85.
 \end{aligned}$$

$$\begin{aligned}
 n = 3 \text{ Deskriptoren: } SC_1, SC_9, SC_{15}, \\
 f &= -6,3067X_0 + 3,1880X_1 + 0,97551X_2 + 221,80 \\
 &= -10,646X_0^* + 6,5053X_1^* + 4,1075X_2^* + 157,85.
 \end{aligned}$$

$n = 4$  Deskriptoren:  $SC_1, SC_5, SC_6, SC_{19}$ ,

$$\begin{aligned} f &= -9,1531X_0 + 1,8270X_1 - 2,0494X_2 + 3,8403X_3 + 258,03 \\ &= -15,451X_0^* + 10,602X_1^* - 2,6580X_2^* + 5,2283X_3^* + 157,85. \end{aligned}$$

$n = 5$  Deskriptoren:  $SC_1, SC_5, SC_6, SC_{16}, SC_{19}$ ,

$$\begin{aligned} f &= -9,2522X_0 + 1,8724X_1 - 2,3131X_2 + 0,56818X_3 + 3,2492X_4 + 260,76 \\ &= -15,618X_0^* + 10,866X_1^* - 2,9999X_2^* + 1,1390X_3^* + 4,4235X_4^* + 157,85. \end{aligned}$$

Tabelle 4.9 enthält die verschiedenen Charakteristika zur Beurteilung der Modelle. Bei gleicher Anzahl von Deskriptoren ergeben sich mit Ausnahme der Modelle für  $n = 2$  und  $n = 3$  durch Verwendung von Substruktur-Vielfachheiten bessere  $R^2$  als bei topologischen Indizes. In Abbildungen 4.8 bis 4.10 sind die Kennwerte für Modelle mit SC als kopfstehende Dreiecke ( $\nabla$ ) eingetragen. CV-Werte sind wiederum grau unterlegt.

Bei der Auswahl eines auf Substruktur-Vielfachheiten basierenden Modells sprechen einige der oben erwähnten Argumente für das Modell mit 4 Deskriptoren. Es liegen minimale Differenzen  $R^2 - R_{CV}^2$  und  $SCV - S$  vor,  $F$  wird maximal. Der Scatterplot in Abbildung 4.12 zeigt durch dieses Modell vorhergesagte Siedepunkte sowie durch LOO-CV berechnete Werte.

### Lineare Modellierung durch TI und SC

Schließlich wollen wir die 18 topologischen Indizes *und* die 20 Substruktur-Vielfachheiten zur Berechnung bester linearer Modelle heranziehen. Zunächst berechnen wir wiederum die Korrelationsmatrix. Wir finden dabei ein Paar vollständig korrelierter Deskriptoren: Für alle  $M$  in unserer realen Bibliothek (und allgemein für Decane) gilt:  $D_{SC_1}(M) = \frac{1}{2}D_{mwc^{(2)}}(M) - 9$ . Wir können deshalb  $SC_1$  vernachlässigen. Unter Verwendung aller 37 verbleibenden Deskriptoren erhält man ein lineares Modell mit  $R^2 = 0,98756$ . Der wesentlich niedrigere Wert von  $R_{CV}^2 = 0,88667$  für Kreuzvalidierung ist hier ein deutlicher Hinweis auf Overfitting. Die besten linearen Modelle mit  $n = 1, \dots, 5$  Deskriptoren sind:

$n = 1$  Deskriptoren:  $SC_{14}$ ,

$$\begin{aligned} f &= -1,7205X_0 + 163,08 \\ &= -5,5175X_0^* + 157,85. \end{aligned}$$

$n = 2$  Deskriptoren:  $mwc^{(4)}, mwc^{(8)}$ ,

$$\begin{aligned} f &= -1,2961X_0 + 0,026540X_1 + 287,83 \\ &= -42,917X_0^* + 41,312X_1^* + 157,85. \end{aligned}$$

$n = 3$  Deskriptoren:  $W, SC_2, SC_{19}$ ,

$$\begin{aligned} f &= 1,1519X_0 + 6,3865X_1 + 2,1586X_2 - 58,907 \\ &= 14,612X_0^* + 13,349X_1^* + 2,9387X_2^* + 157,85. \end{aligned}$$

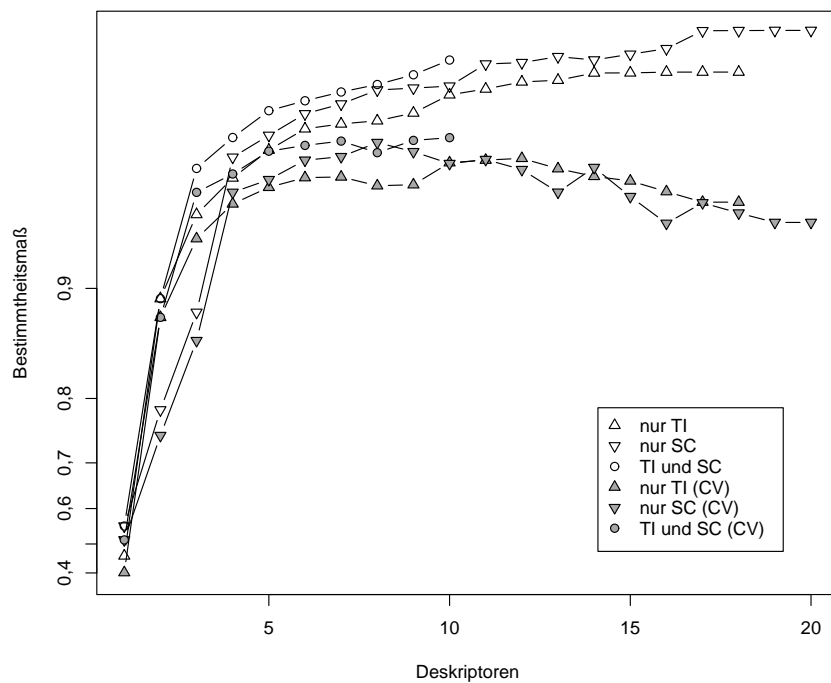


Abbildung 4.8: Bestimmtheitsmaß für die BP-Modelle

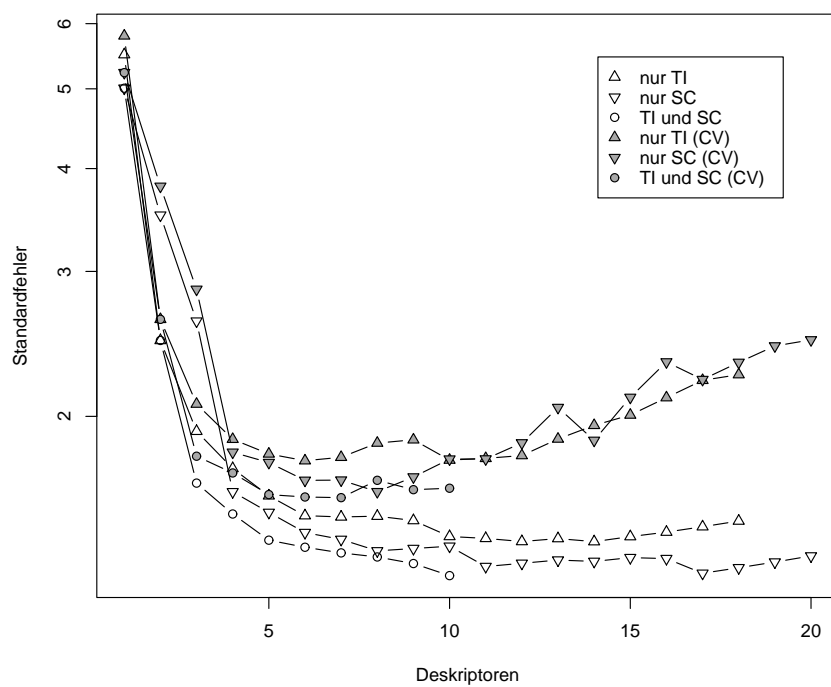
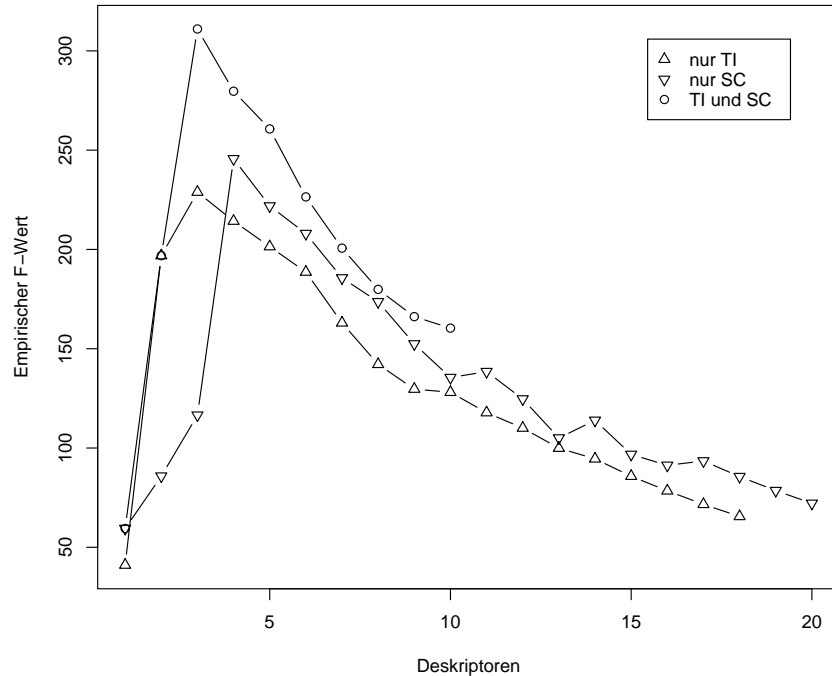


Abbildung 4.9: Standardfehler für die BP-Modelle



Abbildung 4.10: Empirischer  $F$ -Wert für die BP-Modelle

$n = 4$  Deskriptoren:  $W$ ,  ${}^2\chi^v$ ,  $SC_5$ ,  $SC_{19}$ ,

$$\begin{aligned} f &= 0,67431X_0 - 11,032X_1 + 1,7445X_2 + 2,1229X_3 + 98,079 \\ &= 8,5535X_0^* - 6,9138X_1^* + 10,123X_2^* + 2,8901X_3^* + 157,85. \end{aligned}$$

$n = 5$  Deskriptoren:  $W$ ,  ${}^3\chi^v$ ,  $MTI$ ,  $SC_5$ ,  $SC_{19}$ ,

$$\begin{aligned} f &= 11,320X_0 + 6,6378X_1 - 2,6968X_2 + 1,6403X_3 + 2,7967X_4 + 77,497 \\ &= 143,60X_0^* + 3,0598X_1^* - 129,06X_2^* + 9,5189X_3^* + 3,8075X_4^* + 157,85. \end{aligned}$$

Bereits ab  $n = 3$  Deskriptoren verwenden die Modelle mit besten  $R^2$  sowohl topologische Indizes als auch Substruktur-Vielfachheiten. Bei gleicher Anzahl  $n \geq 3$  von Deskriptoren haben Modelle, die sowohl TI als auch SC verwenden, höhere  $R^2$  als solche, die jeweils auf eine der beiden Arten von Deskriptoren beschränkt sind. Ein Blick auf Abbildung 4.8 bestätigt dies. Dort sind die  $R^2$ -Werte für Modelle mit TI und SC markiert durch Kreise ( $\circ$ ) eingetragen. Datenpunkte für  $R_{CV}^2$  sind wiederum grau unterlegt. Offensichtlich spiegelt sich der Vorteil einer Verwendung beider Arten von Deskriptoren auch in den kruzvalidierten Werten wieder, was auf eine bessere Konsistenz der Modelle schließen lässt.

Zur Modell-Selektion werfen wir einen Blick auf Tabelle 4.10. Für  $n = 3$  sind  $S_{CV} - S$  minimal und  $F$  maximal, für  $n = 5$  gibt es eine minimale Differenz  $R^2 - R_{CV}^2$  und für  $n = 7$  ist  $S$  minimal. Scatterplots der Modelle mit 3 und 7 Deskriptoren werden in Abbildungen 4.13 und 4.14 gezeigt.

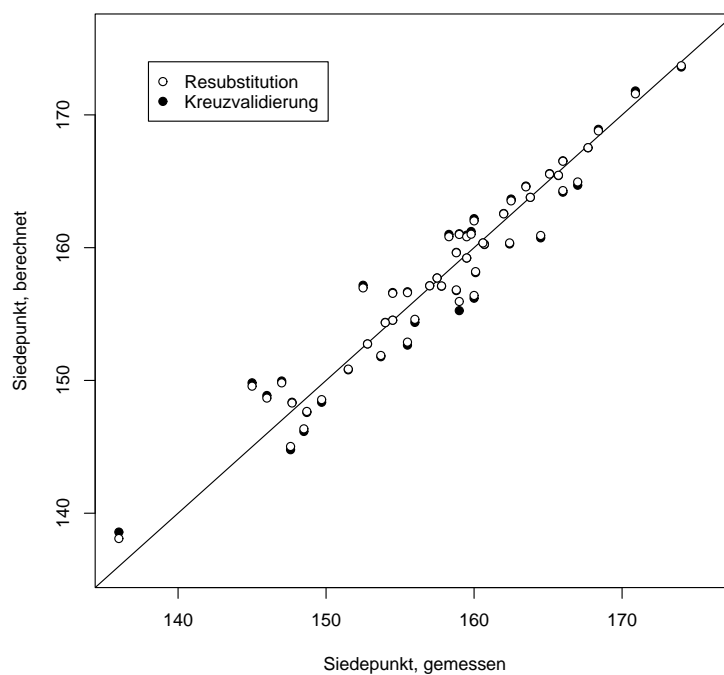


Abbildung 4.11: Scatterplot gemessener und vorhergesagter BP für das beste LM mit 3 Deskriptoren (nur TI)

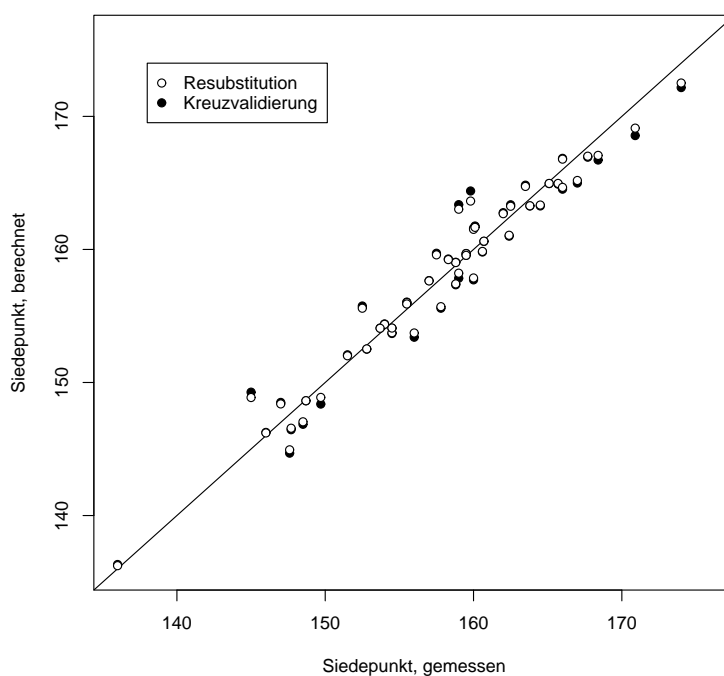


Abbildung 4.12: Scatterplot gemessener und vorhergesagter BP für das beste LM mit 4 Deskriptoren (nur SC)

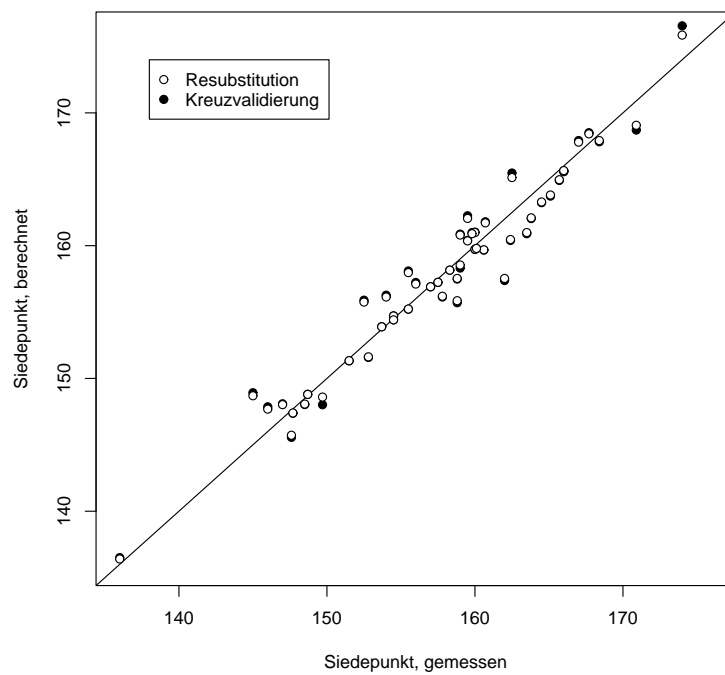


Abbildung 4.13: Scatterplot gemessener und vorhergesagter BP für das beste LM mit 3 Deskriptoren (TI und SC)

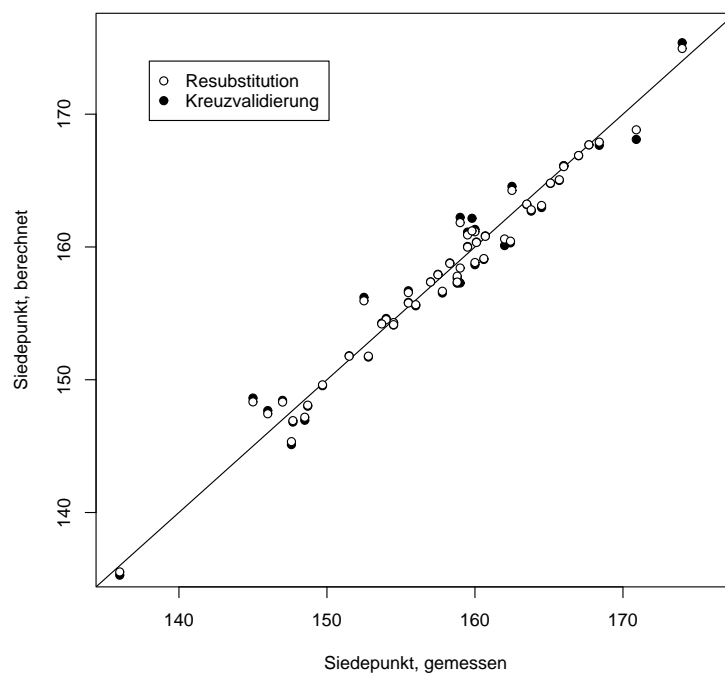


Abbildung 4.14: Scatterplot gemessener und vorhergesagter BP für das beste LM mit 7 Deskriptoren (TI und SC)

$n$	$R^2$	$R_{CV}^2$	$R^2 - R_{CV}^2$	$S$	$S_{CV}$	$S_{CV} - S$	$F$
1	0,55334	0,51266	0,040681	5,0085	5,2316	0,22312	59,47
2	0,89336	0,87999	0,013366	2,4732	2,6236	0,15042	196,87
3	0,95302	0,94538	0,007637	1,6594	1,7891	0,12978	311,02
4	0,96133	0,95135	0,009978	1,5220	1,7072	0,18511	279,67
5	0,96734	0,95785	0,009491	1,4145	1,6069	0,19245	260,68
6	0,96932	0,95936	0,009961	1,3868	1,5961	0,20934	226,45
7	0,97097	0,96045	0,010512	1,3651	1,5932	0,22808	200,66
8	0,97230	0,95746	0,014840	1,3496	1,6724	0,32288	179,89
9	0,97395	0,96062	0,013332	1,3250	1,6292	0,30415	166,16
10	0,97626	0,96129	0,014965	1,2810	1,6357	0,35465	160,37

Tabelle 4.10: Charakteristika der besten LM mit  $n$  Deskriptoren (aus 18 TI und 19 SC) für Siedepunkte von Decanen

### Zusammenfassung und Interpretation

Wir haben gesehen, dass Substruktur-Vielfachheiten in dem vorliegenden Beispiel gut zur QSPR-Findung geeignet sind, und somit eine Alternative zu topologischen Indizes darstellen. Noch bessere Modelle erhält man aber unter Verwendung beider Arten von Indizes. Sowohl unter den TI als auch bei den SC gibt es Deskriptoren, die sich besonders zur Modellierung des BP eignen, und andere, die kaum Einfluss auf die Modellbildung nehmen. In Tabelle 4.11 sind für die jeweils besten LM mit  $n = 1, \dots, 10$  Deskriptoren die verwendeten Deskriptoren durch Kreuze ( $\times$ ) markiert. In den linken Spalten werden Modelle repräsentiert, die auf TI bzw. SC beschränkt sind, rechts diejenigen, welche beide Arten von Deskriptoren verwenden. Die Striche ( $-$ ) in der Zeile von  $SC_1$  signalisieren, dass dieser Deskriptor wegen der oben erwähnten vollständigen Korrelation zu  $mwc^{(2)}$  bei der Suche nach TI-SC-Modellen nicht berücksichtigt wurde.

Insbesondere fällt bei der Analyse von Tabelle 4.11 auf, dass  $SC_{19}$  in jedem SC-Modell mit  $n \geq 4$  und jedem berechneten TI-SC-Modell mit  $n \geq 3$  Deskriptoren enthalten ist. Dabei ist die Gewichtung von  $SC_{19}$  eher gering, wie man den Vorhersagefunktionen mit autoskalierten Vorhersagevariablen entnehmen kann. Offenbar handelt es sich bei  $SC_{19}$  um einen wichtigen „Korrektur-Term“. Einen weiteren wichtigen SC zur Modellierung des BP liefert  $SC_5$ . Unter den TI sind Molecular Walk Counts,  ${}^3\chi^v$  und  $W$  als besonders geeignet zu nennen. Kaum Einfluss auf die besten LM nehmen hingegen  $J$ ,  $\lambda_1^A$  sowie die beiden verbliebenen informationstheoretischen Indizes  $IC_1$  und  $IC_2$ . Dies ist insbesondere bemerkenswert, da  $IC_1$  unter den betrachteten Indizes den drittbesten Korrelationskoeffizienten zu BP hat.

	<i>n</i> (nur TI bzw. SC)										<i>n</i> (TI und SC)									
	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
<i>W</i>	.	.	.	.	×	.	×	.	.	.	.	.	×	×	×	×	.	.	.	.
$\chi^v_0$	.	.	.	.	.	×	×	.	×	×	.	.	.	.	.	.	.	×	×	×
$\chi^v_1$	.	.	.	.	.	.	.	.	×	×	.	.	.	.	.	.	.	×	×	×
$\chi^v_2$	×	.	.	.	.	.	.	.	×	.	×	.	.	.	×	.	.	.	×	×
$\chi^v_3$	.	.	×	×	×	.	.	.	.	×	.	.	.	.	×	×	.	.	.	.
<i>J</i>	.	.	.	.	.	.	.	.	.	.	×	.	.	.	.	.	.	.	.	.
<i>MTI</i>	.	.	.	.	.	×	.	×	.	.	.	.	.	.	×	×	×	.	.	.
<i>twc</i>	.	.	×	.	×	×	×	.	×	×	.	.	.	.	.	.	.	.	.	.
<i>mwc</i> <sup>(2)</sup>	.	.	.	.	.	.	.	.	.	×	×	.	.	.	.	.	.	.	.	.
<i>mwc</i> <sup>(3)</sup>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	×	.	.	.
<i>mwc</i> <sup>(4)</sup>	.	×	.	.	×	.	.	×	×	×	×	.	×	.	.	.	×	.	.	×
<i>mwc</i> <sup>(5)</sup>	.	.	×	.	.	×	×	.	.	×	.	.	.	.	.	.	.	.	.	.
<i>mwc</i> <sup>(6)</sup>	.	.	.	×	.	×	×	×	×	×	.	.	.	.	.	.	×	.	.	.
<i>mwc</i> <sup>(7)</sup>	.	.	.	×	.	.	.	×	×	.	.	.	.	.	.	.	.	.	.	.
<i>mwc</i> <sup>(8)</sup>	.	×	.	×	×	×	×	×	×	×	.	×	.	.	.	.	.	.	.	.
$\lambda_1^A$	.	.	.	.	.	.	×	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>IC</i> <sub>1</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>IC</i> <sub>2</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>1</sub>	.	×	×	×	×	×	×	×	×	×	.	—	—	—	—	—	—	—	—	—
<i>SC</i> <sub>2</sub>	.	.	.	.	.	.	.	×	×	×	.	.	.	×	.	.	.	.	.	.
<i>SC</i> <sub>3</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>4</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	×
<i>SC</i> <sub>5</sub>	.	×	.	×	×	×	×	×	×	×	.	.	.	.	×	×	×	.	.	×
<i>SC</i> <sub>6</sub>	.	.	.	×	×	×	×	×	×	×	.	.	.	.	.	.	.	×	×	.
<i>SC</i> <sub>7</sub>	.	.	.	.	.	.	.	×	×	×	.	.	.	.	.	.	×	.	×	.
<i>SC</i> <sub>8</sub>	.	.	.	.	.	×	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>9</sub>	.	.	×	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>10</sub>	.	.	.	.	.	.	.	×	×	×	.	.	.	.	.	.	.	×	×	×
<i>SC</i> <sub>11</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>12</sub>	.	.	.	.	.	.	×	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>13</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	×	×	×
<i>SC</i> <sub>14</sub>	×	.	.	.	.	.	.	.	.	.	.	×	.	.	.	.	.	.	.	.
<i>SC</i> <sub>15</sub>	.	.	×	.	.	.	×	×	×	×	.	.	.	.	.	.	×	.	.	.
<i>SC</i> <sub>16</sub>	.	.	.	.	×	.	×	.	×	×	.	.	.	.	.	.	.	×	×	×
<i>SC</i> <sub>17</sub>	.	.	.	.	.	.	.	.	.	×	.	.	.	.	.	.	.	.	.	.
<i>SC</i> <sub>18</sub>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	×	.	.	.	.
<i>SC</i> <sub>19</sub>	.	.	.	×	×	×	×	×	×	×	.	.	.	×	×	×	×	×	×	×
<i>SC</i> <sub>20</sub>	.	.	.	.	.	×	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Tabelle 4.11: Beste Teilmengen von *n* Deskriptoren für BP-Modelle

Verfahren	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
MLR	0,55334	0,89336	0,95302	0,96133	0,96734
ANN, 1HN	0,57729	0,30807	0,95443	0,96074	0,85732
ANN, 2HN	0,57958	0,89365	0,95443	0,96148	0,96838
ANN, 3HN	0,57981	0,88842	0,95380	0,96126	0,96632
SVM, lin	0,55046	0,62414	0,94951	0,95835	0,83026
SVM, pol	0,56943	0,71620	0,94885	0,95292	0,84941
SVM, rad	0,53903	0,59737	0,89956	0,89266	0,82629

Tabelle 4.12:  $R^2$  für Modellierung des BP durch verschiedene Regressionsverfahren

### Andere Deskriptoren und Regressionsverfahren

Wir wollen überprüfen, ob durch Verwendung geometrischer Deskriptoren bessere lineare Modelle berechnet werden können. Bei Hinzunahme der 35 geometrischen Indizes aus Anhang B und Berechnung bester linearer Modelle mit  $n = 1, \dots, 5$  Deskriptoren wird nur in einem Modell ein geometrischer Deskriptor einbezogen:

$n = 4$  Deskriptoren:  $W$ ,  $SC_2$ ,  $SC_{19}$ ,  $ssSHWD3$ ,

$$f = 1,1830X_0 - 6,3133X_1 + 2,3076X_2 + 0,23098X_3 + 70,914 \\ = 15,006X_0^* - 13,196X_1^* + 3,1416X_2^* + 0,88281X_3^* + 157,85.$$

Dabei ist  $R^2 = 0,96358$ ,  $R_{CV}^2 = 0,95549$ ,  $S = 1,4772$ ,  $S_{CV} = 1,63299$  und  $F = 297,61$ .

Ebenso liefern andere Regressionsverfahren kaum bessere Modelle. Tabelle 4.12 zeigt  $R^2$ -Werte für neuronale Netze mit 1–3 verborgenen Neuronen und Support-Vektor-Maschinen mit linearem, polynomialem ( $\text{degree} = 2$ ) und radialem Kernel. Um die Reproduzierbarkeit der ANN zu gewährleisten, wurden sie mit Startgewichten 0 trainiert. Deshalb ergeben sich in einigen Fällen sehr schlechte  $R^2$ . Dies ändert sich, wenn man eine zufällige Belegung der Startgewichte zulässt. Der Vergleich mit den linearen Modellen leidet natürlich an der Tatsache, dass bei der Berechnung der nichtlinearen Modelle keine Suche nach besten Teilmengen von Deskriptoren durchgeführt wurde. Es wurden lediglich die für beste LM gefundenen Deskriptoren-Sätze verwendet.

Der Algorithmus zur Berechnung von Regressionsbäumen verfolgt eine eigene Strategie zur Auswahl von Vorhersagevariablen (Abschnitt 3.2.4). Für das vorliegende Beispiel wurde ein RT mit 9 terminalen Knoten unter Verwendung von 5 Deskriptoren  $^1\chi^v$ ,  $^3\chi^v$ ,  $SC_6$ ,  $SC_{14}$ ,  $SC_{16}$  mit  $R^2 = 0,84386$  konstruiert. RT sind aufgrund ihres endlichen Wertebereichs eher dann konkurrenzfähig, wenn andere Verfahren kaum brauchbare Korrelationen liefern

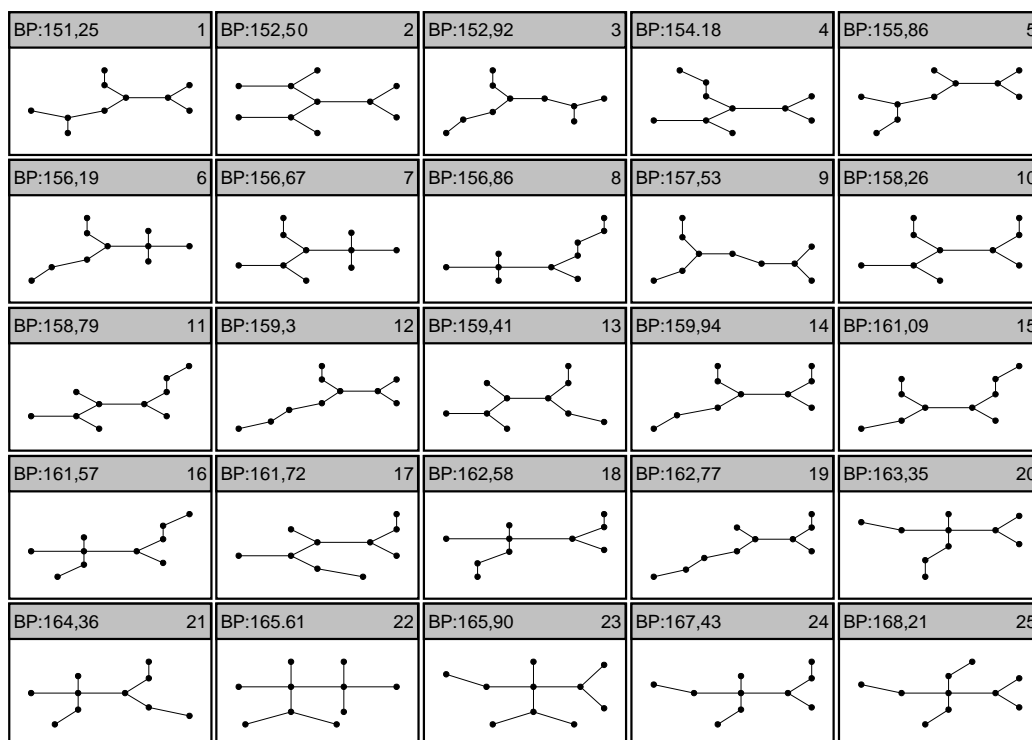


Abbildung 4.15: Rein virtuelle Bibliothek von Decanen mit vorhergesagten Siedepunkten

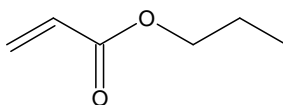
(siehe Abschnitt 4.4.3). Lineare Regression mit BSS für die Deskriptoren scheint für die Modellierung der BP von Decanen eine sehr geeignete Methode zu sein.

### Vorhersage

Insgesamt gibt es 75 Konstitutionsisomere zur Bruttoformel  $C_{10}H_{22}$ . Wir generieren diese virtuelle Bibliothek und entfernen mit Hilfe kanonischer Nummerierung die 50 Strukturen der realen Bibliothek. Zu den verbleibenden Verbindungen existieren laut *Beilstein*-Datenbank keine experimentellen Werte für die Siedepunkte oder die Verbindungen an sich sind nicht im Datenbestand enthalten. Wir wählen das beste 3-Deskriptoren-Modell mit TI und SC zur Vorhersage von Siedepunkten. Diese sind zusammen mit den Strukturen in Abbildung 4.15 aufgeführt. Die Strukturen sind dabei nach aufsteigenden berechneten Siedepunkten angeordnet.

#### 4.4.2 Beispiel: Physikalische Dichte von Propylacrylaten

Eine grundlegende Eigenschaft jeder Substanz ist ihre *physikalische Dichte* (kurz *PD*). Sie ist definiert als Quotient aus Masse und Volumen unter bestimmten äußeren Bedingungen, in unserem Fall 20°C und Normaldruck. Wir werden im Folgenden versuchen, PD (in g/cm<sup>3</sup>) als QSPR zu modellieren. Gegenstand unserer Untersuchung sind 166 Propylacrylate, die zusammen mit Angaben für PD in der *Beilstein*-Datenbank gefunden wurden. Die Substanzklasse der Propylacrylate wird durch folgende Substruktur definiert:



Da in der *Beilstein*-Datenbank aromatische Bindungen nicht markiert sind, liefert die Suche auch Verbindungen, bei denen die C=C-Doppelbindung Teil eines aromatischen Systems ist. Obwohl es sich bei diesen Verbindungen nicht um Propylacrylate im eigentlichen Sinne der Definition handelt, wollen wir sie trotzdem in unsere QSPR-Suche einbeziehen. 5 Verbindungen wurden wegen falscher Temperatur oder anderen zweifelhaften Angaben von unserer Untersuchung ausgeschlossen, eine zinnhaltige Verbindung wurde ebenfalls entfernt. Es verbleiben 160 Verbindungen in der realen Bibliothek.

#### Vorverarbeitung der Strukturen

Wir wollen in diesem Beispiel die Bewertung gefundener QSPR anhand einer Teststichprobe demonstrieren. Zu diesem Zweck teilen wir die Bibliothek per Zufall in einen Lern- und einen Testsatz von je 80 Verbindungen auf. Bevor wir mit unserer QSPR-Suche beginnen, müssen die Strukturen mehreren vorverarbeitenden Maßnahmen unterzogen werden:

- *H-Atome*: In dem von *Beilstein* exportierten *SD-File*<sup>5</sup> sind die Strukturen *ohne* Wasserstoff-Atome kodiert. Für die Berechnung mancher Deskriptoren und insbesondere zur Bestimmung einer 3D-Platzierung ist das Vorhandensein der H-Atome jedoch notwendig. H-Atome werden gemäß der bekannten Wertigkeiten der Nicht-Wasserstoff-Atome hinzugefügt.

---

<sup>5</sup>MDL SD-Files sind ein gebräuchliches Austauschformat für molekulare Strukturen. Eine Spezifikation des Formats ist bei [www.mdl.com](http://www.mdl.com) erhältlich.



Anzahl	H	C	N	O	F	Si	P	S	Cl	Br	SB	DB	TB	AB
1	0	0	31	0	0	3	14	7	13	4	0	33	1	0
2	0	0	2	53	0	0	0	0	5	1	0	56	0	0
3	0	0	0	22	7	0	0	0	1	0	0	46	0	0
4	0	0	0	43	0	0	0	0	1	0	0	18	0	0
5	0	0	0	11	0	0	0	0	1	0	10	3	0	8
6 – 10	12	60	0	31	0	0	0	0	0	0	70	5	0	48
11 – 15	53	64	0	0	0	0	0	0	0	0	50	0	0	12
16 – 20	46	23	0	0	0	0	0	0	0	0	25	0	0	1
21 – 25	21	10	0	0	0	0	0	0	0	0	3	0	0	0
26 – 30	19	2	0	0	0	0	0	0	0	0	2	0	0	0
≥ 31	9	1	0	0	0	0	0	0	0	0	0	0	0	0
$\Sigma$	160	160	33	160	7	3	14	7	21	5	160	160	1	69

Tabelle 4.13: Atomares Profil der realen Bibliothek von Propylacrylaten

- *Aromaten*: Manche Deskriptoren berücksichtigen aromatische Bindungen. Damit diese korrekt berechnet werden können, müssen aromatische Bindungen zuvor gemäß Definition 1.6.3 markiert werden.
- *3D-Platzierungen*: Wir wollen für die PD-Modellierung auch geometrische Deskriptoren heranziehen. Voraussetzung dafür ist eine dreidimensionale Platzierung, die wir gemäß Abschnitt 1.6.2 berechnen. Um unrealistische Platzierungen, insbesondere Durchdringungen von Ringssystemen zu vermeiden, wenden wir den Optimierungsalgorithmus für jede Struktur mehrfach an. Ausgehend von verschiedenen Zufallsplatzierungen wird nach der Energieoptimierung für diejenige Platzierung entschieden, welche den kleinsten Energiewert liefert. Damit können unrealistische Platzierungen zwar nicht generell ausgeschlossen, ihre Häufigkeit aber zumindest deutlich gesenkt werden.

Aus Platzgründen wollen wir unsere reale Bibliothek nicht im Einzelnen graphisch darstellen. Trotzdem möchten wir einen groben Überblick gewinnen. Tabelle 4.13 zeigt die elementare Zusammensetzung sowie die Anzahlen von Verbindungen klassifiziert nach Anzahlen von Einzel-, Doppel-, Dreifachbindungen (engl. *Single, Double, Triple Bond*, kurz *SB, DB, TB*) und aromatischen Bindungen (kurz *AB*). Wir nennen diese Darstellung *atomares Profil*. Die Bibliothek umfasst also Verbindungen mit 10 verschiedenen chemischen Elementen. Gemäß der vorgegebenen Substruktur enthalten alle Verbindungen Sauerstoff. Andere Heteroatome kommen seltener vor. 69 Verbindungen sind aromatisch.

		Min.	1. Quart.	Median	Mittel	3. Quart.	Max.
PD [g/cm <sup>3</sup> ]		0,873	1,010	1,066	1,081	1,116	1,534
	LS	0,873	1,009	1,049	1,080	1,116	1,534
	TS	0,883	1,016	1,073	1,081	1,114	1,453
<i>A (incl. H)</i>		18,00	27,00	33,00	35,74	43,00	82,00
	LS	18,00	24,75	34,00	35,75	43,25	70,00
	TS	20,00	28,00	33,00	35,73	41,00	82,00
<i>MW (incl. H)</i> [amu]		114,1	190,2	238,8	251,8	300,4	580,7
	LS	114,1	183,3	242,7	249,5	305,3	454,6
	TS	128,2	190,2	233,7	254,2	295,7	580,7
<i>V<sub>vdw</sub></i> [Å <sup>3</sup> ]		121,3	183,5	220,6	242,6	280,5	535,5
	LS	121,3	176,1	230,2	241,9	282,0	462,1
	TS	138,9	191,1	218,6	243,3	279,4	535,5

Tabelle 4.14: PD, Atomanzahl, Molekulargewicht und Van der Waals Volumen für die reale Bibliothek von Propylacrylaten

Tabelle 4.14 gibt Aufschluss über die Verteilungen von experimenteller Dichte sowie von Deskriptoren, die auf verschiedene Weisen die „Größe“ der einzelnen Moleküle beschreiben. Es sind jeweils Minimum, Maximum, erstes und drittes Quartil, Median und arithmetisches Mittel für die gesamte Bibliothek sowie für Lern- und Testsatz im Einzelnen angegeben. Wir können uns vergewissern, dass Lern- und Testsatz ähnlich zusammengesetzt sind. Die Anzahl von Atomen, das Molekulargewicht und das Van der Waals Volumen werden über Quotientenbildung wichtige Vorhersagevariablen für die Modellierung der PD liefern.

## Deskriptoren

Im Gegensatz zu dem Beispiel im vorherigen Abschnitt haben wir hier in unserer realen Bibliothek eine Vielzahl verschiedener Summenformeln. Es ist deshalb sinnvoll, arithmetische Indizes zu verwenden. Insgesamt stehen in *MOLGEN-QSPR* 48 arithmetische Indizes zur Verfügung (siehe Anhang B.3). Da sich in unserer Bibliothek keine Verbindungen mit Jod, Radikallstellen oder Ladungen befinden, liefern  $N_I$ , *rel. N<sub>I</sub>*, *rad* und *cha* keine relevanten Informationen. Des Weiteren gibt es nur eine Struktur mit einer Dreifachbindung. Wir werden deshalb auch *n#* und *rel. n#* vernachlässigen. Fluor-, schwefel- und bromhaltige Verbindungen treten nur mit sehr geringen Häufigkeiten in unserer Bibliothek auf, so dass eine gleichmäßige Verteilung auf Lern- und Testsatz fraglich ist. Wir entfernen deshalb auch  $N_F$ , *rel. N<sub>F</sub>*,

$N_S$ , *rel.*  $N_S$ ,  $N_{Br}$ , *rel.*  $N_{Br}$ . Es verbleiben 35 arithmetische Indizes.

Bei den topologischen Indizes wollen wir wieder auf die 30 zu Beginn von Abschnitt 4.4.1 aufgezählten Deskriptoren zurückgreifen, und wegen der gezeigten Identitäten auf  $M_1$  und  $M_2$  verzichten. Unter den verbleibenden 28 TI gibt es für die Bibliothek von Propylacrylaten keine paarweisen vollständigen Korrelationen. Vorsicht ist bei der Verwendung von *twc* geboten. Dessen Werte können stark exponentiell mit der Anzahl von Bindungen und der *topologischen Dichte*<sup>6</sup> eines molekularen Graphen wachsen. In unserer realen Bibliothek sind die vier größten Werte für *twc*:

69959869638977272  
2924196666599052  
130752536580352  
4707532422380

Der größte Werte ist über 10000 mal größer als der viertgrößte. Dies kann fatale Auswirkungen für lineare (und auch nichtlineare) Modelle haben: Falls Verbindungen mit den größten *twc* nicht im LS enthalten sind, kann es zu völlig unrealistischen Vorhersagen für diese Strukturen kommen. Wir lösen dieses Problem, indem wir *twc* durch dessen natürlichen Logarithmus  $\ln(\textit{twc})$  ersetzen.

Die physikalische Dichte wird mit Sicherheit in irgendeiner, uns unbekanntem Weise von der räumlichen Gestalt der einzelnen Moleküle abhängen. Deshalb schließen wir die 35, in Anhang B.3 zusammengestellten geometrischen Indizes in unsere QSPR-Suche ein. Insgesamt gehen wir also von 98 Deskriptoren aus.

Tabelle 4.15 zeigt einen Ausschnitt der Korrelationsmatrix von PD und den oben aufgezählten molekularen Deskriptoren. Vorzeichen der Korrelationskoeffizienten wurden dabei vernachlässigt. Korrelationskoeffizienten sind auf Basis *aller* Beobachtungen berechnet. Den besten Korrelationskoeffizienten zu PD hat die Van der Waals Dichte  $\rho_{vdw}$ , die als Quotient aus Molekulargewicht und Van der Waals Volumen definiert ist. Nahezu genauso gut mit PD korreliert das im Vergleich zu  $\rho_{vdw}$  mit deutlich geringerem algorithmischen Aufwand berechenbare durchschnittliche Atomgewicht *mean AW (incl. H)*. Dieses ist schlicht als Quotient aus dem Molekulargewicht und der Anzahl von Atomen definiert.  $\rho_{vdw}$  und *mean AW (incl. H)* sind ihrerseits mit einem Korrelationskoeffizienten von 0,980 ebenfalls stark korreliert. Wir werden die Auswirkungen dieser Korrelation bei der QSPR-Berechnung bemerken. Mit Ausnahme von  $\rho_{vdw}$  zeigen geometrische Indizes eher schwache Korrelationen zu PD.

---

<sup>6</sup>Unter der topologischen Dichte eines molekularen Graphen verstehen wir hier den Quotienten aus Bindungs- und Atomanzahl.

	$PD$	$\rho_{vdw}$	$mean\ AW$ ( <i>incl. H</i> )	$IC_0$	$rel. N_H$	$IC_1$	$mean\ AW$	$SIC_0$	$SIC_1$	$IC_2$	$N_{Cl}$
$PD$	1,000	0,937	0,934	0,801	0,787	0,784	0,772	0,634	0,620	0,514	0,498
$\rho_{vdw}$	0,937	1,000	0,980	0,825	0,715	0,706	0,902	0,722	0,634	0,365	0,601
$mean\ AW$ ( <i>incl. H</i> )	0,934	0,980	1,000	0,849	0,808	0,756	0,847	0,787	0,722	0,394	0,629
$IC_0$	0,801	0,825	0,849	1,000	0,778	0,851	0,628	0,823	0,715	0,446	0,615
$rel. N_H$	0,787	0,715	0,808	0,778	1,000	0,847	0,376	0,695	0,748	0,531	0,491
$IC_1$	0,784	0,706	0,756	0,851	0,847	1,000	0,432	0,689	0,796	0,725	0,445
$mean\ AW$	0,772	0,902	0,847	0,628	0,376	0,432	1,000	0,611	0,469	0,161	0,504
$SIC_0$	0,634	0,722	0,787	0,823	0,695	0,689	0,611	1,000	0,921	0,271	0,486
$SIC_1$	0,620	0,634	0,722	0,715	0,748	0,796	0,469	0,921	1,000	0,465	0,363
$IC_2$	0,514	0,365	0,394	0,446	0,531	0,725	0,161	0,271	0,465	1,000	0,065
$N_{Cl}$	0,498	0,601	0,629	0,615	0,491	0,445	0,504	0,486	0,363	0,065	1,000
$rel. N_{Cl}$	0,496	0,607	0,641	0,626	0,501	0,456	0,516	0,534	0,412	0,088	0,978
$G_2$	0,484	0,410	0,343	0,400	0,291	0,421	0,278	0,120	0,136	0,389	0,290
$G_1$	0,477	0,419	0,357	0,409	0,309	0,404	0,278	0,117	0,152	0,335	0,382
$rel. N_{Br}$	0,456	0,548	0,478	0,112	0,031	0,016	0,755	0,217	0,147	0,036	0,056
$N_{Br}$	0,455	0,542	0,471	0,107	0,028	0,013	0,747	0,201	0,132	0,042	0,054
$CIC_1$	0,430	0,492	0,583	0,528	0,576	0,562	0,405	0,889	0,942	0,253	0,282
$G_2$ ( <i>incl. H</i> )	0,429	0,351	0,280	0,345	0,230	0,368	0,233	0,185	0,201	0,365	0,249
$G_1$ ( <i>incl. H</i> )	0,423	0,360	0,293	0,352	0,246	0,350	0,234	0,183	0,216	0,316	0,331
$rel. N_O$	0,413	0,344	0,337	0,433	0,380	0,373	0,187	0,408	0,374	0,084	0,094
$MW$	0,378	0,288	0,226	0,284	0,212	0,292	0,162	0,263	0,291	0,310	0,242
$SIC_2$	0,374	0,390	0,478	0,401	0,503	0,529	0,317	0,756	0,873	0,508	0,104
$\lambda_1^A$	0,344	0,214	0,220	0,279	0,399	0,508	0,017	0,090	0,056	0,473	0,221
$MW$ ( <i>incl. H</i> )	0,332	0,241	0,175	0,239	0,162	0,249	0,127	0,310	0,337	0,289	0,207

Tabelle 4.15: Ausschnitt der Korrelationsmatrix für die physikalische Dichte und Deskriptoren der realen Bibliothek von Propylacrylaten

### Lineare Modellierung durch beste Teilmengen–Selektion

Wir berechnen basierend auf den Beobachtungen in LS die besten linearen Modelle bzgl.  $R_{LS}^2$  mit  $n = 1, \dots, 5$  Deskriptoren:

$n = 1$  Deskriptor:  $\rho_{vdw}$ ,

$$f = 0,89996X_0 + 0,14597,$$

$$R_{LS}^2 = 0,87968, S_{LS} = 0,047187, F_{LS} = 570,26.$$

$n = 2$  Deskriptoren:  $IC_2, \rho_{vdw}$ ,

$$f = 0,12217X_0 + 0,81813X_1 - 0,21835,$$

$$R_{LS}^2 = 0,91688, S_{LS} = 0,039474, F_{LS} = 424,68.$$

$n = 3$  Deskriptoren:  $N_O, rel. n - (incl. H), \rho_{vdw}$ ,

$$f = 0,015912X_0 - 0,30250X_1 + 0,84571X_2 + 0,39659,$$

$$R_{LS}^2 = 0,93813, S_{LS} = 0,034280, F_{LS} = 384,12.$$

$n = 4$  Deskriptoren:  $A, rel. N_O, {}^0\chi, \rho_{vdw}$ ,

$$f = 0,064792X_0 + 0,67865X_1 - 0,084221X_2 + 0,89515X_3 + 0,062349,$$

$$R_{LS}^2 = 0,95060, S_{LS} = 0,030836, F_{LS} = 360,76.$$

	$n = 5$	$n = 4$	$n = 3$	$n = 2$	$n = 1$
LS	0,95481	0,95059	0,93813	0,91688	0,87968
TS	0,92018	0,92645	0,93244	0,89716	0,87341
LS	0,95474	0,94903	0,93770	0,91524	0,87278
TS	0,92266	0,91832	0,89164	0,91582	0,87117
LS	0,95440	0,94753	0,93724	0,91415	0,60970
TS	0,93862	0,93797	0,92980	0,88833	0,57425
LS	0,95397	0,94716	0,93720	0,91241	0,59306
TS	0,94236	0,91611	0,92598	0,90536	0,69813
LS	0,95374	0,94664	0,93712	0,91082	0,57784
TS	0,94436	0,90873	0,93826	0,90041	0,67580

Tabelle 4.16: Bestimmtheitsmaß für LS und TS der besten PD-Modelle bzgl.  $R_{LS}^2$  mit  $n$  Deskriptoren

$n = 5$  Deskriptoren:  ${}^1\chi$ ,  ${}^0\chi^v$ ,  $CIC_0$ ,  $IC_1$ ,  $\rho_{vdw}$ ,

$$f = 0,034734X_0 - 0,041353X_1 + 0,14269X_2 + 0,12314X_3 + 1,0380X_4 - 0,66073,$$

$$R_{LS}^2 = 0,95481, S_{LS} = 0,029690, F_{LS} = 312,70.$$

Tabelle 4.16 vergleicht  $R_{LS}^2$  und  $R_{TS}^2$  für die jeweils 5 besten linearen Modelle mit  $n = 1, \dots, 5$  Deskriptoren.  $R_{LS}^2$  ist erwartungsgemäß für die meisten Modelle größer als  $R_{TS}^2$ . Zudem sehen wir, dass beste Modelle bzgl. LS in der Regel nicht beste Werte bzgl. TS besitzen. Abbildung 4.16 visualisiert dieses Phänomen. Für  $n = 1$  Deskriptor liegen dabei nur die zwei besten Modelle innerhalb des zur Darstellung gewählten Bereichs für das Bestimmtheitsmaß. Unter den betrachteten QSPR hat das Modell mit größtem  $R_{TS}^2$

$n = 5$  Deskriptoren:  $A$ ,  $rel. N_O$ ,  $W$ ,  ${}^0\chi$ ,  $\rho_{vdw}$ ,

$$f = 0,065886X_0 + 0,68347X_1 - 0,000039413X_2 - 0,079592X_3 + 0,88018X_4 + 0,028970,$$

$$R_{LS}^2 = 0,95374, S_{LS} = 0,030041, F_{LS} = 305,10.$$

Abbildung 4.17 zeigt zu diesem Modell gemessene und vorhergesagte PD für LS und TS als Scatterplot.

Die verwendeten Deskriptoren für beste LM bzgl. LS kann man Tabelle 4.17 entnehmen. Jeder Spalte ist dabei ein Modell zugeordnet. Im Spaltenkopf steht jeweils  $R_{LS}^2$ . Die Modelle sind von links nach rechts bzgl. fallender  $R_{LS}^2$  angeordnet. Die Zeilen repräsentieren Deskriptoren, Kreuze an Position  $(i, j)$  bedeuten, dass der Deskriptor in Zeile  $i$  für das Modell aus Spalte  $j$  benötigt wird. Die Tabelle reflektiert die hohe Korrelation von PD,  $\rho_{vdw}$  und  $mean AW$  (*incl. H*): In allen Modellen mit  $n \geq 2$  Deskriptoren und

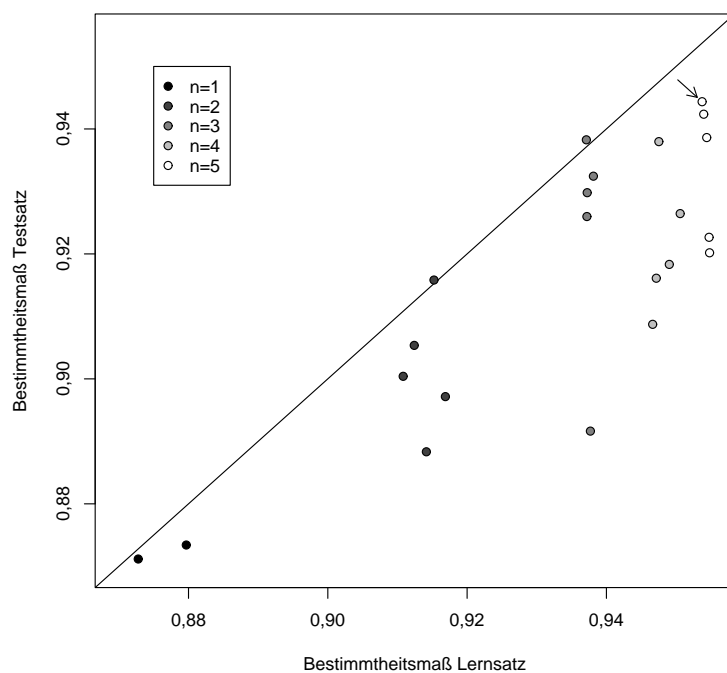


Abbildung 4.16: Scatterplot von  $R_{LS}^2$  und  $R_{TS}^2$  für die besten LM bzgl.  $R_{LS}^2$  mit  $n = 1, \dots, 5$  Deskriptoren

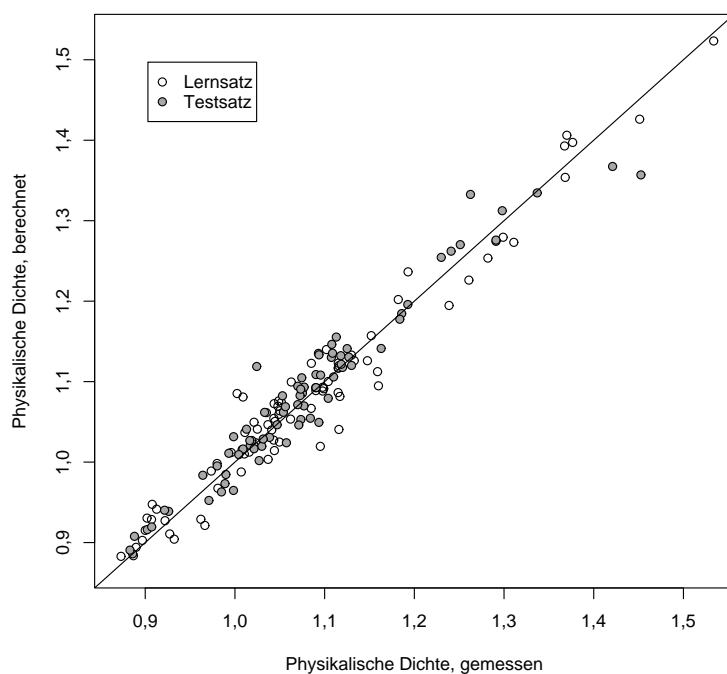


Abbildung 4.17: Scatterplot gemessener und vorhergesagter PD für das markierte Modell aus Abbildung 4.16

	$n = 5$					$n = 4$				$n = 3$				$n = 2$		$n = 1$									
$R_{LS}^2 \rightarrow$	0,95481	0,95474	0,95440	0,95397	0,95374	0,95059	0,94903	0,94753	0,94716	0,94664	0,93813	0,93770	0,93724	0,93720	0,93712	0,91688	0,91524	0,91415	0,91241	0,91082	0,87968	0,87278	0,60970	0,59306	0,57784
$\rho_{vdw}$	x	x	x	x	x	x	.	.	.	.	x	x	x	x	x	x	.	.	.	.	x	.	.	.	.
$\chi^1$	x	x	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.
$\chi^0$	x	x	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.
$CIC_0$	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$IC_1$	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$SIC_1$	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$rel. N_O$	.	.	x	x	x	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$B$	.	.	x	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$\ln(twc)$	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$\lambda^A$	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$\chi^0$	.	.	.	x	x	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$A$	.	.	.	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$MTI$	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$W$	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$N_O$	.	.	.	.	.	.	x	x	x	x	x	.	.	x	.	.	.	x	.	.	.	.	.	.	.
$mean AW (incl. H)$	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.	x	.	x	x	.	x	.	.	.
$C$	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$rel. N_{Cl}$	.	.	.	.	.	.	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$N_{Cl}$	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$J$	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$rel. n - (incl. H)$	.	.	.	.	.	.	.	.	.	.	x	.	.	.	x	.	.	.	.	.	.	.	.	.	.
$rel. n arom. (incl. H)$	.	.	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	.	.
$HBA$	.	.	.	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	.
$IC_2$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$mean AW$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$SHDW1$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$IC_0$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x
$rel. N_H$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x

Tabelle 4.17: Beste Teilmengen von  $n$  Deskriptoren für PD-Modelle

den beiden besten 1-Deskriptor-Modellen sind jeweils entweder  $\rho_{vdw}$  oder  $mean AW (incl. H)$ , nie aber beide enthalten.

### Lineare Modellierung durch schrittweise Teilmengen-Selektion

Es ist zu erwarten, dass unter Verwendung von mehr als 5 Deskriptoren Modelle mit höherem  $R_{TS}^2$  gefunden werden können. Leider stoßen wir bei der großen Grundmenge von 98 Deskriptoren recht bald an die Grenze der algorithmischen Möglichkeiten bei der Suche nach besten Teilmengen von Deskriptoren. Wir wollen an dieser Stelle von dem, in Abschnitt 3.1.4 beschriebenen schrittweisen Verfahren zur Variablen-Selektion Gebrauch machen.

Wir generieren Modelle mit  $n = 1, \dots, 20$  Deskriptoren. In jedem Schritt werden 50 Teilmengen von Deskriptoren ausgewählt, für die man beste lineare Modelle (bzgl.  $R_{LS}^2$ ) erhält. Diese werden auf alle Möglichkeiten um je einen Deskriptor erweitert, wiederum die besten 50 LM selektiert u.s.w. Als Ergebnis erhält man für jede Anzahl  $n = 1, \dots, 20$  von Deskriptoren 50 Modelle.

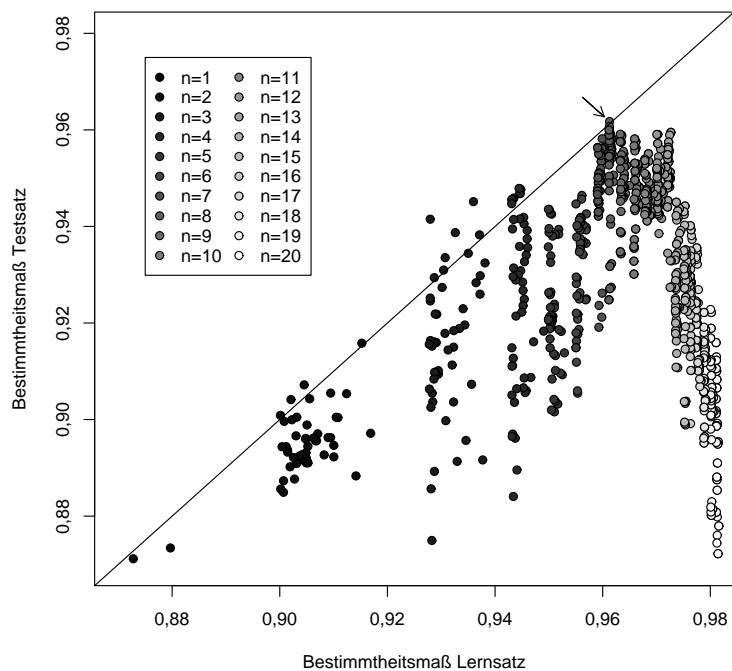


Abbildung 4.18: Scatterplot von  $R^2_{LS}$  und  $R^2_{TS}$  für die besten LM bzgl.  $R^2_{TS}$  konstruiert nach 50-fachem schrittweisem Verfahren

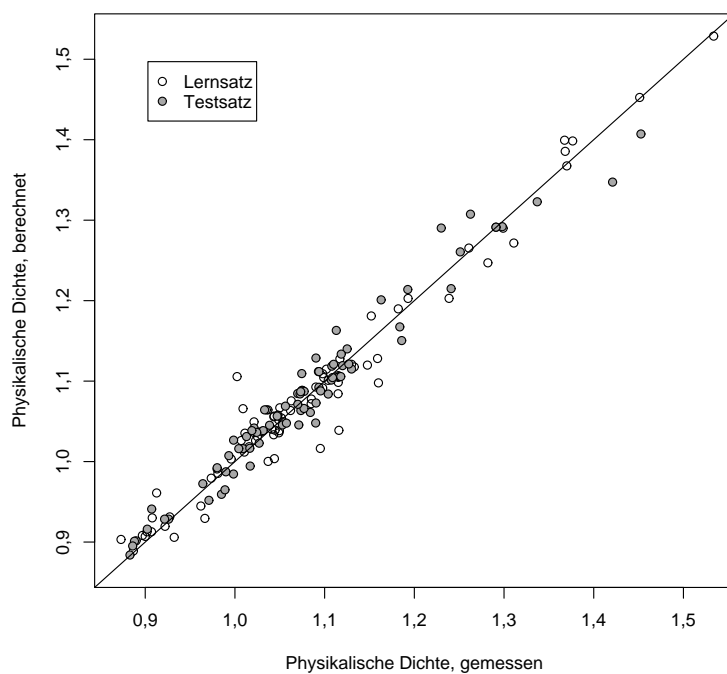


Abbildung 4.19: Scatterplot gemessener und vorhergesagter PD für das markierte Modell aus Abbildung 4.18



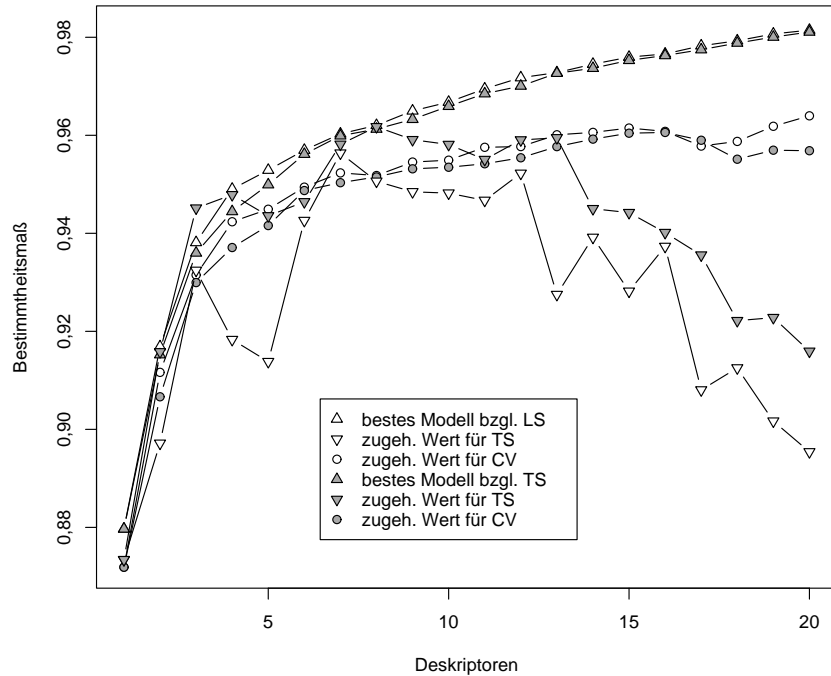


Abbildung 4.20: Beste Modelle bzgl.  $R_{LS}^2$  (bzw.  $R_{TS}^2$ ) und zugehörige Werte für  $R_{TS}^2$  (bzw.  $R_{LS}^2$ ) sowie  $R_{CV}^2$  nach Anzahl von Deskriptoren

Obwohl diese nicht notwendig die bestmöglichen LM für jedes  $n$  umfassen, ist diese Vorgehensweise geeignet, um mit geringem Rechenaufwand gute Modelle zu erhalten.

In Abbildung 4.18 ist  $R_{TS}^2$  gegen  $R_{LS}^2$  für die so bestimmten Modelle abgetragen. Die Färbung der Datenpunkte spiegelt die Anzahl der Deskriptoren wieder: Modelle mit einem Deskriptor sind schwarz, Modelle mit 20 Deskriptoren weiß und alle anderen in entsprechenden Graustufen dargestellt. Bei den Modellen mit hoher Anzahl von Deskriptoren ( $n > 13$ ) ist Overfitting zu erkennen: Die Werte von  $R_{TS}^2$  sind deutlich schlechter als die für  $R_{LS}^2$ . In Abbildung 4.20 wurden für jedes  $n = 1, \dots, 20$  das beste Modell bzgl.  $R_{LS}^2$  ausgewählt und der zugehörige Wert für  $R_{TS}^2$  eingetragen. Ebenso wurden die besten Modelle bzgl.  $R_{TS}^2$  und deren  $R_{LS}^2$  bestimmt. Man sieht in beiden Fällen, dass die Werte für LS mit wachsendem  $n$  streng monoton steigen, wohingegen das Bestimmtheitsmaß für TS bei 7 bzw. 8 Deskriptoren sein Maximum annimmt. Bis etwa 13 Deskriptoren bleiben die Werte für TS auf hohem Niveau, fallen dann aber wieder ab.

Zur Vorhersage von PD entscheiden wir uns für das Modell mit größtem  $R_{TS}^2$ . Es verwendet

	$R_{TS}^2 \rightarrow$	0,96174	0,96069	0,96004	0,95997	0,95949	0,95949	0,95939	0,95910	0,95906	0,95905	0,95905	0,95896	0,95896	0,95895	0,95893	0,95892	0,95891	0,95883	0,95881	0,95863	0,95854	0,95818	0,95816	0,95815
<i>mean AW (incl. H)</i>		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
<i>rel. N<sub>Cl</sub></i>		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
<i>N<sub>O</sub></i>		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
<i>C</i>		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
<i>mwc<sup>(3)</sup></i>		x	.	.	x	x	x	x	x	x	x	x	x	x	x	x	.	x	x	x	.	x	x	x	x
<i>CIC<sub>0</sub></i>		x	.	.	x	x	x	x	.	x	x	x	x	x	.	x	.	x	.	x	.	x	x	x	x
<i>IC<sub>1</sub></i>		x	.	.	x	x	x	x	.	.	.	x	.	x	.	x	.	x	.	x	.	x	x	x	x
<i>N<sub>C</sub></i>		x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>mwc<sup>(6)</sup></i>		.	x	x	.	x	x	.	.	.	x	x	.	.	.	.	x	x	.	.	x	.	.	.	.
<i>MW (incl. H)</i>		.	x	x	.	.	.	.	x	.	x	.	.	.	x	.	x	.	x	.	x	x	.	.	.
<i>mwc<sup>(5)</sup></i>		.	x	x	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	x	.	.	.	.
<i>MTI</i>		.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	x	.	.	x	.	.	.
<i>W</i>		.	.	x	.	.	.	.	x	.	.	.	.	.	x	.	.	.	.	.	x	.	.	.	.
<i>rel. n aromatic</i>		.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>IC<sub>0</sub></i>		.	.	.	.	x	x	.	.	x	x	x	x	.	.	.	x	.	.	.	.	.	.	.	.
<i><sup>1</sup>χ<sup>v</sup></i>		.	.	.	.	x	x	.	.	x	x	.	.	.	.	.	x	.	.	.	x	.	.	.	.
<i><sup>0</sup>χ</i>		.	.	.	.	x	x	.	.	x	x	.	.	.	.	.	x	.	.	.	.	.	.	.	.
<i>mwc<sup>(8)</sup></i>		.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.
<i><sup>3</sup>χ<sup>v</sup></i>		.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.
<i>CIC<sub>1</sub></i>		.	.	.	.	.	x	.	.	.	.	.	x	x	.	.	.	.	.	.	.	.	.	.	.
<i>SIC<sub>1</sub></i>		.	.	.	.	.	.	x	.	x	x	.	.	.	.	.	.	x	.	.	.	.	.	.	.
<i>rel. n-</i>		.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.
<i>SHDW5</i>		.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>rel. N<sub>O</sub></i>		.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.
<i>MW</i>		.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>HBD</i>		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.
<i>rel. n - (incl. H)</i>		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
<i>SAS<sub>H<sub>2</sub>O</sub></i>		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.
<i>n-</i>		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x
Anzahl Deskriptoren		8	8	8	8	13	13	8	9	8	12	12	8	8	8	8	8	8	9	13	8	8	9	8	8

Tabelle 4.18: Deskriptoren der 25 besten PD-Modelle bzgl.  $R_{TS}^2$  berechnet durch 50-fache schrittweise Selektion

$n = 8$  Deskriptoren:

$$\begin{aligned}
 &N_C, N_O, \text{rel. } N_{Cl}, C, \text{mean AW (incl. H)}, mwc^{(3)}, CIC_0, IC_1, \\
 &f = -0,0071639X_0 + 0,018563X_1 - 0,78530X_2 + 0,047740X_3 + \\
 &\quad + 0,11221X_4 - 0,00060864X_5 + 0,12967X_6 + 0,072563X_7 - 0,27686, \\
 &R_{LS}^2 = 0,96127, S_{LS} = 0,028062, F_{LS} = 220,25.
 \end{aligned}$$

Kritisch betrachtet werden muss hierbei die Tatsache, dass für dieses Modell  $R_{TS}^2 = 0,96174 > R_{LS}^2$  ist. Abbildung 4.19 zeigt gemessene und berechnete Werte für dieses Modell. Beobachtungen des LS sind weiß, die des TS grau markiert. Für die 25 besten Modelle bzgl. TS sind in Tabelle 4.18 die verwendeten Deskriptoren zusammengestellt. Darunter sind mit Ausnahme von *SHDW5* und *SAS<sub>H<sub>2</sub>O</sub>* keine geometrischen Deskriptoren vertreten. In allen Modellen wird *mean AW (incl. H)* verwendet. Wegen der starken Korrelation zwischen *mean AW (incl. H)* und  $\rho_{vdw}$  tritt die Van der Waals Dichte

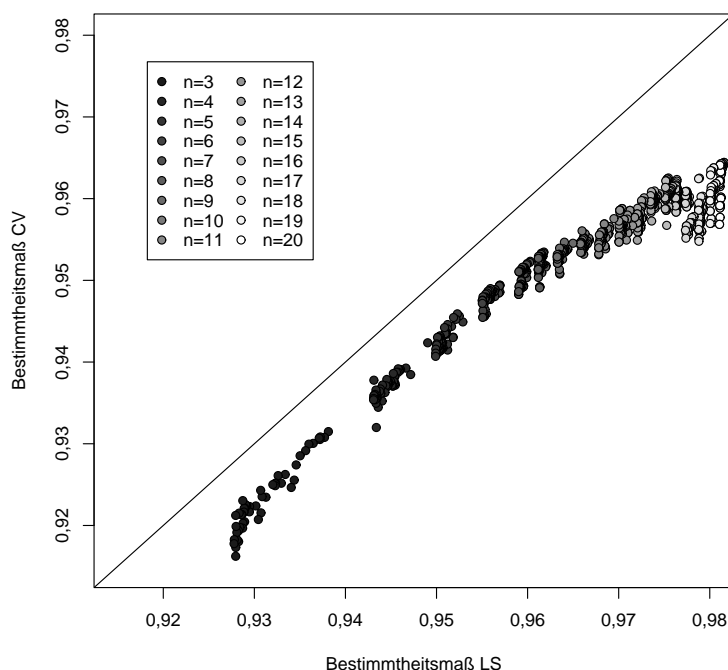


Abbildung 4.21: Scatterplot von  $R_{LS}^2$  und  $R_{CV}^2$  für die besten LM bzgl.  $R_{LS}^2$  konstruiert nach 50-fachem schrittweisem Verfahren

nicht auf. Dagegen wurden vermehrt arithmetische Deskriptoren verwendet, insbesondere Anzahlen und relative Anzahlen von Atomen verschiedener Elemente. Auch dass die zyklomatische Zahl als Deskriptor zur Modellierung der physikalischen Dichte herangezogen wird, ist nachvollziehbar. Schließlich hängt insbesondere von der Anzahl der Kreise ab, inwiefern sich die Atomkugeln überlappen. Sie nimmt damit Einfluss auf Van der Waals Volumen und Dichte.

Wir haben gesehen, dass mit Hilfe arithmetischer und geometrischer Deskriptoren und OLS die physikalische Dichte von Propylacrylaten gut modellierbar ist. Dies kommt natürlich insbesondere der Anwendbarkeit der Modelle zugute, da auf algorithmisch aufwändige Berechnung von 3D-Platzierungen und geometrischer Deskriptoren verzichtet werden kann. Anhand einer Teststichprobe wurde für diese Beispiel nachgewiesen, dass bei Verwendung von mehr als 13 Deskriptoren die Vorhersagefähigkeit der berechneten Modelle deutlich nachlässt. Es kommt zu Overfitting.

An dieser Stelle wäre es interessant zu untersuchen, ob auch mittels Kreuzvalidierung dieser Effekt erkennbar gewesen wäre. Wir haben für die Deskriptoren zu den schrittweise berechneten Modellen eine LOO-CV auf dem Lernsatz durchgeführt. Die  $R_{CV}^2$ -Werte für die besten Modelle bzgl. LS und TS wurden ebenfalls in Abbildung 4.20 eingetragen. Abbildung 4.21 zeigt einen

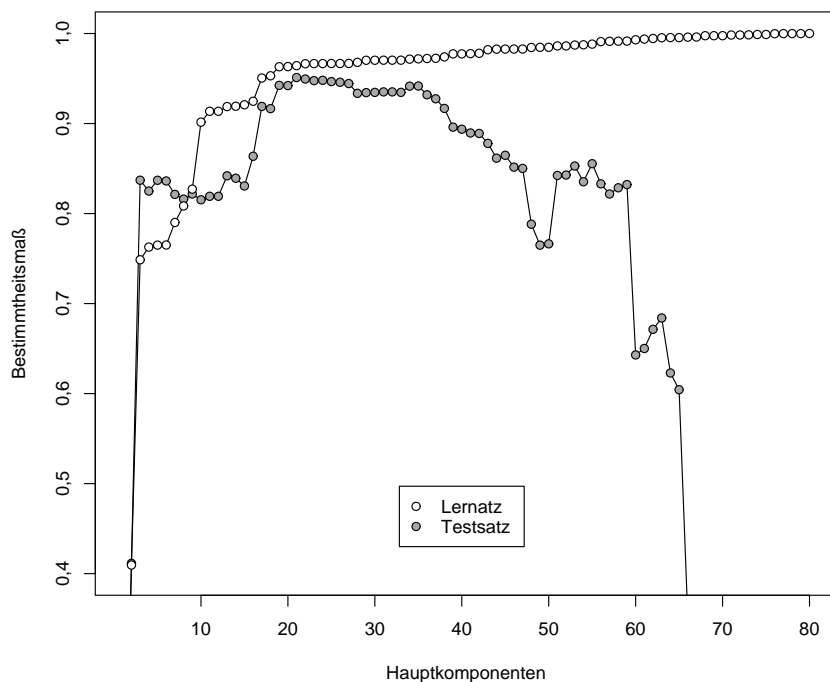


Abbildung 4.22:  $R_{LS}^2$  und  $R_{TS}^2$  für LM bestimmt durch PCR in Abhängigkeit der Anzahl verwendeter Hauptkomponenten

Scatterplot der  $R^2$ -Werte für LS und CV der Modelle mit  $n > 3$  Deskriptoren. Offenbar ist für  $R_{CV}^2$  kein ähnlich deutliches Absinken wie für  $R_{TS}^2$  ab  $n = 14$  Deskriptoren zu erkennen. Bei der Auswahl von Deskriptoren-Teilmenen durch Kreuzvalidierung ist hier Vorsicht geboten!

### Lineare Modellierung durch Hauptkomponenten-Regression

Mit autoskalierten Deskriptoren- und Eigenschaftswerten wurde eine PCR durchgeführt. Abbildung 4.22 zeigt die  $R^2$ -Werte in Abhängigkeit von der Anzahl verwendeter Hauptkomponenten für LS und TS. Für das beste Modell bzgl. TS ist  $R_{TS}^2 = 0,95111$  und  $R_{LS}^2 = 0,96428$ . Es verwendet 21 Hauptkomponenten. Hinsichtlich seiner Vorhersagefähigkeit ist es schlechter als die zuvor mit OLS und schrittweiser Deskriptoren-Selektion bestimmten Modelle. Die Komplexität der Modelle kann aufgrund der verschiedenen Methoden nicht direkt verglichen werden. Es ist jedoch zu bedenken, dass zur Vorhersage mit Hilfe des PCR-Modells zunächst alle 98 Deskriptoren berechnet werden müssen, die dann gemäß der Zusammensetzung der Hauptkomponenten und ihrer Gewichtung den Vorhersagewert liefern. Dieser Mehraufwand spricht in jedem Fall für das zuvor bestimmte 8-Deskriptoren-Modell.

### 4.4.3 Beispiel: Antibakterielle Aktivität von Quinolonen

Im Folgenden Beispiel wollen wir Struktur-Eigenschafts-Beziehungen für eine biologische Eigenschaft, die *antimycobakterielle Aktivität* (kurz *ABA*) von Quinolon-Derivaten berechnen. Die untersuchten Verbindungen beschreibt folgende generische Strukturformel:

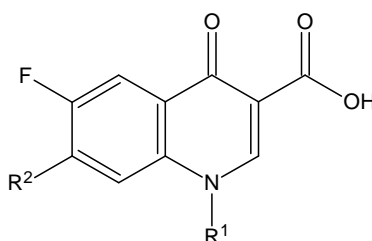


Abbildung 4.23 zeigt die Substituenten für  $R^1$  (obere Reihe) und  $R^2$  (untere drei Reihen). Das mit „Z“ beschriftete Atom repräsentiert dabei das Zentralmolekül (s.o.). ABA wird angegeben durch die minimale Konzentration (in  $\mu\text{g/ml}$ ), die notwendig ist, um eine hemmende Wirkung gegen *Mycobacterium fortuitum* zu erreichen (engl. *Minimum Inhibitory Concentration*, kurz *MIC*).

Substituent $R^2$	Substituent $R^1$					
	1	2	3	4	5	6
1	0,50	1,00	0,03	0,06	0,50	0,25
2				0,06		0,25
3	1,00	0,50	0,03	0,06	0,25	0,13
4	0,25	1,00	0,13	0,06	0,50	0,50
5				0,13		0,25
6	0,13	0,50	0,06	0,06	0,13	0,25
7				0,25		0,50
8				1,00		2,00
9				1,00		1,00
10	1,00	1,00	0,06	0,03	1,00	0,13
11				0,03		
12				0,13		
13	2,00	2,00	0,13	0,25	2,00	0,13
14				0,50		0,50
15				0,25		

Tabelle 4.19: Gemessene MIC für die reale Bibliothek von Quinolonen

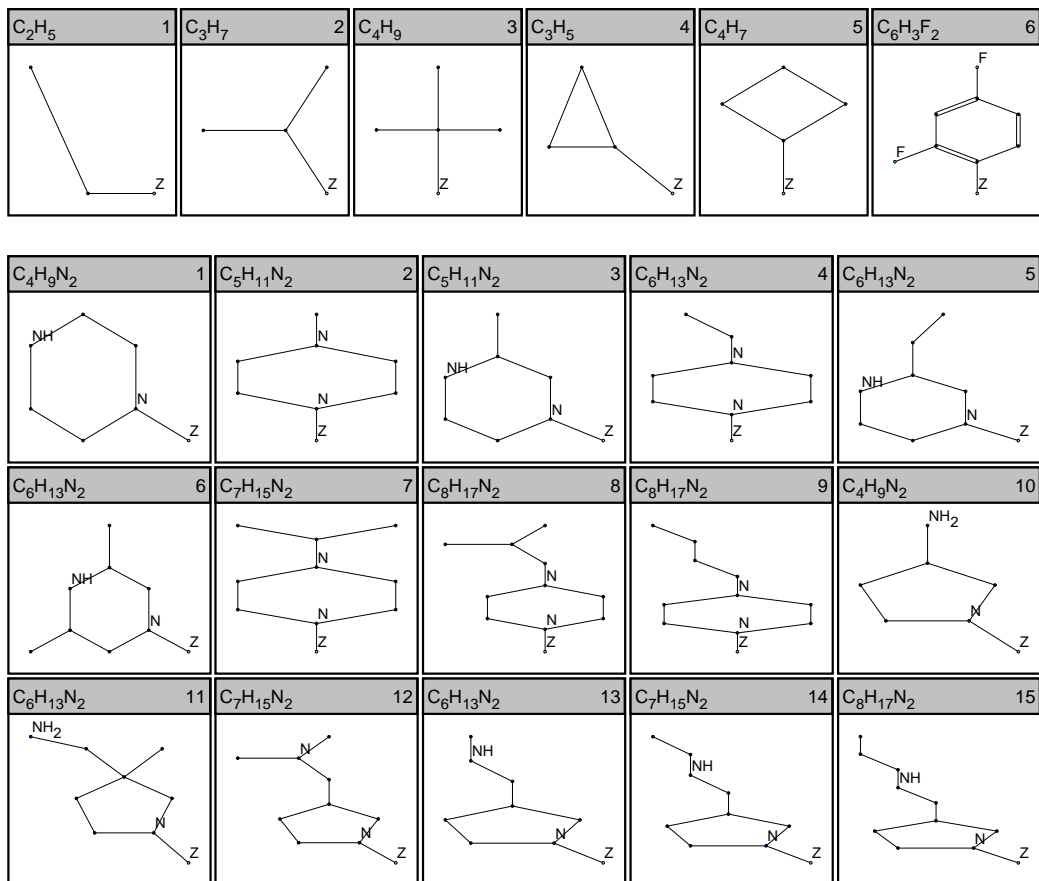
Abbildung 4.23: Substituenten für  $R^1$  (obere Reihe) und  $R^2$ 

Tabelle 4.19 zeigt die reale Bibliothek von  $m = 51$  Quinolonen in Form von Paaren aus  $R^1$  und  $R^2$  sowie die zugehörigen gemessenen MIC (nach [113]). Wir werden auf verschiedene Weisen Modelle zur Vorhersage von MIC aufstellen.

### Deskriptoren

Ausgangspunkt für unsere Untersuchungen bilden die 48 arithmetischen und 170 topologischen Indizes aus Anhang B. Von diesen 218 Deskriptoren entfernen wir 25, die auf der realen Bibliothek konstante Werte annehmen:

$$N_O, N_N, N_S, \text{rel. } N_S, N_{Cl}, \text{rel. } N_{Cl}, N_{Br}, \text{rel. } N_{Br}, N_I, \text{rel. } N_I, N_P, \text{rel. } N_P, n=, n\#, \text{rel. } n\#, \text{rel. } n\# (\text{incl. } H), \text{cha}, \text{rad}, HBA, {}^7rings, {}^8rings, \geq^9rings, T_3, \text{con. comp.}, \text{gt planar}$$

Unter den verbleibenden 193 Indizes suchen wir nach vollständigen Korrelationen. Vollständige Korreliertheit ist eine Äquivalenzrelation. Im Folgenden sind alle 19 Äquivalenzklassen mit mehr als einem Element aufgeführt. Die meisten der vollständigen Korrelationen ergeben sich aus der speziellen Beschaffenheit der realen Bibliothek. Solche, die allgemeine Gültigkeit haben, sind durch „ $\simeq$ “ ausgezeichnet:

$$\begin{aligned} &rel.N_O \sim rel.N_N, N_F \sim n_{aroma}, B \sim loc.B, B(incl.H) \sim \\ &loc.B(incl.H), C \simeq rings, M_1 \simeq mwc^{(2)}, M_2 \simeq mwc^{(3)}, {}^0\chi \sim \\ &{}^0\chi^s, {}^1\chi \sim {}^1\chi^s, {}^2\chi \sim {}^2\chi^s, {}^3\chi^s \sim {}^3\chi_p, {}^3\chi^s(cluster) \sim {}^3\chi_c, {}^3\chi^v \sim \\ &{}^3\chi_p^v, F \simeq N_{GS} \simeq {}^2P_{acyc} \simeq {}^2P, {}^7P_{acyc} \sim {}^7P, {}^8P_{acyc} \sim {}^8P, \\ &{}^{\geq 9}P_{acyc} \sim {}^{\geq 9}P, {}^3rings \sim {}^3\chi_{ch} \sim {}^3\chi_{ch}^v, {}^6\chi_c \sim {}^6\chi_c^v. \end{aligned}$$

Aus jeder dieser Äquivalenzklassen verwenden wir jeweils nur den erstgenannten Deskriptor, die verbleibenden 22 werden von den weiteren Untersuchungen ausgeschlossen. Es verbleiben 171 nicht konstante, paarweise nicht vollständig korrelierte Indizes.  $twc$  und  $twc_{unsat}$  ersetzen wir wie zuvor in Abschnitt 4.4.2 durch ihre natürlichen Logarithmen.

### Regression

Wir wollen in diesem Beispiel verschiedene Methoden des überwachten Lernens demonstrieren. Zunächst betrachten wir ABA als kontinuierliche Variable, repräsentiert durch MIC, und ermitteln Vorhersagefunktionen durch Regression.

Wie in den vorangegangenen Abschnitten bestimmen wir beste lineare Modelle mit OLS-Regression und BSS. Auf diese Weise erhalten wir als bestes Modell mit

$$\begin{aligned} n = 5 \text{ Deskriptoren: } &X_0 = {}^2\kappa_\alpha, X_1 = {}^6P_{acyc}, X_2 = ch, J_4, X_3 = {}^6\chi_p, X_4 = {}^4\chi_c, \\ &f = 1,7782X_0 + 0,26966X_1 + 142,04X_2 - 8,9882X_3 - 5,8961X_4 - 11,528, \\ &R^2 = 0,64597, S = 0,34676, F = 16,421. \end{aligned}$$

Offenbar lassen sich für MIC nicht ähnlich gute lineare Modelle aufstellen, wie wir dies für BP und PD gesehen haben. Eine Ursache dafür ist, dass MIC nur 7 verschiedene Werte mit folgenden Häufigkeiten annimmt:

MIC	0,03	0,06	0,13	0,25	0,50	1,00	2,00
Häufigkeit	4	7	9	9	9	9	4

Wir untersuchen, ob Regressionsbäume besser geeignet sind, um MIC zu modellieren. Zu den Standard-Parametern (`mincut` = 5, `minsize` = 10, `mindev` = 0,01) berechnet  $R$  einen RT (Abbildung 4.24) mit 7 terminalen Knoten und

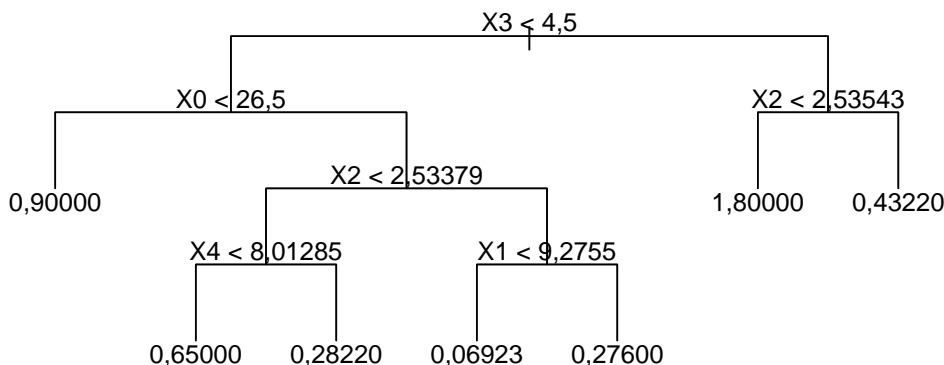


Abbildung 4.24: Regressionsbaum für MIC mit 5 Deskriptoren

$n = 5$  Deskriptoren:  $X_0 = B$ ,  $X_1 = {}^1\chi^v$ ,  $X_2 = \lambda_1^A$ ,  $X_3 = FRB$ ,  $X_4 = {}^5\chi_{pc}$ ,  
 $R^2 = 0,81748$ .

Er zeigt damit eine wesentlich höhere Anpassung an die experimentellen Werte als das beste lineare Modell mit gleicher Anzahl von Deskriptoren.

Wir verwenden die beiden Deskriptorensätze *LM* und *RT*, die für das beste 5-Deskriptoren-LM und den RT bestimmt wurden, zur Berechnung neuronaler Netze und Support-Vektor-Maschinen. Tabelle 4.20 enthält die  $R^2$ -Werte für SVM mit linearem, polynomialem (*degree* = 2) und radialem Kernel sowie für ANN mit einem, zwei und drei HN. Um die Reproduzierbarkeit der ANN zu gewährleisten, wurden Startgewichte 0 gewählt. Initialisiert man die Startgewichte per Zufall, so erhält man in der Regel bessere Modelle. Auf diese Weise wurde für Deskriptorensatz RT ein ANN mit zwei HN und  $R^2 = 0,75380$  gefunden. Das beste Modell war ein ANN mit drei HN und  $R^2 = 0,87415$  unter Verwendung des Deskriptorensatzes RT:

$n = 5$  Deskriptoren:  $X_0 = B$ ,  $X_1 = {}^1\chi^v$ ,  $X_2 = \lambda_1^A$ ,  $X_3 = FRB$ ,  $X_4 = {}^5\chi_{pc}$ ,  
 $f^* = -4,40/(1+\exp(-1,88X_0^*+1,10X_1^*-7,41X_2^*-18,5X_3^*-3,81X_4^*+19,4))+$   
 $+4,43/(1+\exp(-2,23X_0^*-1,38X_1^*+1,19X_2^*-17,8X_3^*+0,709X_4^*+11,1))+$   
 $+0,309/(1+\exp(-6,65X_0^*+3,88X_1^*+21,2X_2^*-3,51X_3^*+5,40X_4^*-14,7))+$   
 $+0,0176,$   
 $R^2 = 0,87415$ ,  $S = 0,25753$ ,  $F = 9,5924$ .

Dabei bezeichnen  $X_j^*$ ,  $j \in 4$  die bereichsskalierten Deskriptorenwerte.  $f^*$  liefert den bereichsskalierten Eigenschaftswert. Um die Vorhersage für MIC zu erhalten, muss dieser noch rücktransformiert werden.

Die beiden bislang betrachteten Deskriptorensätze wurden so gewählt, dass sie für lineare Modelle bzw. Regressionsbäume beste Vorhersagefunktionen



Regressionsverfahren	Deskriptorensatz			
	<i>LM</i>	<i>RT</i>	<i>BCC</i>	<i>HCC</i>
MLR	0,64597	0,39654	0,42980	0,43927
RT	0,39654	0,81748	0,75118	0,65189
ANN, 1HN	0,49117	0,45177	0,43712	0,56780
ANN, 2HN	0,71535	0,45995	0,43712	0,56904
ANN, 3HN	0,71284	0,47453	0,43712	0,56903
SVM, lin	0,39005	0,37659	0,38318	0,32952
SVM, pol	0,37598	0,50894	0,48433	0,60133
SVM, rad	0,38718	0,55660	0,54426	0,59496

Tabelle 4.20:  $R^2$  für Modellierung von MIC durch verschiedene Deskriptoren und Regressionsverfahren

liefern. Wir testen zwei weitere Deskriptorensätze, die anhand der Korrelationskoeffizienten zu MIC ausgewählt wurden. Zum einen wurden die 5 Deskriptoren mit betragsmäßig größten Korrelationskoeffizienten verwendet:

$$\lambda_1^A (-0,402), R (0,369), FRB (0,346), {}^4\chi_c^v (-0,313), {}^4\chi_c (-0,310).$$

Die  $R^2$ -Werte für diesen Deskriptorensatz sind in Spalte *BCC* von Tabelle 4.20 eingetragen.

Die  $n$  Deskriptoren mit betragsmäßig größten Korrelationskoeffizienten zur Zielvariable bilden mitunter deutlich schlechtere Deskriptorensätze für MLR als die mit BSS bestimmte Menge von Deskriptoren. Ein Grund dafür kann eine starke Korrelation zwischen den Deskriptoren sein, wie wir in Abschnitt 4.4.2 gesehen haben. Um dieser Tatsache Rechnung zu tragen, ermitteln wir für eine Teilmenge  $\Omega$  von Deskriptoren folgenden Vergleichswert:

$$|\Omega|^{-1} \sum_{i \in \Omega} |R(X_i, Y)| - \left| \binom{\Omega}{2} \right|^{-1} \sum_{\{i,j\} \subset \Omega} |R(X_i, X_j)|,$$

wobei  $R(X, Z)$  den Korrelationskoeffizienten von  $X$  und  $Z$  bezeichne. Die 5-Teilmenge mit größtem Vergleichswert (0,18520) umfasst

$$\Phi, {}^4rings, {}^5rings, \lambda_1^A, {}^6\chi_c.$$

Für den nach dieser Heuristik gefundenen Deskriptorensatz stehen die  $R^2$ -Werte in Spalte *HCC*. Man sieht, dass die zu *HCC* ermittelten Modelle bis auf zwei Ausnahmen bessere  $R^2$ -Werte besitzen als zu *BCC* bestimmte Modelle. Insbesondere zur Berechnung von SVM mit polynomialen und radialen Kernel liefert *HCC* den besten der vier untersuchten Deskriptorensätze.

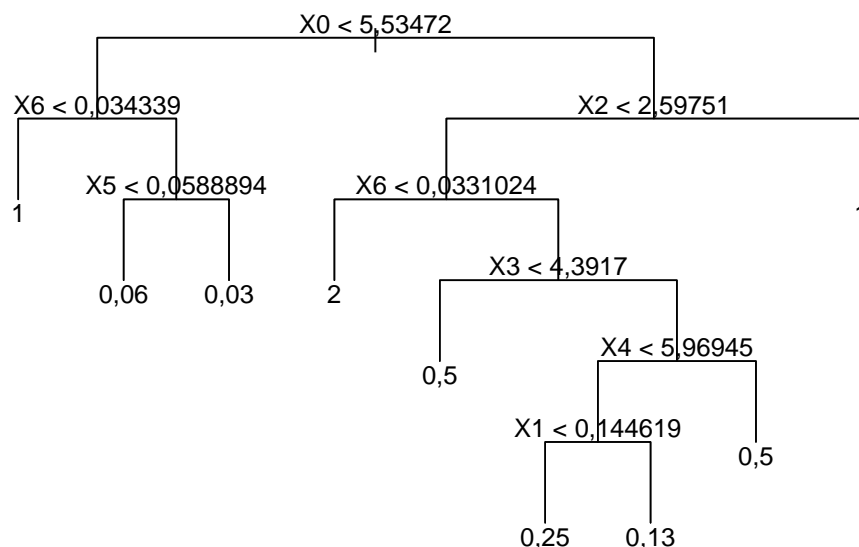


Abbildung 4.25: Mehrklassen-CT für MIC mit 7 Deskriptoren

Allerdings scheint Regression angesichts der niedrigen Werte für  $R^2$  in dem vorliegenden Beispiel eine wenig Erfolg versprechende Methode zur Vorhersage von ABA zu sein. Aufgrund der bereits erwähnten Verteilung der Eigenschaftswerte wollen wir Klassifikationsverfahren testen.

### Multi-Klassifikation

Wie schon oben erwähnt, nimmt MIC nur 7 verschiedene Werte an. Man kann die Fälle mit gleicher MIC als Klassen auffassen und die SAR-Suche als Klassifikationsproblem mit 7 Klassen formulieren. Abbildung 4.25 zeigt einen CT für MIC mit 8 inneren Knoten unter Verwendung von

$$\begin{aligned}
 n = 7 \text{ Deskriptoren: } X_0 &= {}^m M_2, X_1 = TIC_1, X_2 = CIC_1, X_3 = IC_2, \\
 X_4 &= MSD, X_5 = ch. J_5, X_6 = ch. J_6, \\
 MCE &= \frac{16}{51} = 0,31373.
 \end{aligned}$$

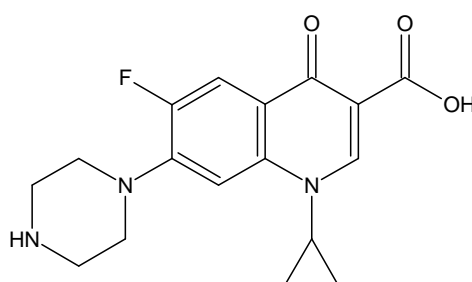
Tabelle 4.21 zeigt die Verteilung gemessener und berechneter MIC für diesen CT. Aufgrund der durch den CT berechneten Klassen kann man nun wiederum einen Wert für  $R^2$  berechnen. Dieser ist verständlicherweise mit 0,34811 deutlich schlechter als für die zuvor berechneten RT. Oft ist bei der Vorhersage pharmazeutisch-biologischer Eigenschaften nur interessant, ob eine Verbindung aktiv oder inaktiv ist. Wir wollen diesen Sachverhalt eingehend mit binärer Klassifikation untersuchen.

gemessene Klasse	berechnete Klasse						
	0,03	0,06	0,13	0,25	0,50	1,00	2,00
0,03	4	0	0	0	0	0	0
0,06	1	4	0	0	2	0	0
0,13	0	0	4	2	1	2	0
0,25	0	0	1	5	1	0	2
0,50	0	1	0	0	6	2	0
1,00	0	0	0	0	0	9	0
2,00	0	0	0	0	0	1	3

Tabelle 4.21: Verteilung gemessener und berechneter MIC für den CT aus Abbildung 4.25

### Binäre Klassifikation

Ein häufig verwendeter antimycobakterieller Wirkstoff ist Ciprofloxacin:



Ciprofloxacin ist ebenfalls in unserer realen Bibliothek enthalten und hat eine MIC von 0,06. [125] folgend werden wir alle Strukturen als *antimycobakteriell aktiv* betrachten, für die  $MIC \leq 0,06$  ist. Damit können wir unsere SAR-Suche als binäres Klassifikationsproblem auffassen. Wir werden dies mit verschiedenen Methoden zur Deskriptoren-Selektion und verschiedenen Klassifikationsverfahren lösen. Die Berechnung von Klassifikationsbäumen liefert einen CT (Abbildung 4.26) unter Verwendung von

$$n = 3 \text{ Deskriptoren: } X_0 = {}^m M_2, X_1 = {}^3 \chi^s, X_2 = ch. J_6, \\ MCE = \frac{2}{51} + \frac{0}{51} = \frac{2}{51} = 0,039216, MCE_{CV} = \frac{10}{51} = 0,19608.$$

MCE setzt sich zusammen aus zwei Arten von Fehlern:

- *Fehler erster Art*: Als falsch (inaktiv) berechnete wahre (aktive) Beobachtungen (Strukturen).
- *Fehler zweiter Art*: Als wahr (aktiv) berechnete falsche (inaktive) Beobachtungen (Strukturen).

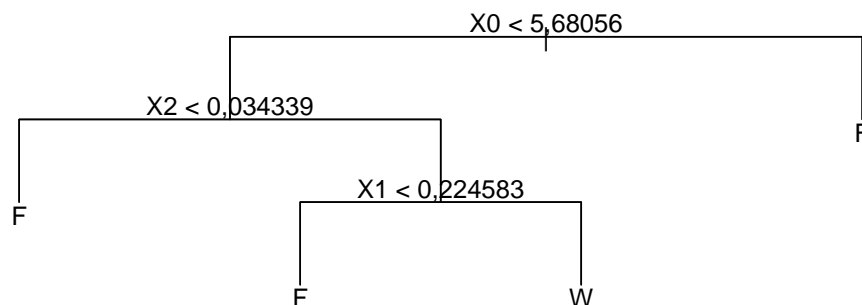


Abbildung 4.26: 2-Klassen-CT für ABA mit 3 Deskriptoren

Tabelle 4.22 zeigt Vierfeldertafeln für verschiedenen Deskriptorensätze und Klassifikationsverfahren. Die Tafel für Deskriptoren-Selektion durch CT und Klassifikation durch CT befindet sich oben links. Tabelle 4.23 gibt  $TCE$  und in Klammern den mit LOO-CV ermittelten  $TCE_{CV}$  für die verschiedenen Deskriptorensätze und Klassifikationsverfahren an.

Wir wollen ABA über Klassifikation durch Regression (vgl. Abschnitt 3.1.1) modellieren. Dazu erhalten Verbindungen mit  $MIC \leq 0,06$  den Eigenschaftswert 1, solche mit  $MIC > 0,06$  Eigenschaftswert  $-1$ . Liefert die Vorhersagefunktion Werte größer 0, so wird die entsprechende Struktur als aktiv, anderenfalls als inaktiv klassifiziert. Wir ermitteln beste lineare Modelle durch OLS-Regression mit  $n \in \underline{3}$  Deskriptoren, verwenden jedoch zur Bewertung nicht  $RSS$  oder  $R^2$ , sondern  $MCE$ . Wir erhalten Modelle mit

$n = 1$  Deskriptor:  $ch. J_1$ ,

$$\tilde{f} = 9,7362X_0 - 2,9152,$$

$$MCE = \frac{8}{51} + \frac{0}{51} = \frac{8}{51} = 0,15686, MCE_{CV} = \frac{11}{51} = 0,21569.$$

$n = 1$  Deskriptor:  ${}^4\chi_c^v$ ,

$$\tilde{f} = 5,1999X_0 - 0,72345,$$

$$MCE = \frac{6}{51} + \frac{2}{51} = \frac{8}{51} = 0,15686, MCE_{CV} = \frac{10}{51} = 0,19608.$$

$n = 2$  Deskriptoren:  ${}^3\kappa, {}^2\kappa_\alpha$ ,

$$\tilde{f} = 2,3886X_0 - 1,9619X_1 + 3,2971,$$

$$MCE = \frac{1}{51} + \frac{1}{51} = \frac{2}{51} = 0,039216, MCE_{CV} = \frac{2}{51} = 0,039216.$$

$n = 2$  Deskriptoren:  $\lambda_1^A, \chi_T$ ,

$$\tilde{f} = 65,570X_0 + 7623,5X_1 - 167,42,$$

$$MCE = \frac{2}{51} + \frac{0}{51} = \frac{2}{51} = 0,039216, MCE_{CV} = \frac{6}{51} = 0,11765.$$

Bemerkenswert ist die kleine Missklassifikationsrate für LOO-CV des erstgenannten 2-Deskriptoren-Modells. Vollständig separiert werden können die beiden Klassen durch folgende drei Modelle unter Verwendung von je

gemessene Klasse	berechnete Klasse										Klass.-verfahren		
	F	W	F	W	F	W	F	W	F	W		F	W
F	40	0	40	0	39	1	37	3	39	1	37	3	CT
W	2	9	3	8	2	9	2	9	6	5	1	10	
F	39	2	40	0	40	0	40	0	39	1	40	0	MLR
W	4	7	0	11	0	11	0	11	9	2	9	2	
F	38	2	38	2	40	0	38	2	38	2	39	1	LDA
W	4	7	0	11	0	11	0	11	6	5	6	5	
F	38	2	38	2	40	0	39	1	33	7	35	5	QDA
W	1	10	0	11	0	11	0	11	0	11	1	10	
F	40	0	40	0	39	1	37	3	38	2	40	0	KNN
W	5	6	5	6	3	8	1	10	4	7	0	11	
F	39	5	40	0	40	0	40	0	39	1	40	0	ANN 1HN
W	0	11	0	11	0	11	0	11	1	10	11	0	
F	39	1	40	0	40	0	40	0	39	1	34	6	ANN 2HN
W	0	11	0	11	0	11	0	11	1	10	0	11	
F	39	1	40	0	40	0	40	0	39	1	34	6	ANN 3HN
W	0	11	0	11	0	11	0	11	1	10	0	11	
F	39	1	39	1	40	0	40	0	37	3	37	3	SVM lin
W	4	7	0	11	0	11	0	11	4	7	4	7	
F	40	0	40	0	40	0	40	0	39	1	35	5	SVM pol
W	4	7	0	11	0	11	0	11	2	9	2	9	
F	40	0	40	0	40	0	40	0	39	1	35	5	SVM rad
W	2	9	1	10	0	11	0	11	5	6	2	9	
Deskriptoren	CT		LM <sub>0</sub>		LM <sub>1</sub>		LM <sub>2</sub>		FR		[125]		

Tabelle 4.22: Verteilung gemessener und berechneter ABA für verschiedene Klassifikationsverfahren und Deskriptorensätze

$$\begin{aligned}
n = 3 \text{ Deskriptoren: } & {}^1\chi^v, \text{ } CIC_2, \lambda_1^A, \\
& \tilde{f} = -0,86763X_0 + 0,95011X_1 + 55,068X_2 - 133,16, \\
& MCE = 0, MCE_{CV} = \frac{4}{51} = 0,078431.
\end{aligned}$$

$$\begin{aligned}
n = 3 \text{ Deskriptoren: } & {}^2\kappa_\alpha, {}^8P_{acyc}, {}^6P, \\
& \tilde{f} = -0,89054X_0 - 0,042156X_1 + 0,095820X_2 - 1,4373, \\
& MCE = 0, MCE_{CV} = \frac{4}{51} = 0,078431.
\end{aligned}$$

$$\begin{aligned}
n = 3 \text{ Deskriptoren: } & IC_2, MSD, \lambda_1^A, \\
& \tilde{f} = -0,87302X_0 - 1,0691X_1 + 44,867X_2 - 104,19, \\
& MCE = 0, MCE_{CV} = \frac{4}{51} = 0,078431.
\end{aligned}$$

Wir nennen letztere drei Modelle  $LM_0$ ,  $LM_1$  und  $LM_2$ . In Tabellen 4.22 und 4.23 bezeichnen die entsprechenden Spalten die jeweils verwendeten Deskriptorensätze. Die Ergebnisse für Klassifikation durch multiple lineare Regression findet man bei diesen Tabellen in den mit MLR betitelten Zeilen. Des

Klass.- verfahren	Deskriptorensatz					
	CT	LM <sub>0</sub>	LM <sub>1</sub>	LM <sub>2</sub>	FR	[125]
CT	2 (10)	3 (8)	3 (12)	5 (12)	7 (10)	4 (8)
MLR	6 (7)	0 (4)	0 (4)	0 (4)	10 (10)	9 (10)
LDA	6 (6)	2 (3)	0 (3)	2 (4)	8 (12)	7 (9)
QDA	3 (5)	2 (4)	0 (2)	1 (3)	7 (9)	6 (8)
KNN	5 (5)	5 (5)	4 (5)	4 (5)	6 (8)	0 (8)
ANN, 1HN	5 (6)	0 (2)	0 (1)	0 (1)	2 (2)	11 (7)
ANN, 2HN	1 (3)	0 (1)	0 (0)	0 (1)	2 (2)	6 (1)
ANN, 3HN	1 (4)	0 (0)	0 (0)	0 (0)	2 (2)	6 (3)
SVM, lin	5 (5)	1 (3)	0 (0)	0 (2)	7 (7)	7 (7)
SVM, pol	4 (5)	0 (1)	0 (0)	0 (1)	3 (3)	7 (8)
SVM, rad	2 (4)	1 (2)	0 (0)	0 (1)	6 (7)	7 (7)

Tabelle 4.23:  $TCE$  und  $TCE_{CV}$  für verschiedene Klassifikationsverfahren und Deskriptorensätze

Weiteren wurden die Deskriptoren mit größten Fisher-Quotienten

$${}^2\kappa_\alpha(1, 32722), {}^2\kappa(1, 2909), \Phi_\alpha(1, 1641)$$

zur Berechnung von Vorhersagefunktionen herangezogen (Spalte  $FR$  in Tabellen 4.22 und 4.23). Schließlich wollen wir noch die in [125] gewählten Deskriptoren

$$M_1, M_2, \xi^c$$

in unsere Untersuchungen einbeziehen.

Als Klassifikationsverfahren wurden neben CT und LDA auch *quadratische Diskriminanzanalyse* (kurz *QDA*), KNN, ANN mit einem, zwei und drei HN sowie SVM mit linearem, polynomialem (`degree = 2`) und radialem Kernel getestet. Die Deskriptorenwerte wurden für all diese Verfahren autoskaliert. Die Anzahl  $k$  betrachteter Nachbarn für KNN-Klassifikation wurde dabei über LOO-CV ermittelt. Tabellen 4.22 und 4.23 enthalten die Ergebnisse für  $k$  mit kleinstem  $TCE_{CV}$ . Für die verschiedenen Deskriptorensätze wurden folgende  $k$  ermittelt:

Deskriptorensatz	CT	LM <sub>0</sub>	LM <sub>1</sub>	LM <sub>2</sub>	FR	[125]
Anzahl $k$ von Nachbarn	15	7	5	5	9	1

Wegen  $k = 1$  für die Deskriptoren aus [125] erhält man trivialerweise eine vollständige Separierung der beiden Klassen. Dies ist beim Lesen von Tabelle 4.22 zu berücksichtigen und erlaubt keine Aussage über die Eignung dieses Deskriptorensatzes für KNN-Klassifikation.

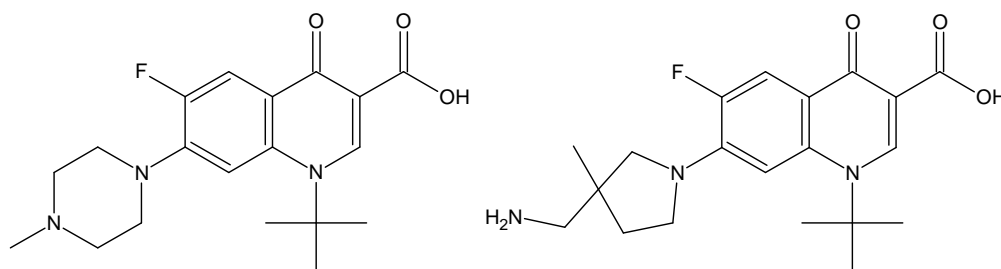
Bei ANN wurden je 10 per Zufall gewählte Startgewichtungen getestet und diejenige mit kleinstem  $TCE_{CV}$  ausgewählt. Zu den Deskriptoren aus [125] konnte kein ANN mit einem HN gefunden werden, welches die beiden Klassen separiert. Bessere Resultate konnten für diesen Fall mit bereichsskalierten Deskriptorenwerten erzielt werden ( $TCE = 3$ ,  $TCE_{CV} = 6$ ).

Auch beim Trainieren von Support-Vektor-Klassifikatoren können mehrere Parameter variiert werden. Dies wirkt sich mitunter deutlich auf die Anpassung und Vorhersagefähigkeit der SVM aus. Wir haben diesbezüglich unter der  $R$ -Umgebung die Parameter  $\text{cost} = 2^i$ ,  $i \in \{-1, \dots, 7\}$  und  $\text{gamma} = 2^j$ ,  $j \in \{-3, \dots, 3\}$  getestet und diejenige Parameter-Kombination mit kleinstem  $TCE_{CV}$  ausgewählt.

Ein Blick auf Tabelle 4.23 zeigt, dass die Deskriptorensätze für beste LM auch gute ANN und SVM liefern. Die Deskriptoren mit besten Fisher-Quotienten sind eher schlecht geeignet zur Berechnung von Klassifikatoren. Dies liegt wohl nicht zuletzt an den hohen Korrelationen zwischen den Deskriptoren:  $R(^2\kappa_\alpha, ^2\kappa) = 0,99140$ ,  $R(^2\kappa_\alpha, \Phi_{\bar{\alpha}}) = 0,97880$  und  $R(^2\kappa, \Phi_{\bar{\alpha}}) = 0,98505$ .

### Vorhersage

Wir wollen die linearen Modelle  $LM_0$ ,  $LM_1$  und  $LM_2$  zur Vorhersage von ABA für eine virtuelle Bibliothek von Quinolonen heranziehen. Wir lassen für  $R^1$  die 6 Substituenten aus der oberen Reihen von Abbildung 4.23 und für  $R^2$  die verbleibenden 15 Substituenten zu. Die resultierende virtuelle Bibliothek umfasst  $6 \cdot 15 = 90$  Strukturen. Abzüglich der realen Bibliothek verbleiben 39 zuvor nicht vermessene Verbindungen in der rein virtuellen Bibliothek. Die Vorhersagen unserer drei Klassifikatoren sind bemerkenswert konsistent. Von  $LM_0$  und  $LM_2$  werden übereinstimmend folgende beiden Strukturen als antimycobakteriell aktiv prognostiziert:



$LM_1$  klassifiziert in der rein virtuellen Bibliothek nur eine Struktur als aktiv. Dies ist gerade die linke der beiden oben gezeigten Verbindungen. Sie erscheint damit als aussichtsreicher Kandidat für einen neuen Wirkstoff.

## 4.5 Bemerkungen zur realen Bibliothek

Bislang haben wir noch nicht über die Auswahl der realen Bibliothek gesprochen. Bei den hier gezeigten Beispielen wurden Strukturen und Eigenschaften existierenden Datenbanken oder der Literatur entnommen. Im Idealfall sollte aber bereits die Auswahl der realen Bibliothek das Fundament für die erfolgreiche Optimierung eines kombinatorisch-chemischen Experimentes bilden. Die reale Bibliothek wird dann so gewählt werden, dass in ihr möglichst diverse Strukturen enthalten sind.

### 4.5.1 Diversität als Auswahlkriterium für die reale Bibliothek

Mathematische Werkzeuge zur Suche nach Teilbibliotheken hoher *Diversität* bilden Methoden des *unüberwachten* statistischen Lernens. Unüberwachtes statistisches Lernen wird im Gegensatz zum überwachten Lernen *ohne* die Verwendung einer abhängigen Variablen durchgeführt. Das Ziel des unüberwachten Lernens besteht darin, die Beobachtungen gemäß Ähnlichkeiten in den Werten der unabhängigen Variablen zu strukturieren und zu klassifizieren.

Für unsere Anwendungen in der kombinatorischen Chemie können wir als unabhängige Variablen molekulare Deskriptoren heranziehen, und diese entweder zunächst nur auf die Bausteine, oder aber auch direkt auf die virtuelle kombinatorische Bibliothek anwenden. Wichtige Verfahren zur Durchführung unüberwachten Lernens bilden *Hauptkomponentenanalyse* (engl. *Principal Component Analysis*, kurz *PCA*) und *Clusteranalyse*.

Beispielsweise werden in [164] topologische Indizes und PCA herangezogen, um die 20 Aminosäuren in Gruppen ähnlicher Strukturen einzuteilen. J. Biebl verwendet in [17] Fuzzy-Clusteralgorithmen zur Konformationsanalyse.

Das in Abschnitt 4.3.2 beschriebene Verfahren zur Generierung aller Substrukturen sowie deren Vielfachheiten für eine gegebene Bibliothek chemischer Verbindungen scheint besonders geeignet zur Berechnung unabhängiger Variablen, da diese auf sehr kanonische Weise und nicht durch die subjektive Auswahl des Benutzers bestimmt werden. In dem Beispiel 4.3.8 haben wir alle 20 Substrukturen mit 2 bis 6 Kanten sowie deren Vielfachheiten für eine reale Bibliothek von Decanen berechnet. Die Anwendung des Algorithmus auf die virtuelle Bibliothek aller 75 Decane liefert keine weiteren Substrukturen. Wir verwenden die autoskalierten Substruktur-Vielfachheiten für eine hierarchische Clusteranalyse. Abbildung 4.27 zeigt das Ergebnis der Clusteranalyse als *Dendrogramm*. Die Blätter sind gemäß der Nummerierung der Strukturen in Abbildungen 4.6 (R01–R50) und 4.15 (V01–V25) beschriftet.



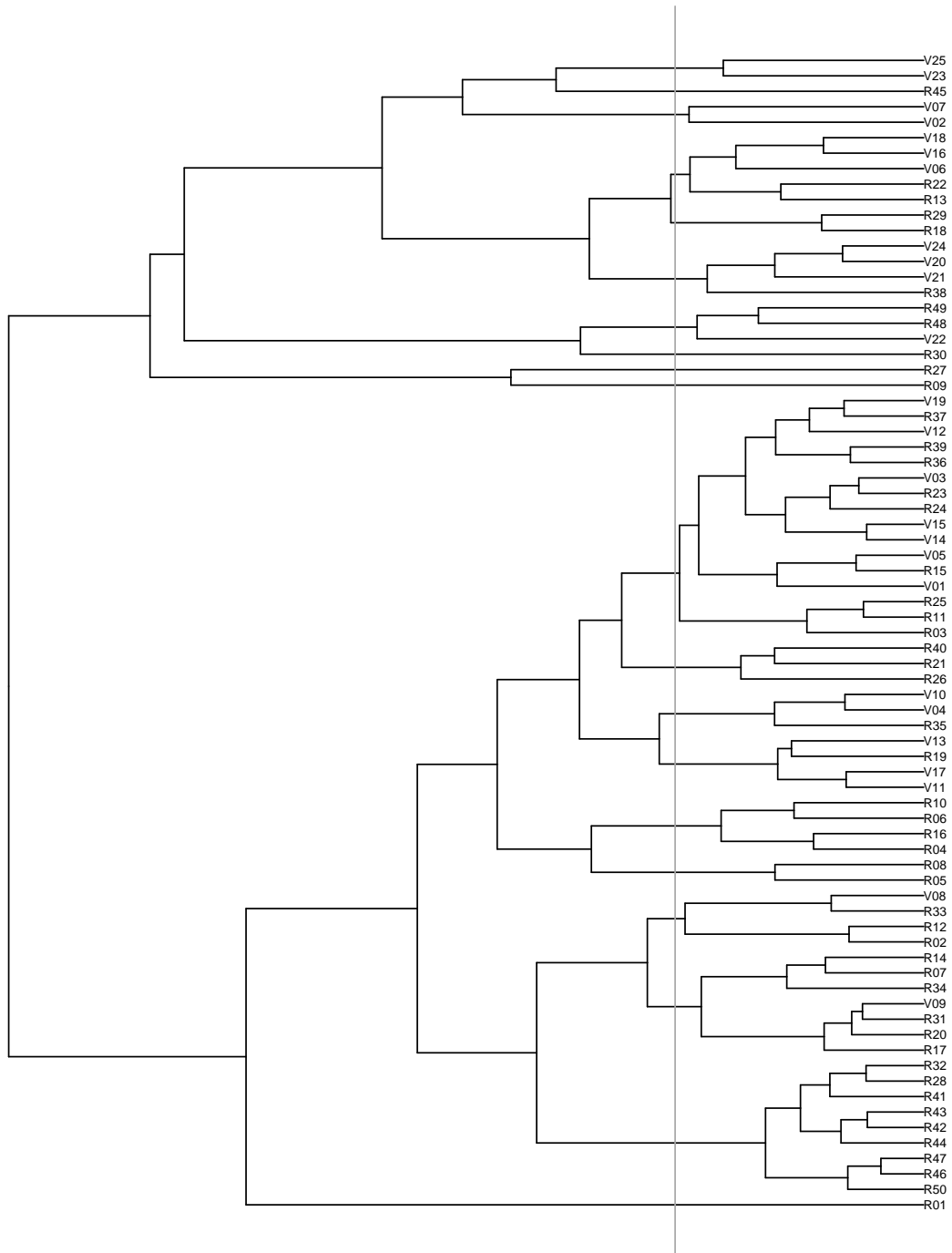


Abbildung 4.27: Dendrogramm der virtuellen Bibliothek von Decanen

Bei der hierarchischen Clusteranalyse werden nahe beieinander liegende Beobachtungen sukzessive zusammengefasst. Die Entfernung zweier Beobachtungen wird dabei durch eine Metrik auf dem Raum der Vorhersagevariablen  $X_i$ ,  $i \in n$  bestimmt. In unserem Beispiel wurde die euklidische Metrik auf  $\mathbb{R}^{20}$  verwendet. Das Dendrogramm visualisiert die Reihenfolge, nach der die Verbindungen zusammengefasst wurden. Wollte man nun aufgrund der Clusteranalyse eine reale Bibliothek mit  $m$  Verbindungen für Synthese und Screening auswählen, könnte man anhand des Dendrogramms zunächst eine Klasseneinteilung in  $m$  Cluster vornehmen, und je eine Struktur aus jedem Cluster für die reale Bibliothek selektieren. In Abbildung 4.27 ist diese Vorgehensweise durch die graue vertikale Linie skizziert. Die Teilbäume rechts der Linie repräsentieren 20 Cluster. Entnimmt man jedem Cluster eine Verbindung, erhält man eine Teilbibliothek hoher Diversität.

#### 4.5.2 Überprüfung der Teilmengenrelation von realer und virtueller Bibliothek

Im Normalfall — das hat sich in mehreren Kooperationsprojekten mit Industriepartnern immer wieder bestätigt — bilden zunächst vorhandene Datenbestände den Ausgangspunkt für kombinatorisch-chemische Experimente. In dieser Situation ist es sinnvoll, zu überprüfen, ob die bestehende reale Bibliothek auch tatsächlich Teilmenge einer mit kombinatorischen Methoden generierten virtuellen Bibliothek ist.

Dazu werden zuerst in beiden Bibliotheken aromatische Bindungen identifiziert. Dann werden die molekularen Graphen auf kanonische Form gebracht. Eine mögliche Vorgehensweise ist nun, die beiden Bibliotheken als Sequenzen molekularer Graphen aufzufassen. Bei der Implementierung wird man die Strukturen dabei in möglichst komprimierter Form darstellen. Anhang A zeigt die in der aktuellen *MOLGEN*-Version verwendete lineare Darstellung molekularer Strukturen. Die einzelnen Strukturen können dann als Zeichenfolgen aufgefasst und lexikographisch sortiert werden. Die sortierten Sequenzen kann man mit linearem Aufwand auf Dubletten und Teilmengenrelationen überprüfen. Ebenso lassen sich Schnitt- und Differenzmengen zwischen den Bibliotheken bilden.

# Kapitel 5

## Molekulare Strukturaufklärung

Ein Großteil der chemischen Laborarbeiten, sowohl in der Forschung als auch der Industrie beschäftigt sich mit Fragen der Analytik. Wichtige Teilprobleme sind dabei die Auftrennung von *Stoffgemischen* in chemische *Reinsubstanzen* und die Ermittlung der Struktur dieser Reinsubstanzen. Obwohl Konfiguration und Konformation einer chemischen Verbindung größere Information liefern, ist man oft nur auf die Bestimmung der Konstitution konzentriert. Wir fassen diese Bestrebungen unter dem Begriff der *molekularen Strukturaufklärung* zusammen.

Es gibt viele Situationen, in denen es von enormem Nutzen ist, die Strukturformel(n) einer chemischen Substanz zu kennen. Als Beispiele seien hier die Qualitätskontrolle einer chemischen Synthese oder Suche nach Giftstoffen in der Umweltanalytik genannt. Oft können anhand der chemischen Struktur Modelle zur Vorhersage physiko-chemischer oder biologisch-pharmazeutischer Eigenschaften aufgestellt werden, wie wir in Kapitel 4 sahen. Und nicht zuletzt ist die Kenntnis der Strukturformel einer chemischen Verbindung schon deshalb wichtig, um sie datentechnisch identifizieren und verarbeiten zu können.

Die Eingabe für ein Strukturaufklärungsproblem stammt zum einen aus dem Vorwissen des Chemikers über die unbekannte Substanz und zum anderen aus experimentellen Messungen, denen der Analyt unterzogen wurde. Vorkenntnisse des Chemikers können etwa aus dem Syntheseweg oder allgemeiner aus der Herkunft einer Substanz resultieren. Auf Seite der experimentellen Messungen liefern Methoden der *Spektroskopie* den größten Informationsgehalt.

## 5.1 Spektroskopische Methoden

Die chemische Analytik hat eine Vielzahl derartiger Methoden hervorgebracht. Die wichtigsten darunter sind:

- *Kernresonanzspektroskopie* (engl. *Nuclear Magnetic Resonance Spectroscopy*, kurz *NMR*),
- *Infrarot- und Ultraviolett-spektroskopie* (*IR, UV*),
- *Atomabsorptions- und Atomemissions-spektroskopie* (*AAS, AES*),
- *Massenspektrometrie* (*MS*).

Prinzipiell beruhen diese Methoden darauf, dass die Moleküle des Analyten Energiezustände wechseln. Mit Ausnahme der Massenspektrometrie wird detektiert, ob und in welchem Maße Energie aufgenommen oder abgegeben wird.

Bei der NMR-Spektroskopie werden bestimmte Atome in einem starken homogenen Magnetfeld angeregt. Dabei beobachtet man Übergänge zwischen Energiezuständen von bestimmten Atomkernen, die in Form chemischer Verschiebungen und Spin-Spin-Kopplungen gemessen werden. Diese lassen sehr genaue Rückschlüsse auf die Umgebungen der einzelnen Atome im Molekül zu. IR- und UV-Spektroskopie messen, wie stark der Analyt Licht verschiedener Wellenlängen absorbiert. Bestimmte Substrukturen können gemäß ihres charakteristischen Absorptionsverhaltens auf diese Weise erkannt werden. AAS und AES ermöglichen den Nachweis einzelner Elemente. Einen Überblick über die Funktionsweise der verschiedenen spektroskopischen Methoden findet man u.a. in [18, 32, 66]. In [109] sind einfache Beispiele von Strukturaufklärungsproblemen mit Anleitungen zur manuellen bzw. *interaktiven* computergestützten Bearbeitung zusammengestellt.

Für uns ist in erster Linie von Interesse, welche Art struktureller Information spektroskopische Methoden liefern, wie diese aus den experimentellen Daten gewonnen und zur molekularen Strukturaufklärung eingesetzt werden können.

## 5.2 Prinzip der automatisierten molekularen Strukturaufklärung

Bestrebungen, den Strukturaufklärungsprozess zu automatisieren, werden bereits seit über drei Jahrzehnten vorangetrieben. Beschleunigt wurden Entwicklungen auf dem Gebiet der automatisierten Strukturaufklärung durch die Verfügbarkeit immer leistungsfähigerer Rechner und dem einhergehenden Fortschritt in verschiedenen Bereichen der elektronischen Datenverarbeitung. Insbesondere begann man in verstärktem Maße chemische Strukturen und ihre Eigenschaften digital zu kodieren und in Datenbanken zu sammeln. Man kann zwei Vorgehensweisen zur automatisierten Strukturaufklärung unterscheiden:

- Datenbank basierte Strukturaufklärung,
- De Novo-Strukturaufklärung.

Es gibt mittlerweile für die verschiedenen spektroskopischen Methoden große Datenbanken, in denen Paare von Spektren und Strukturen abgelegt sind ([159], Kapitel 9). In solchen Spektren-Datenbanken kann man zu einem experimentell ermittelten Spektrum einer unbekanntem chemischen Verbindung nach ähnlichen Spektren suchen, Datenbank-Spektren hinsichtlich ihrer Übereinstimmung mit dem experimentellen Spektrum abfallend sortieren, und die zugehörigen Strukturen gemäß dieser Sortierung als mögliche Kandidaten für die Unbekannte ausgegeben lassen. Die Algorithmen zum *Spektrenvergleich* sind inzwischen soweit entwickelt, dass die korrekte Struktur mit hoher Wahrscheinlichkeit die erste Position in einer solchen *Hitliste* einnimmt, vorausgesetzt sie ist in der Datenbank vorhanden.

Hiermit haben wir bereits das größte Problem der Strukturaufklärung mit Hilfe von Spektren-Datenbanken angesprochen. Selbst die größten Datenbanken umfassen derzeit nur mehrere Zehntausend bis wenige Hunderttausend Spektren. So sind in der im Rahmen dieser Arbeit verwendeten MS-Datenbank *NIST '98* [137] 107888 Spektren zu 107812 Strukturen verzeichnet. Die in Abschnitt 1.7 verwendete Datenbank *Beilstein* umfasst derzeit 8711107 als existent nachgewiesene Strukturen, die Anzahl mathematisch möglicher Konstitutionsisomere kann schon bei Summenformeln zu kleinen Molekülmassen (bis 150 amu) über 100 Millionen liegen (vgl. Anhang E). So ist beispielsweise die hinsichtlich ihres Molekulargewichts kleinste Summenformel, für die mehr als 100 Millionen Konstitutionsisomere existieren,  $C_8H_6N_2O$  (146 amu, 109.240.025 Isomere).

Gerade im Zusammenhang mit kombinatorisch-chemischen Methoden muss man davon ausgehen, dass Syntheseprodukte im Normalfall nicht in Spek-

tren-Datenbanken vorliegen. Auf der Suche nach neuen Wirkstoffen und Materialien ist man aber gerade an diesen Strukturen interessiert. Wegen der hohen Durchsatzraten neuer Synthese- und Screeningmethoden ist der Bedarf an automatisierten Verfahren zur Strukturerkennung größer denn je. So wird in [70] das Verfahren des *Hochauflösungs-Screenings* (engl. *High Resolution Screening*, kurz *HRS*) beschrieben, bei dem in einem Arbeitsgang Stoffgemische in Reinsubstanzen zerlegt, die Reinsubstanzen auf ihre biologische Aktivität getestet und im Falle aktiver Substanzen deren Massenspektren aufgenommen werden. Auf diese Weise können in einem Arbeitsgang mehrere 10000 Verbindungen innerhalb weniger Stunden analysiert werden. Die manuelle Auswertung der spektroskopischen Daten kann für jede aktive Verbindung mehrere Tage oder gar Monate in Anspruch nehmen. Automatisierte Verfahren zur Strukturbestimmung sind dringend erforderlich, um diesen Engpass zu überwinden.

De Novo Strukturaufklärung verfolgt das Ziel der molekularen Strukturbestimmung *ohne* Datenbanksuche. Eines der ersten *Expertensysteme* für diese Problemstellung wurde im Rahmen des *DENDRAL*-Projekts [87] realisiert. Dabei wurden ausschließlich MS verwendet. Seither wurde eine Vielzahl solcher Expertensysteme basierend auf der Kombination verschiedener spektroskopischer Methoden entwickelt (*RASTR* [37], *X-PERT* [35, 36], *StrE-luc* [34], *SESAMI* [27], *CHEMICS* [43, 44], *SpecSolv* [166], *EXPEC* [88, 89], u.s.w.). Im Wesentlichen arbeiten diese Systeme unabhängig von den verwendeten spektroskopischen Methoden nach dem gleichen Grundprinzip, welches in Abbildung 5.1 skizziert ist. Dieses Prinzip kann grob in drei Teilprobleme gegliedert werden, die einer mathematischen Modellierung bedürfen:

- Mit *Spektreninterpretation* bezeichnen wir die Extraktion struktureller Eigenschaften des Analyten aus den spektroskopischen Daten. Hierbei kommen Methoden der *Mustererkennung* und des *überwachten statistischen Lernens* zum Einsatz (Abschnitt 5.5).
- Im zweiten Schritt werden durch *molekulare Strukturgenerierung* alle Strukturformeln konstruiert, die den extrahierten strukturellen Eigenschaften genügen (Abschnitt 2.1).
- Mittels *Spektrensimulation* werden ausgehend von den Struktur-Kandidaten virtuelle Spektren berechnet. Diese werden mit den experimentellen Daten verglichen. Aufgrund der Vergleichswerte können gute Strukturkandidaten angeordnet und ausgewählt werden (Abschnitt 5.4). Wir fassen Simulation, *Vergleich*, *Ranking* und *Selektion* unter dem Begriff *Struktur-Verifikation* zusammen.

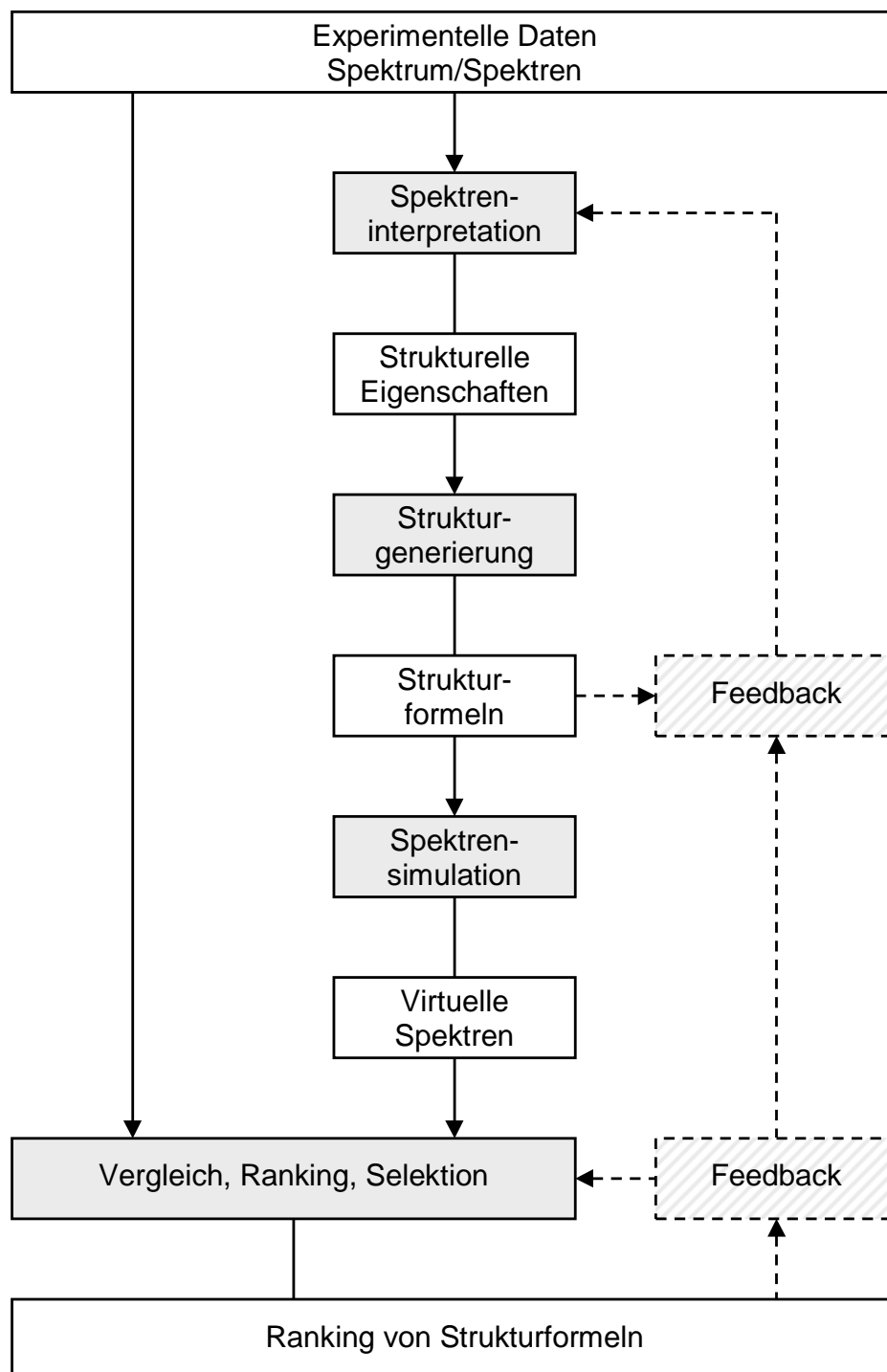


Abbildung 5.1: Vorgehensweise bei der automatisierten Strukturaufklärung

Im günstigsten Fall ist diejenige Struktur, deren virtuelles Spektrum die beste Übereinstimmung zu den gemessenen Daten liefert, die Lösung des Strukturaufklärungsproblems. In der Regel erhält man aber entweder nach der Strukturgenerierung oder nach der Verifikation sehr große oder leere Strukturräume. Dann müssen Parameter der Interpretation oder der Selektion modifiziert werden. Zudem muss es möglich sein, Vorkenntnisse des Chemikers über den Analyten in den Strukturaufklärungsprozess einfließen zu lassen. Solche Vorkenntnisse können beispielsweise aus der Kenntnis des Synthesewegs resultieren.

Während die Strukturgenerierung streng nach den mathematischen Vorgaben durchgeführt werden sollte, benötigt man zur Interpretation und Verifikation eine chemische Wissensbasis. Im Extremfall ist diese wiederum eine Spektren-Datenbank. So sucht *SpecInfo* [86, 102] zunächst nach ähnlichen Spektren in einer Datenbank und extrahiert aus den zugehörigen Strukturen gemeinsame Substrukturen als Eingabe für die Strukturgenerierung. Andere Verfahren [154, 160, 163] trainieren Spektren-Klassifikatoren mit Hilfe von Datenbank-Spektren. Für den eigentlichen Strukturaufklärungsprozess sind solche Klassifikatoren jedoch unabhängig von Datenbank-Spektren. Good- und Badlist Einträge für die Strukturgenerierung werden direkt aus den Prognosen der Vorhersagefunktionen gewonnen [151, 155].

In [135] und [40] werden an- und abwesende Substrukturen aus IR-Spektren ermittelt, indem man versucht, die Banden des experimentellen Spektrums möglichst gut durch Substrukturen zu erklären. Dabei werden solche Substrukturen verwendet, für die bekannt ist, dass sie ein charakteristisches Absorptionsverhalten in IR-Spektren zeigen. Derartige Erfahrungswerte sind in Form von Intervallen angegeben [66, 108] und wurden ursprünglich aus Spektren-Datenbanken gewonnen.

Auch für die Struktur-Verifikation ist eine Datenbasis aufgeklärter Spektren unentbehrlich. So werden statistische Modelle zur Vorhersage chemischer Verschiebungen in NMR-Spektren [75, 92, 93, 94] mittels vorhandener Datenbestände trainiert. In Abschnitt 5.4 werden wir sehen, wie man Datenbank-Spektren dazu verwendet, um die Selektion von Strukturkandidaten anhand eines MS durchzuführen.



## 5.3 Grundlagen der Massenspektrometrie

Bereits der Name lässt erahnen, dass sich die Massenspektrometrie von den übrigen Spektroskopie-Methoden unterscheidet. Im Gegensatz zur NMR- und IR-Spektroskopie werden bei der Massenspektrometrie nicht die Moleküle des Analyten selbst vermessen, sondern geladene Teilchen, die durch Reaktionen aus dem Analyten hervorgehen. Die Anzahl und Struktur der verschiedenen geladenen Teilchen ist abhängig von Ionisierungsart und -energie. Neu erwecktes Interesse an der Massenspektrometrie begründet sich aus

- der Möglichkeit, sie in einen weitgehend automatisierten Prozess von Synthese und Screening einzubinden.
- Dies wird durch die hohe Sensitivität von MS ermöglicht, d.h. Massenspektren können
  - mit sehr geringen Stoffmengen und
  - in sehr kurzer Zeit aufgenommen werden.
- Massenspektrometrie liefert „vollständige“ Strukturinformation.

Allerdings ist diese Strukturinformation nur schwer zugänglich. Wir nehmen diese Argumente zum Anlass, uns eingehend mit dieser Methode zu beschäftigen. Wir sind im Folgenden besonders an der *niedrig auflösenden* (engl. *Low Resolution*, kurz *LR*) 70eV *Elektronenstoß-* (engl. *Electron Impact*, kurz *EI*) Massenspektrometrie interessiert. Zu dieser Methode existieren die umfangreichsten Datenbanken mit Referenzspektren, und man kann davon ausgehen, dass zu gleichen Substanzen auf verschiedenen Spektrometern annähernd gleiche Spektren aufgenommen werden. Abbildung 5.2 zeigt ein Beispiel eines solchen Massenspektrums. Dabei ist nach rechts das Verhältnis von Masse und Ladung, nach oben die Intensität abgetragen. In der Chemie ist die Intensität typischerweise mit Werten zwischen 0 und 100 angegeben. Für unsere mathematischen Betrachtungen werden wir der einfacheren Schreibweise wegen im Folgenden einer Skalierung zwischen 0 und 1 den Vorzug geben.

### 5.3.1 Funktionsweise des EI-Massenspektrometers

Abbildung 5.3 zeigt schematisch die Funktionsweise eines EI-Massenspektrometers. Die Moleküle (a) der aufzuklärenden Substanz werden in einer Ionisationskammer mit Elektronen (b) beschossen. Dabei verlieren die Teilchen des Analyten ein Elektron aus der äußeren Schale. Dieser Vorgang wird als *Ionisation* bezeichnet und es entsteht das positiv geladene *Molekülion* (c).

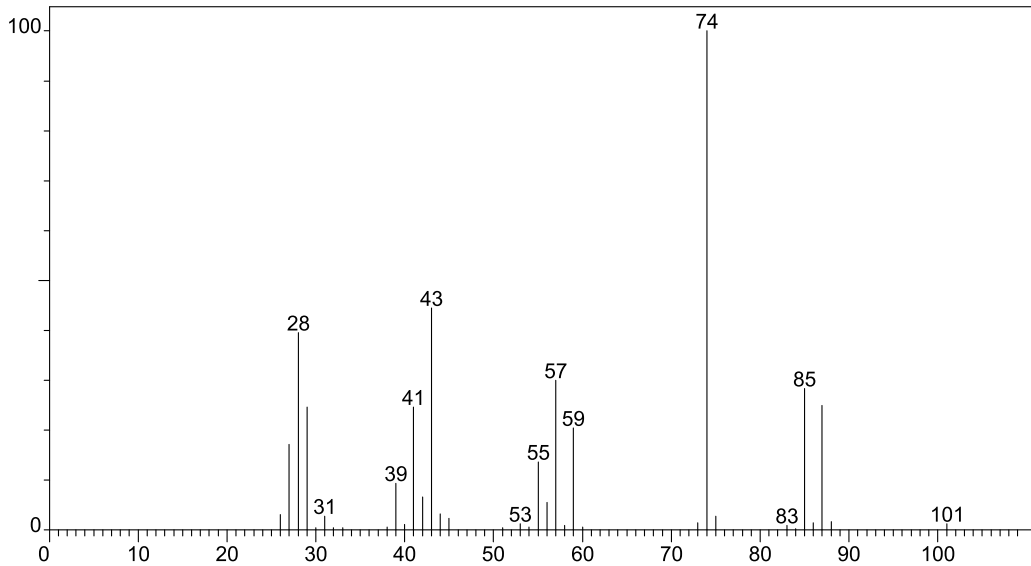


Abbildung 5.2: Beispiel eines Elektronenstoß-Massenspektrums

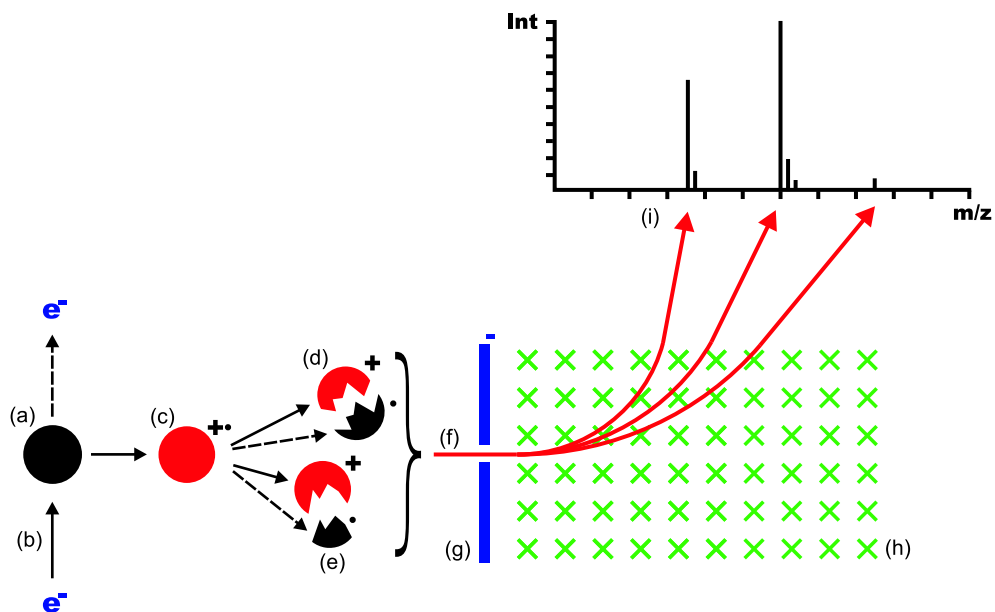


Abbildung 5.3: Funktionsweise eines EI-Massenspektrometers

Dieses wiederum zerfällt weiter in positiv geladene Ionen (d) und *Neutralteilchen* (e). Die *Fragmente*, die direkt aus dem Molekülion hervorgehen, nennen wir *primäre (Fragment-)Ionen* bzw. *primäre Neutralteilchen*. Je nach Stabilität zerfallen die primären Ionen weiter in *sekundäre (Fragment-)Ionen* und *sekundäre Neutralteilchen*. Dieser Vorgang, genannt *Fragmentierung*, kann bei Ionen beliebiger Ordnung stattfinden. Zudem können *Umlagerungsreaktionen* eine noch stärkere Diversifizierung des Teilchengemischs in der Ionisationskammer hervorgerufen.

Während die Neutralteilchen aus diesem Teilchengemisch durch eine Vakuumpumpe abgesogen werden, wirkt auf die Ionen (f) eine Spannung (g), die sie in Richtung eines Magnetfelds (h) beschleunigt. Dort unterliegt der Ionenstrahl der *Lorenzkraft*. Teilchen mit kleinerem *Masse-Ladungs-Verhältnis* erfahren eine stärkere Ablenkung als solche mit größerem. Dementsprechend werden die Teilchen an verschiedenen Orten von einem Detektor registriert und auf Peaks im Massenspektrum abgebildet. Die Lage der Peaks gibt das Masse-Ladungs-Verhältnis der Ionen an. Die Höhe der Peaks steht für die (relative) Anzahl der Teilchen, die zu dem jeweiligen  $m/z$ -Wert registriert wurden, und wird als *Intensität* bezeichnet.

### 5.3.2 Problemstellungen der EI-Massenspektrometrie

Die automatische Strukturaufklärung anhand eines EI-Massenspektrums findet auf mehreren Ebenen statt. Dabei sind

- die Molekülmasse,
- die Bruttoformel und
- die Strukturformel

des Analyten zu bestimmen. Dass ausgerechnet die Bestimmung der Molekülmasse im EI-Massenspektrometer Probleme bereitet, bedarf einer kurzen Erklärung: Leider ist bei manchen Substanzen das Molekülion derart instabil, dass es sofort weiter in Fragmente zerfällt. Das Molekülion selbst ist dann nicht in genügend hoher Konzentration vorhanden, um im Massenspektrum einen Peak bei der Molekülmasse zu hinterlassen. Dies betrifft etwa 15% aller Substanzen, die in MS-Datenbanken erfasst sind. Zudem gibt es kein deterministisches Kriterium um für ein EI-Massenspektrum festzustellen, ob die größte auftretende Masse tatsächlich der Molekülmasse entspricht. Dennoch lassen bestimmte Anhaltspunkte Rückschlüsse vom Spektrum auf die Molekülmasse zu. Bereits in [87] wurde ein heuristisches Verfahren zur Bestimmung von Kandidaten für die Molekülmasse beschrieben. Statistische

Ergebnisse liefert [100] für eine Methode zur Vorhersage der Molekülmasse. Dabei konnte für unbekannte Spektren in 91% der Fälle die Molekülmasse korrekt vorhergesagt werden, in 95% der Fälle war der korrekte Kandidat für die Molekülmasse unter den ersten beiden Vorschlägen. Gewissermaßen komplementäre Information zur Molekülmasse können Verfahren wie [101] liefern, die die Parität der Molekülmasse vorhersagen. Neuartige Massenspektrometer ermöglichen eine hardwareseitige Bestimmung der Molekülmasse durch *weiche Ionisierung* (engl. *Soft Ionisation*, kurz *SI*). So wird das Molekülion in hinreichend hoher Konzentration erhalten. Dafür liefern SI-MS in der Regel jedoch deutlich weniger Fragmente als EI-MS.

Abbildung 5.4 zeigt schematisch unsere Vorgehensweise bei der Strukturaufklärung mittels MS. Schwerpunkt bilden in dieser Arbeit die Bestimmung von Brutto- und Strukturformel. Für die Interpretation verwenden wir MS-Klassifikatoren. Diese können sowohl Informationen zur elementaren Zusammensetzung als auch zur Struktur des Analyten liefern (vgl. Anhang C). Dazu ziehen wir Klassifikatoren nach K. Varmuza und W. Werther aus früheren Arbeiten [154, 160, 163] heran, entwickeln aber auch neue Klassifikatoren unter Verwendung zuvor nicht betrachteter Klassifikationsmethoden (Abschnitt 5.5.2) und neuer struktureller Eigenschaften (Abschnitt 5.5.3).

Zu gegebener Molekülmasse kann man alle Bruttoformeln berechnen, die dieser Masse entsprechen. Die Generierung aller Strukturformeln zu einer vorgeschriebenen Bruttoformel wurde in Abschnitt 2.1 behandelt.

Auf den ersten Blick liefert ein Massenspektrum Informationen über die Massen und Intensitäten im Ionengemisch des Analyten. Ein wichtiges Bindeglied zwischen diesen primären Informationen stellen Isotopenverteilungen und theoretische Isotopenmuster dar. Wir werden diesen Zusammenhang in Abschnitt 5.3.4 erläutern. Ideen aus [55, 83, 132] werden weiterentwickelt, um aus den vorhandenen Informationen über Masse, Intensitäten und Isotopenmuster *Vergleichswerte* für Bruttoformel-Kandidaten zu berechnen. Diese werden zum *Ranking* und zur *Selektion* von Bruttoformel-Kandidaten herangezogen (Abschnitt 5.4.1).

Sind Kandidaten für die Strukturformel gegeben, kann man das Verfahren zur Berechnung von Vergleichswerten verfeinern, indem man die Kenntnis über MS-Fragmentierungsreaktionen zur Generierung der möglichen Fragmentionen heranzieht. Die so gewonnenen *Strukturformel-Vergleichswerte* werden zum *Ranking* und zur *Selektion* von Strukturkandidaten verwendet (Abschnitt 5.4.2).

Schließlich werden die einzelnen Schritte gekoppelt, um anhand zweier Beispiele die automatisierte Strukturaufklärung mit MS zu demonstrieren (Abschnitt 5.6). Aufgrund der festgestellten Missklassifikationsraten der MS-Klassifikatoren, der Größe der Strukturräume und der Unzulänglichkeiten

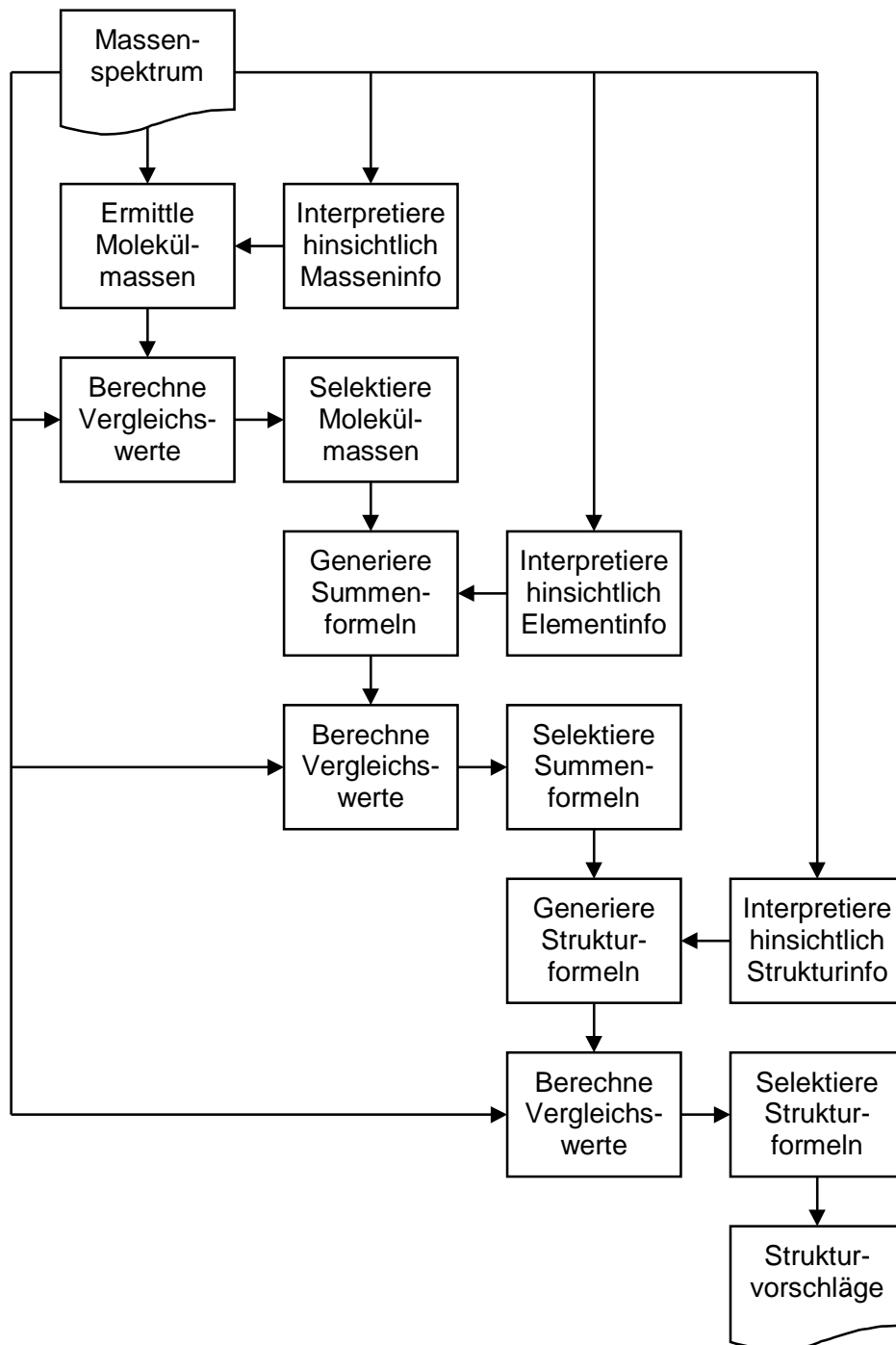


Abbildung 5.4: Vorgehensweise bei der Strukturaufklärung mittels MS

bei der Kandidaten-Selektion kann ein Expertensystem, welches ausschließlich auf niedrig aufgelösten EI-MS beruht, nach derzeitigem Stand der Forschung noch nicht hinreichend zuverlässig arbeiten, um es in der Praxis im automatischen Betrieb einzusetzen.

Allerdings können entscheidende Fehlerquellen durch gezielten Hardwareeinsatz beseitigt werden. Die Bestimmung der Bruttoformel kann hardwareseitig durch eine höhere Massenauflösung verbessert werden (Abschnitt 5.8). Stehen AAS oder AES zur Verfügung, so kann mit deren Hilfe die empirischen Formel bestimmt werden. Dann ist die Ermittlung der Summenformel meist eindeutig möglich, wie Studien unter Verwendung von Ergebnissen aus [139] zeigten.

### 5.3.3 Massenspektrum und Peakcluster

#### 5.3.1 Definition:

Ein niedrig aufgelöstes *Massenspektrum* ist eine Abbildung

$$I : \mathbb{N}^* \longrightarrow \mathbb{R}_0^+, \quad m \longmapsto I(m)$$

für die gilt:

$$(i) \quad \exists \hat{m} : \forall m > \hat{m} : I(m) = 0.$$

$\hat{m}$  ist die größte Masse mit Intensität größer 0. Oft wird zudem gefordert, dass

$$(ii) \quad \exists ! \tilde{m} : \forall m \neq \tilde{m} : I(m) < I(\tilde{m}).$$

Die Paare  $P = (m, I(m))$  mit  $I(m) \neq 0$  heißen *Peaks* des Spektrums.  $m$  ist die *Masse*,  $I(m)$  die *Intensität* des Peaks.  $\tilde{P} := (\tilde{m}, I(\tilde{m}))$  wird *Basispeak* des Spektrums genannt.  $\hat{P} := (\hat{m}, I(\hat{m}))$  ist der *Peak größter Masse*.

#### 5.3.2 Beispiel:

Tabelle 5.1 zeigt die Peaks des Massenspektrums aus Abbildung 5.2. Basispeak ist  $\tilde{P} = (74; 1, 000)$ , Peak größer Masse ist  $\hat{P} = (101; 0, 013)$ .

**5.3.3 Bemerkung:**

Bedingung (ii) aus Definition 5.3.1 besagt, dass das Spektrum korrekt angesteuert sein soll. Weisen zwei oder mehr Peaks maximale Intensität auf, so deutet dies auf eine übersteuerte Aufnahme hin. Eine weitere Erscheinung bei nicht optimal aufgenommenen Spektren ist so genanntes „Rauschen“, d.h. es existieren Peaks geringer Intensität bei nahezu jeder Masse. Für die Korrektur dieses Phänomens stehen mehrere Methoden zur Verfügung.

- *Wähle die  $k$  höchsten Peaks:* Seien die Peaks  $P_i = (m_i, I(m_i))$  des Spektrums  $I$  abfallend geordnet nach ihrer Intensität, d.h.  $I(m_i) \geq I(m_{i+1})$ . Dann setze

$$I'(m) := \begin{cases} I(m), & \text{falls } I(m) \geq I(m_k), \\ 0, & \text{sonst.} \end{cases}$$

- *Lösche alle Peaks unterhalb der  $p$ -fachen Intensität des Basispeaks:* Sei  $p \in ]0, 1[$ . Dann setze

$$I'(m) := \begin{cases} I(m), & \text{falls } I(m) \geq p \cdot I(\tilde{m}), \\ 0, & \text{sonst.} \end{cases}$$

Es ist zu beachten, dass bei der Anwendung dieser Methoden wichtige Informationen, z.B. in Form von Isotopenpeaks verloren gehen können. Eine andere, nicht mit Informationsverlust behaftete vorverarbeitende Manipulation des Spektrums ist die *Normierung* bezüglich der Intensität des Basispeaks. Während es in der Chemie üblich ist, die Intensität des Basispeaks auf 100 zu normieren, wollen wir der einfacheren Schreibung wegen einer Normierung des Basispeaks auf 1 den Vorzug geben:

$$I'(m) := \frac{I(m)}{I(\tilde{m})}.$$

Eine weitere vorverarbeitende Maßnahme ist die Partitionierung des Spektrums in Cluster dicht zusammen liegender Peaks. Solche Cluster lassen sich in ihrer Entstehung oft *einem* Ion zuordnen. Entsprechend wird bei der Interpretation versucht, jedem Cluster nach Möglichkeit *eine* Bruttoformel oder besser noch Strukturformel zuzuordnen (siehe z.B. [141]).

**5.3.4 Definition:**

Sei  $I$  ein niedrig aufgelöstes Massenspektrum und  $d > 0$ ,  $h > 0$ . Eine Menge von Peaks  $\mathcal{P} = \{(m_i, I(m_i)) \mid i \in k\}$  mit  $m_i < m_{i+1}$  für alle  $i$  heißt *Peakcluster* zur *Homogenität*  $h$  und *Diversität*  $d$ , wenn

$m$	$I(m)$	$m$	$I(m)$	$m$	$I(m)$	$m$	$I(m)$	$m$	$I(m)$
26	0,031	33	0,005	44	0,032	57	0,300	83	0,009
27	0,172	38	0,006	45	0,024	58	0,009	84	0,003
28	0,395	39	0,094	51	0,005	59	0,204	85	0,284
29	0,246	40	0,011	53	0,013	60	0,006	86	0,015
30	0,005	41	0,246	54	0,006	73	0,015	87	0,250
31	0,028	42	0,066	55	0,136	74	1,000	88	0,017
32	0,005	43	0,445	56	0,055	75	0,028	101	0,013

Tabelle 5.1: Peaks des Massenspektrums aus Abbildung 5.2

- $\forall i : m_i - m_{i-1} \leq h \wedge (m \in ]m_{i-1}, m_i[ \Rightarrow I(m) = 0)$
- $\forall m \in [m_0 - d, m_0[ \cup ]m_{k-1}, m_{k-1} + d] : I(m) = 0.$

Ein Peakcluster  $\mathcal{P}$  kann auch durch folgende Abbildung dargestellt werden:

$$I_{\mathcal{P}} : \mathbb{N}^* \longrightarrow \mathbb{R}_0^+, \quad m \longmapsto I_{\mathcal{P}}(m) := \begin{cases} I(m_i), & \text{falls } \exists i : m_i = m, \\ 0, & \text{sonst.} \end{cases}$$

### 5.3.5 Bemerkung:

Die Massen aufeinander folgender Peaks eines Peakclusters mit Homogenität  $h$  und Diversität  $d$  unterscheiden sich also höchstens um  $h$  Einheiten, während Peakcluster von Peaks außerhalb des Clusters durch mindestens  $d$  Massen ohne Intensität getrennt sein müssen. Damit jeder Peak genau einem Peakcluster zugeordnet werden kann, wählt man  $d = h$ . [87] und [161] empfehlen  $d = h = 2$ .

Das Spektrum aus Beispiel 5.3.2 kann demnach in 6 Peakcluster zerlegt werden, die sich über folgende Massenbereiche erstrecken: 26–33, 38–45, 51–60, 73–75, 83–88, 101.

Im nächsten Abschnitt wird gezeigt, wie sich die Intensitätsverhältnisse in den Peakclustern theoretisch aus der Bruttoformel lassen.

## 5.3.4 Isotope und theoretische Isotopenmuster

Die Atome eines chemischen Elements  $X \in \mathcal{E}$  besitzen nicht notwendig gleiche Atommassen. Die Masse eines Atoms wird im Wesentlichen durch den Atomkern bestimmt. Im Atomkern gibt es zwei Arten von Elementarteilchen mit ganzzahliger Masse 1: Protonen und *Neutronen*. Die Anzahl der Protonen ist durch das chemische Element festgelegt. Im Gegensatz zu Protonen sind Neutronen ungeladen und es können Atome des gleichen Elements  $X$



$X$	$\check{m}_X$	$\hat{m}_X$	$I_X(\check{m}_X)$	$I_X(\check{m}_X+1)$	$I_X(\check{m}_X+2)$
H	1	1	1	0	0
C	12	13	0,989	0,011	0
N	14	15	0,9963	0,0037	0
O	16	18	0,9976	0,0004	0,0020
F	19	19	1	0	0
Si	28	30	0,9223	0,0467	0,0310
P	31	31	1	0	0
S	32	34	0,9504	0,0075	0,0421
Cl	35	37	0,7577	0	0,2423
Br	79	81	0,5069	0	0,4931
I	127	127	1	0	0

Tabelle 5.2: Natürliche Isotopenverteilungen für die Elemente aus  $\mathcal{E}_{11}$ 

mit unterschiedlicher Anzahl von Neutronen auftreten. Solche Atome unterschiedlicher Masse bezeichnet man als *Isotope* von  $X$  (Schreibweise für das Isotop der Masse  $m$ :  ${}^mX$ , z.B.  ${}^{13}\text{C}$ ). Aus der Tatsache, dass Isotope stets mit bekannten, naturgegebenen relativen Häufigkeiten auftreten, kann man mit Hilfe der Massenspektrometrie Informationen über die elementare Zusammensetzung einer unbekanntes Substanz gewinnen.

### 5.3.6 Definition:

Sei  $X \in \mathcal{E}$  ein chemisches Element. Die *natürliche Isotopenverteilung* von  $X$  ist eine Abbildung

$$I_X : \mathbb{N}^* \longrightarrow \mathbb{R}_0^+, \quad m \longmapsto I_X(m)$$

für die gilt:

- (i)  $\exists \hat{m}_X : I_X(\hat{m}_X) > 0 \wedge \forall m > \hat{m}_X : I_X(m) = 0$ ,
- (ii)  $\exists ! \check{m}_X : \forall m \neq \check{m}_X : I_X(m) < I_X(\check{m}_X)$ .

$\check{m}_X$  ist die *nominale Masse*,  $\hat{m}_X$  die *größte* und  $\check{m}_X := \min\{m \mid I_X(m) \neq 0\}$  die *kleinste Isotopenmasse* von  $X$ . Zudem fordern wir eine Normierung der Intensitätssumme auf 1:

- (iii)  $\sum_m I_X(m) = 1$ .

Dies wird später in Satz 5.3.11 eine wichtige Rolle spielen.

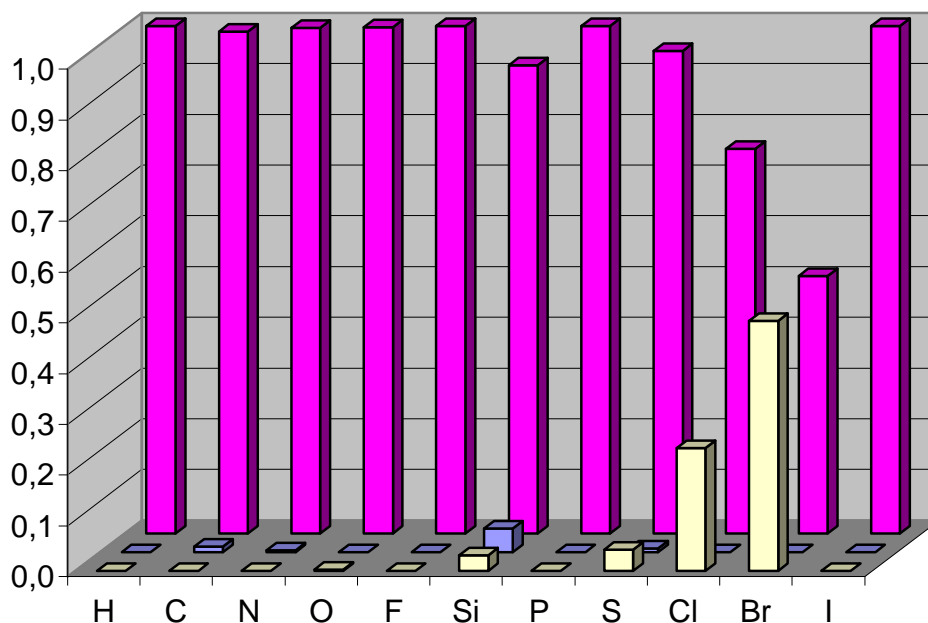


Abbildung 5.5: Graphische Darstellung der natürlichen Isotopenverteilungen

**5.3.7 Bemerkung:**

Tabelle 5.2 enthält die natürlichen Isotopenverteilungen<sup>1</sup> für die Elemente  $X$  aus  $\mathcal{E}_{11}$ . Für diese Elemente ist und  $\tilde{m}_X = \tilde{m}_X$ . Für Massen  $m$ , die in der Tabelle nicht angegeben sind, ist  $I_X(m) = 0$ . Die Elemente aus  $\mathcal{E}_{11}$  können hinsichtlich ihrer Isotopenverteilungen in drei Klassen eingeteilt werden (vgl. auch [91]):

- *Klasse 0*: Nominale Masse und größte Isotopenmasse stimmen überein:  $\hat{m}_X = \tilde{m}_X$ .
- *Klasse 1*: Nominale Masse und größte Isotopenmasse unterscheiden sich um eine Masseneinheit:  $\hat{m}_X = \tilde{m}_X + 1$ .
- *Klasse 2*: Nominale Masse und größte Isotopenmasse unterscheiden sich um zwei Masseneinheiten:  $\hat{m}_X = \tilde{m}_X + 2$ .

Abbildung 5.5 veranschaulicht die Isotopenverteilungen für die Elemente  $X \in \mathcal{E}_{11}$  graphisch. Dabei findet man in der hinteren Reihe  $I_X(\tilde{m}_X)$ , in der mittleren  $I_X(\tilde{m}_X+1)$  und in der vorderen  $I_X(\tilde{m}_X+2)$ . Leider sind die Intensitätsauflösungen realer Massenspektren ebenso begrenzt wie die optische

<sup>1</sup>Quelle: Mass Spectrometry and Chromatography - Scientific Instrument Services, Inc., [www.sisweb.com/referenc/source/exactmaa.htm](http://www.sisweb.com/referenc/source/exactmaa.htm)

Auflösung dieser Graphik. So können die Isotopenpeaks von N und O kaum erkannt werden.

Verschiedene Isotope desselben Elements besitzen exakt gleiche chemische Eigenschaften. Dies gilt insbesondere hinsichtlich ihres Bindungs- und Reaktionsverhaltens. Die einzelnen Moleküle einer chemische Verbindung können gemäß ihrer Zusammensetzung aus Isotopen verschiedene Molekülmassen besitzen. Die relativen Häufigkeiten der verschiedenen Molekülmassen richten sich nach den Isotopenverteilungen der in der Verbindung auftretenden Elemente und der Bruttoformel.

### 5.3.8 Beispiel:

Beispielsweise gibt es  $\text{Cl}_2$ -Moleküle der Masse 70 ( $^{35}\text{Cl}_2$ ), 72 ( $^{35}\text{Cl}^{37}\text{Cl}$ ) und 74 ( $^{37}\text{Cl}_2$ ). Für ihre relativen Häufigkeiten gilt:

$$\begin{aligned}\text{prob}(^{35}\text{Cl}_2) &= \text{prob}(^{35}\text{Cl}) \cdot \text{prob}(^{35}\text{Cl}) = 0,7577 \cdot 0,7577, \\ \text{prob}(^{35}\text{Cl}^{37}\text{Cl}) &= \text{prob}(^{35}\text{Cl}) \cdot \text{prob}(^{37}\text{Cl}) \cdot 2 = 0,7577 \cdot 0,2423 \cdot 2, \\ \text{prob}(^{37}\text{Cl}_2) &= \text{prob}(^{37}\text{Cl}) \cdot \text{prob}(^{37}\text{Cl}) = 0,2423 \cdot 0,2423.\end{aligned}$$

Für die Verteilung der Molekülmassen erhält man (nach Rundung):

$$I_{\text{Cl}_2}(m) = \begin{cases} 0,57410929 & \text{für } m = 70, \\ 0,36718142 & \text{für } m = 72, \\ 0,05870929 & \text{für } m = 74, \\ 0 & \text{sonst.} \end{cases}$$

Diese Vorgehensweise kann auf beliebige Bruttoformeln verallgemeinert werden. Dazu formulieren wir zunächst eine weitere Definition:

### 5.3.9 Definition:

Ein *Isotopenmuster* ist eine Abbildung

$$I : \mathbb{N}^* \longrightarrow \mathbb{R}_0^+, \quad m \longmapsto I(m)$$

für die gilt:

- (i)  $\exists \hat{m} : I(\hat{m}) > 0 \wedge \forall m > \hat{m} : I(m) = 0.$
- (ii)  $\sum_m I(m) = 1.$

$\hat{m}$  heißt die *größte Masse* von  $I$ .

**5.3.10 Bemerkung:**

Die natürlichen Isotopenverteilungen der Elemente sind Beispiele für Isotopenmuster. Auch die in 5.3.8 dargestellte Verteilung der Molekülmassen zu einer gegebenen Bruttoformel ist ein Isotopenmuster. Diese Berechnungen sind von wesentlicher Bedeutung für die Massenspektrometrie und sollen im Folgenden einer genaueren mathematischen Betrachtung unterzogen werden.

**5.3.11 Satz:**

Auf der Menge  $\mathcal{I}$  der Isotopenmuster existiert folgende innere Verknüpfung:

$$\mathcal{I} \times \mathcal{I} \longrightarrow \mathcal{I}, \quad (I_1, I_2) \longmapsto I_1 \cdot I_2,$$

wobei

$$(I_1 \cdot I_2)(m) := \sum_{i=0}^m I_1(i)I_2(m-i).$$

Die Verknüpfung, genannt *Faltung* ist assoziativ und für die größte Masse  $\hat{m}_{12}$  von  $I_1 \cdot I_2$  gilt:

$$\hat{m}_{12} = \hat{m}_1 + \hat{m}_2.$$

*Beweis:*

Zuerst wollen wir die Wohldefiniertheit der Faltung zeigen: Die größte Masse von  $I_1 \cdot I_2$  ist  $\hat{m}_{12} := \hat{m}_1 + \hat{m}_2$ , denn

$$(I_1 \cdot I_2)(\hat{m}_1 + \hat{m}_2) = \sum_{i=0}^{\hat{m}_1 + \hat{m}_2} I_1(i)I_2(\hat{m}_1 + \hat{m}_2 - i) \geq I_1(\hat{m}_1)I_2(\hat{m}_2) > 0.$$

und für  $m > \hat{m}_1 + \hat{m}_2$  gilt:

$$\begin{aligned} (I_1 \cdot I_2)(m) &= \sum_{i=0}^m I_1(i)I_2(m-i) \\ &= \sum_{a+b=m} I_1(a)I_2(b) \\ &= \sum_{\substack{a+b=m \\ a > \hat{m}_1}} \underbrace{I_1(a)}_{=0} I_2(b) + \sum_{\substack{a+b=m \\ a \leq \hat{m}_1}} I_1(a) \underbrace{I_2(b)}_{=0} = 0. \end{aligned}$$

Damit ist Bedingung (i) aus Definition 5.3.9 bewiesen. Zur Wohldefiniertheit bleibt (ii) zu zeigen:

$$\begin{aligned}
 \sum_m (I_1 \cdot I_2)(m) &= \sum_{m=0}^{\hat{m}_1 + \hat{m}_2} \sum_{i=0}^m I_1(i) I_2(m-i) \\
 &= \sum_{m=0}^{\hat{m}_1 + \hat{m}_2} \sum_{a+b=m} I_1(a) I_2(b) \\
 &= \sum_{a+b \leq \hat{m}_1 + \hat{m}_2} I_1(a) I_2(b) \\
 &= \left( \sum_{a=0}^{\hat{m}_1} I_1(a) \right) \left( \sum_{b=0}^{\hat{m}_2} I_2(b) \right) = 1 \cdot 1 = 1.
 \end{aligned}$$

Zur Assoziativität: Sei  $m \in \mathbb{N}^*$ . Dann gilt:

$$\begin{aligned}
 ((I_1 \cdot I_2) \cdot I_3)(m) &= \sum_{i=0}^m (I_1 \cdot I_2)(m-i) I_3(i) \\
 &= \sum_{i=0}^m \left( \sum_{j=0}^i I_1(j) I_2(i-j) \right) I_3(i) \\
 &= \sum_{i=0}^m \sum_{a+b=i} I_1(a) I_2(b) I_3(i) \\
 &= \sum_{a+b+c=m} I_1(a) I_2(b) I_3(c) \\
 &= \sum_{i=0}^m I_1(i) \sum_{b+c=m-i} I_2(b) I_3(c) \\
 &= \sum_{i=0}^m I_1(i) \sum_{j=0}^{m-i} I_2(j) I_3(m-i-j) \\
 &= \sum_{i=0}^m I_1(i) (I_2 \cdot I_3)(m-i) = (I_1 \cdot (I_2 \cdot I_3))(m).
 \end{aligned}$$

□

### 5.3.12 Definition:

Sei  $\mathcal{E} = \{X_i \mid i \in n\}$  eine Menge chemischer Elemente. und  $\beta \in \mathbb{N}^{\mathcal{E}}$  eine Bruttoformel. Dann ist das *theoretische Isotopenmuster* von  $\beta$  definiert durch

$$I_\beta := \prod_{X \in \mathcal{E}} I_X^{\beta(X)}.$$

**5.3.13 Lemma:**

Sei  $\beta \in \mathbb{N}^{\mathcal{E}}$  eine Bruttoformel. Für das theoretische Isotopenmuster von  $\beta$  gilt entweder

$$(i) \quad \exists l : \beta(X_l) = 1 \wedge \forall i \neq l : \beta(X_i) = 0 \\ \implies I_\beta = I_{X_l}$$

oder

$$(ii) \quad \exists l : \beta(X_l) \geq 1. \text{ Sei } \beta - X_l \in \mathbb{N}^{\mathcal{E}} \text{ definiert durch}$$

$$(\beta - X_l)(X) := \begin{cases} \beta(X_l) - 1, & \text{falls } X = X_l, \\ \beta(X), & \text{sonst.} \end{cases}$$

$$\implies I_\beta = I_{\beta - X_l} \cdot I_{X_l}.$$

*Beweis:*

Klar nach Definition von  $I_\beta$ . □

**5.3.14 Bemerkung:**

Lemma 5.3.13 liefert ein rekursives Verfahren zu Berechnung theoretischer Isotopenmuster von Bruttoformeln.

**5.3.15 Definition:**

Sei  $\beta \in \mathbb{N}^{\mathcal{E}}$  eine Bruttoformel. Dann ist

- $m_\beta := \sum_{X \in \mathcal{E}} \tilde{m}_X \beta(X)$  die *nominale* Masse,
- $\tilde{m}_\beta := \min\{m \mid \forall m' : I_\beta(m) \geq I_\beta(m')\}$  die Masse größter Intensität,
- $\hat{m}_\beta := \max\{m \mid I_\beta(m) > 0\}$  die größte Masse und
- $\check{m}_\beta := \min\{m \mid I_\beta(m) > 0\}$  die kleinste Masse

von  $\beta$ .

**5.3.16 Bemerkung:**

Für die größte Masse von  $\beta$  gilt

$$\hat{m}_\beta = \sum_{X \in \mathcal{E}} \hat{m}_X \beta(X)$$

und für die kleinste Masse

$$\check{m}_\beta = \sum_{X \in \mathcal{E}} \check{m}_X \beta(X).$$

Zwischen der Masse größter Intensität von  $\beta$  und den Elementen gibt es keinen ähnlich einfachen Zusammenhang. Insbesondere ist die nominale Masse nicht notwendig gleich der Masse größter Intensität, wie das folgende einfache Beispiel zeigt:

**5.3.17 Beispiel:**

Sei  $\beta = \text{Br}_2$ . Dann ist  $m_\beta = 2 \cdot 79 = 158$  und

$$I_\beta(m) = \begin{cases} 0,25694761 & \text{für } m = 158, \\ 0,49990478 & \text{für } m = 160, \\ 0,24314761 & \text{für } m = 162, \\ 0 & \text{sonst,} \end{cases}$$

d.h.  $\check{m}_\beta = 160 \neq m_\beta$ .

**5.3.5 Datenbasis aufgeklärter EI–Massenspektren**

Sowohl für die Bestimmung der Güte von Rankingfunktionen als auch zur Berechnung von MS–Klassifikatoren ist eine Datenbasis aufgeklärter Spektren unabdingbar. Wir verwenden in dieser Arbeit Spektren und Strukturen der *NIST* Massenspektren–Bibliothek<sup>2</sup>. Diese Bibliothek umfasst 107888 Spektren zu 107812 Strukturen. Spektren und Strukturen werden dabei in verschiedenen Files ausgeliefert und können über numerische Identifikatoren einander zugeordnet werden.

Um mit hoher Sicherheit nur stimmige Paare von Spektren und Strukturen für die folgenden Untersuchungen zu verwenden, wurden die Daten sehr restriktiven Konsistenzprüfungen unterzogen, bevor sie in die hier verwendete Datenbasis aufgenommen wurden:

---

<sup>2</sup>NIST/EPA/NIH Mass Spectral Library, NIST '98 Version, U.S. Department of Commerce, National Institute of Standards and Technology

Anzahl	H	C	N	O	F	Si	P	S	Cl	Br	I
1	176	124	17989	17936	1391	3583	1659	2224	4472	2573	590
2	316	369	11833	19725	494	1960	250	290	1933	692	118
3	510	645	3544	12057	1484	731	62	30	590	85	14
4	874	1266	2356	8485	295	404	10	3	332	62	7
5	1000	1758	980	4070	465	162	1	0	122	11	0
6 – 10	12218	24173	626	6043	1035	223	2	0	231	12	0
11 – 15	19143	25310	4	534	245	4	0	0	2	0	0
16 – 20	19678	16350	0	103	116	0	0	0	0	0	0
21 – 25	10949	8412	0	6	32	0	0	0	0	0	0
26 – 30	7731	4606	0	0	6	0	0	0	0	0	0
≥ 31	13022	2971	0	0	2	0	0	0	0	0	0
$\Sigma$	85617	85984	37332	68959	5565	7067	1984	2547	7682	3435	729

Tabelle 5.3: Atomares Profil der MS–Struktur–Datenbasis zu  $\mathcal{E}_{11}$ 

	$\mathcal{E}$	Minimum	1. Quartil	Median	Mittel	3. Quartil	Maximum
Atom- anzahl	$\mathcal{E}_{11}$	2	25	34	38,67	47	212
	$\mathcal{E}_4$	2	26	35	39,15	47	212
Molekül- masse	$\mathcal{E}_{11}$	2	178	242	267,26	330	1014
	$\mathcal{E}_4$	2	168	226	252,02	312	1014

Tabelle 5.4: Atomanzahlen und Molekülmassen der MS–Struktur–Datenbasis

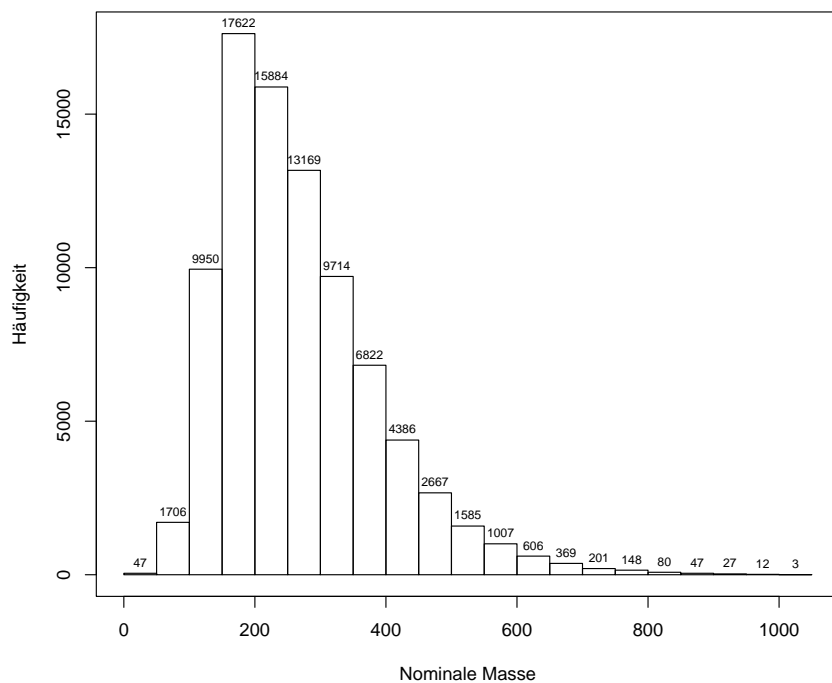
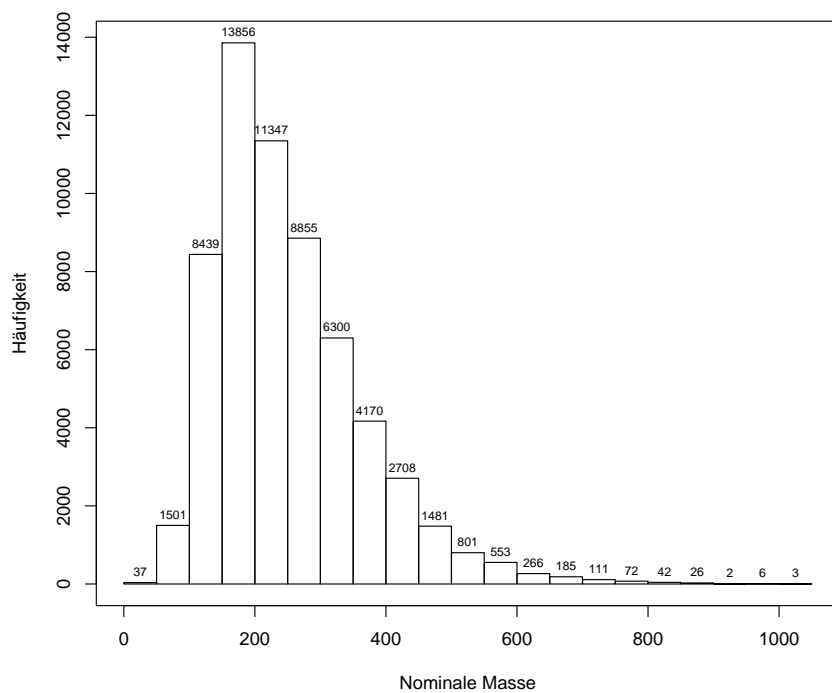
In dem Spektren–File sind neben Peaklisten und Identifikatoren auch Bruttoformeln und Namen der vermessenen Verbindungen abgelegt. In dem Struktur–File sind neben dem Identifikator ebenfalls Namen verzeichnet. Wenn trotz gleichem Identifikator Bruttoformel oder Name abweichen, werden die entsprechenden Daten verworfen.

Von den verbleibenden Struktur–Spektrum–Paaren wurden solche in unsere Datenbasis aufgenommen, die nur Elemente aus  $\mathcal{E}_{11}$ , keine isotopenmarkierten Atome, keine Ladungen und keine ungepaarten Elektronen besitzen. Auf diese Weise verbleiben 86052 Struktur–Spektrum–Paare. 60761 der Strukturen beschränken sich auf Elemente aus  $\mathcal{E}_4$ .

Zunächst wollen wir uns einen Überblick zu dieser Datenbasis gewinnen. Tabelle 5.3 zeigt das atomare Profil, Tabelle 5.4 gibt Auskunft über statistische Kennwerte für Atomanzahlen (inkl. H–Atome) und nominale Massen der Verbindungen unserer Datenbasis. Dabei unterscheiden wir zwischen der gesamten Datenbasis und den Verbindungen, die nur Elemente aus  $\mathcal{E}_4$  enthalten. Die Histogramme in Abbildungen 5.6 und 5.7 geben genaueren Aufschluss über die Verteilung der Molekülmassen in unserer Datenbasis.

Wir nutzen die Gelegenheit, um die Anwesenheit des Molekülions in den Massenspektren statistisch zu untersuchen. 73845 der 86052 Spektren zeigen einen Peak bei der nominalen Molekülmasse. Bei 12207 Spektren ist das



Abbildung 5.6: Molekülmassen der MS-Struktur-Datenbasis zu  $\mathcal{E}_{11}$ Abbildung 5.7: Molekülmassen der MS-Struktur-Datenbasis zu  $\mathcal{E}_4$

Molekülion abwesend. Dies entspricht einem Anteil von 14,186%.

Des Weiteren interessiert uns die Differenz aus der Masse des Basispeaks  $\tilde{m}_{max}$  des schwersten Peakclusters und der nominalen Molekülmasse  $m_M$ . Für diesen Massenunterschied ergibt sich folgende Verteilung

$m_M - \tilde{m}_{max}$	0	1	15	-2	-1	18	31	43	2	29
Häufigkeit	66539	4084	2782	1876	1201	970	542	433	387	383
rel. Häufigkeit	77,32	4,75	3,23	2,18	1,40	1,13	0,63	0,50	0,45	0,45

Die verbleibenden 6855 Struktur–Spektrum–Paare verteilen sich auf 360 weitere Massendifferenzen. Man kann das Wissen über die Verteilung dieser Massendifferenzen nutzen, um Kandidaten für die Molekülmasse zu bestimmen. Allerdings wird durch diese Statistik auch deutlich, dass die Bestimmung der Molekülmasse über entsprechende massenspektrometrische Methoden (SIMS) einen erheblichen Vorteil für die automatisierte Strukturaufklärung bedeutet.

Eingeschränkt auf Struktur–Spektrum–Paare aus  $\mathcal{E}_4$  ist das Molekülion in 7232 Spektren abwesend. Dies entspricht einen Anteil von 11,902%. Die Massendifferenzen  $m_M - \tilde{m}_{max}$  verteilen sich wie folgt:

$m_M - \tilde{m}_{max}$	0	1	15	-1	18	31	2	43	60	29
Häufigkeit	49358	3060	1161	907	901	481	327	309	225	217
rel. Häufigkeit	81,23	5,04	1,91	1,49	1,48	0,79	0,54	0,51	0,37	0,36

Bei den 3815 verbleibenden Struktur–Spektrum–Paare treten weitere 315 Massendifferenzen auf.

## 5.4 Rankingfunktionen für Massenspektren

Wie schon anfangs erwähnt, wird im Massenspektrometer nicht nur der Analyt selbst vermessen, sondern vielmehr ein Ionengemisch, welches aus dem Analyten durch Fragmentierung hervorgeht. Die positiv geladenen Teilchen werden detektiert und hinterlassen entsprechend ihrer Häufigkeit ein mehr oder weniger starkes Signal im Massenspektrum.

Diese Signale verteilen sich für jede in dem Gemisch auftretende Bruttoformel gemäß dem ihr zugeordneten theoretischen Isotopenmuster. Das Massenspektrum ist eine *Linearkombination* theoretischer Isotopenmuster mit positiven Koeffizienten.

### 5.4.1 Definition:

Seien  $\beta_i \in \mathbb{N}^{\mathcal{E}}$ ,  $i \in n$  Bruttoformeln und  $a_i \in \mathbb{R}_0^+$ . Dann ist die *Linearkombination* von  $I_{\beta_i}$  mit Koeffizienten  $a_i$  die Abbildung

$$\sum_{i \in n} a_i I_{\beta_i} : \mathbb{N}^* \longrightarrow \mathbb{R}_0^+, \quad m \longmapsto \sum_{i \in n} a_i I_{\beta_i}(m)$$

### 5.4.2 Bemerkung: Entstehung des Massenspektrums

Seien  $\beta_i$ ,  $i \in n$  die paarweise verschiedenen Bruttoformeln der im Massenspektrum  $I$  auftretenden Fragmentionen  $M_{ij}$ ,  $i \in n$ ,  $j \in n_i$ , wobei  $\beta_{M_{ij}} = \beta_i$ . Weiter nehmen wir an, dass die Fragmentionen  $M_{ij}$  mit Häufigkeiten  $a_{ij}$  vorkommen. Dann ist

$$I = \sum_{i \in n} \sum_{j \in n_i} a_{ij} I_{\beta_i}. \quad (5.1)$$

Die Berechnung der Häufigkeiten  $a_{ij}$  aus den Strukturformeln der Fragmentionen  $M_{ij}$  ist nach heutigem Stand der Wissenschaft nicht möglich. Versuche [11, 46, 47, 62, 68, 130], diese Werte mit selbstlernenden Systemen und neuronalen Netzen zu approximieren waren, nur auf eingeschränkten Substanzklassen erfolgreich.

Erschwerend kommt hinzu, dass selbst modernste Massenspektrometer nicht mit der mathematischen Exaktheit arbeiten, die zur sicheren Erkennung von Bruttoformeln durch ihre theoretischen Isotopenmuster notwendig wäre.

Die folgenden Betrachtungen werden diesen beiden Tatsachen Rechnung tragen, indem die Häufigkeiten als Unbekannte und der Messfehler als Zielfunktion eines Optimierungsproblems modelliert werden. Dazu schreiben wir Gleichung 5.1 als

$$I = \sum_{i \in n} \left( I_{\beta_i} \sum_{j \in n_i} a_{ij} \right). \quad (5.2)$$

Die Summen  $a_i := \sum_{j \in n_i} a_{ij}$ ,  $i \in n$  der Häufigkeiten der Fragmentionen sind in jedem Falle positiv und wir können 5.2 umformulieren zu:

$$I = \sum_{i \in n} a_i I_{\beta_i} \text{ mit } a_i > 0, i \in n. \quad (5.3)$$

Wir fassen diese Erkenntnisse zusammen in

#### 5.4.3 Satz:

Sei  $I$  ein mit mathematischer Exaktheit aufgenommenes Massenspektrum und  $\{\beta_i \mid i \in n\}$  die Menge aller Bruttoformeln von Fragmentionen, die an der Entstehung des Massenspektrums beteiligt sind. Dann existieren  $x_i \in \mathbb{R}_0^+$  für  $i \in n$ , so dass

$$I = \sum_{i \in n} x_i I_{\beta_i}.$$

*Beweis:* Klar nach Bemerkung 5.4.2, Gleichung 5.2 und 5.3.  $\square$

#### Vergleichswerte für Brutto- und Strukturformeln

Wir können obige Beziehung sowohl zur Bestimmung der Bruttoformel als auch der Strukturformel aus einem Massenspektrum  $I$  verwenden. Dazu werden wir mit Hilfe von Satz 5.4.3 zu einem gegebenen Kandidaten  $K$  für die Brutto- oder Strukturformel einen *Vergleichswert* (engl. *Matchvalue*, kurz *MV*) berechnen, der aussagt, wie plausibel  $I$  durch  $K$  erklärbar ist. Dabei soll der Vergleichswert

$$(R) \quad MV(I, K) \in [0, 1]$$

eine reelle Zahl mit großen/kleinen Werten sein, falls  $I$  gut/schlecht durch  $K$  begründet werden kann. Weitere Anforderungen für eine derartige *Rankingfunktion* liegen nahe: Für den korrekten Kandidaten  $K^T$  soll

$$(T) \quad MV(I, K^T) = 1$$

sein. Falschen Kandidaten  $K^F$  sollen Werte

$$(F) \quad MV(I, K^F) < MV(I, K^T)$$

zugeordnet werden. Leider können wir im Falle real gemessener Massenspektren keine Rankingfunktion angeben, die letzteren beiden Anforderungen generell entspricht. Immerhin können wir unter der theoretischen Annahme mathematisch exakt aufgenommener Massenspektren eine Rankingfunktion für Bruttoformeln definieren, die Bedingung (R) und (T) erfüllt.

Seien  $\beta_i$ ,  $i \in n$  die Summenformeln der Fragmentionen eines Kandidaten  $K$  für das Spektrum  $I$  gemäß Satz 5.4.3. Dann ist

$$\min_{\mathbf{x} \geq 0} \sum_m \left( I(m) - \sum_{i \in n} x_i I_{\beta_i}(m) \right) = 0, \quad (5.4)$$

falls  $K = K^T$  der korrekte Kandidat ist. Können Intensitätsanteile durch die theoretischen Isotopenmuster  $I_{\beta_i}$  nicht erklärt werden, so nimmt die linke Seite in 5.4 einen positiven Wert an.

Es ist sinnvoll, große Intensitätsanteile, die nicht erklärt werden können, stärker zu gewichten als mehrere kleine Abweichungen gleichen Gesamtbeitrages. Dies erreicht man beispielsweise durch Quadrieren der Differenzen. Wir erhalten

$$\min_{\mathbf{x} \geq 0} \sum_m \left( I(m) - \sum_{i \in n} x_i I_{\beta_i}(m) \right)^2 \in [0, \Sigma_m (I(m))^2]. \quad (5.5)$$

Durch Normierung können wir eine Rankingfunktion definieren, die Bedingung (R), und für mathematisch exakt aufgenommene Massenspektren auch Bedingung (T) erfüllt:

$$\text{MV}(I, K) = 1 - \sqrt{\left( \sum_m (I(m))^2 \right)^{-1} \min_{\mathbf{x} \geq 0} \sum_m \left( I(m) - \sum_{i \in n} x_i I_{\beta_i}(m) \right)^2} \quad (5.6)$$

Die Quadratwurzel im Funktionsterm bewirkt eine gleichmäßigere Verteilung der Vergleichswerte, da das Minimum aus 5.5 im Allgemeinen näher bei Null als an der oberen Intervallgrenze liegt. Zudem wird so die Bezeichnung „erklärbarer Anteil der Gesamtintensität“ für unsere Rankingfunktion gerechtfertigt. In diesen Vergleichswert fließen zwei wichtige Typen von Informationen, die niedrig aufgelöste EI-Massenspektren liefern auf sehr natürliche Art und Weise ein, nämlich

- die An- und Abwesenheit von Peaks bei bestimmten Massen und
- die Übereinstimmung mit theoretischen Isotopenmustern.

Auf Ebene der Bruttoformel, d.h. ohne Kenntnis der Struktur der Kandidaten, sind dies sogar alle verfügbaren Informationen.

Im Folgenden werden wir alle Teilprobleme beleuchten, die zur Berechnung dieses Vergleichswertes gelöst werden müssen. Zunächst gilt es, die Menge möglicher Fragment-Summenformeln  $\beta_i$  eines Kandidaten für die Brutto- oder Strukturformel zu bestimmen.

Im Falle einer Bruttoformel  $\beta$  kommen zunächst alle  $\beta' \subseteq \beta$  in Frage. Um die Dimension des quadratischen Optimierungsproblems 5.5 klein zu halten, ist es wichtig, nur solche  $\beta'$  zu verwenden, die auch tatsächlich zur Erklärung von Peaks beitragen. Eine erste Verbesserung erhält man, in dem man sich auf Fragment-Bruttoformeln  $\beta'$  beschränkt, bei deren nominalen Massen  $m_{\beta'}$  das untersuchte Spektrum  $I$  einen Peak besitzt, also

$$\{\beta_i \mid i \in n\} = \{\beta' \subseteq \beta \mid I(m_{\beta'}) > 0\}$$

Insbesondere benötigt man natürlich eine Methode, solche Bruttoformeln zu generieren.

### Generierung von Bruttoformeln

Kandidaten  $\beta \in \mathbb{N}^{\mathcal{E}}$  für die Bruttoformel mit einer gegebenen nominalen Masse  $m$  müssen die Diophantische Gleichung

$$\sum_{X \in \mathcal{E}} \tilde{m}_X \beta(X) = m$$

erfüllen. Wir lösen diese Gleichung mit Hilfe eines Backtrack-Algorithmus. Oft werden Bruttoformeln innerhalb eines gegebenen Masse-Intervalls benötigt. Obere und untere Schranken für die Anzahlen von Atomen bestimmter Elemente sollen berücksichtigt werden können. Auf diese Weise kann man Vorwissen über die elementare Zusammensetzung verarbeiten. Der folgende Algorithmus generiert alle Bruttoformeln  $\beta$ , deren nominale Masse innerhalb  $[m_{\min}, m_{\max}]$  liegen und kompatibel sind zu der weichen Bruttoformel  $B$  mit  $B(X_i) = [\beta_{\min}(X_i), \beta_{\max}(X_i)]$ , wobei  $\mathcal{E} = \{X_i \mid i \in n\}$ :

#### 5.4.4 Algorithmus: $GenMolForm(\beta_{\min}, \beta_{\max}, m_{\min}, m_{\max})$

- (1) *begin()*
- (2) **while** *EndFlag = false* **do**
- (3)     *Output*( $\beta$ )
- (4)     *next()*
- (5) **end**

#### Funktion: *begin()*

- (1) *EndFlag*  $\leftarrow$  *false*
- (2)  $\beta \leftarrow \beta_{\min}$
- (3) **if**  $m_{\beta} < m_{\min} \vee m_{\max} < m_{\beta}$
- (4)     *next()*
- (5) **end**

**Funktion:** *next()*

```
(1)  do step()
(2)  while  $\neg(m_{\min} \leq m_{\beta} \leq m_{\max}) \wedge \text{EndFlag} = \text{false}$ 
```

**Funktion:** *step()*

```
(1)  for each  $i \in n$  do
(2)      if  $\beta(X_i) < \beta_{\max}(X_i) \wedge m_{\beta} + m_{X_i} \leq m_{\max}$ 
(3)           $\beta(X_i) \leftarrow \beta(X_i) + 1$ 
(4)          return
(5)      else
(6)           $\beta(X_i) \leftarrow \beta_{\min}(X_i)$ 
(7)      end
(8)  end
(9)   $\text{EndFlag} \leftarrow \text{true}$ 
```

Anhang D enthält Tabellen mit Anzahlen von Bruttoformeln zu verschiedenen nominalen Massen zwischen 1 und 1000. Anhang E listet alle Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_4}^{\mathcal{C}}$  mit Massen zwischen 1 und 150 und mindestens einem C-Atom auf. Für die Berechnung von Vergleichswerten nach Gleichung 5.6 ist es sinnvoll, nur solche Bruttoformeln zu berücksichtigen, deren nominale Massen tatsächlich im Massenspektrum auftreten.

### Bruttoformeln und Ionentypen

Wie schon erwähnt, können die Elemente im Massenspektrometer wegen der außerordentlichen energetischen Bedingungen sehr variable Atomzustände annehmen. Die Notwendigkeit der Existenzkriterien (Gr1) und (Gr2) aus Satz 1.3.19 gilt nicht mehr. Insbesondere können auch einfach positiv geladene Ionen auftreten, die Bedingung (Gr1) nicht erfüllen. Man unterscheidet dabei Teilchen, die ein ungepaartes Elektron besitzen (engl. *Odd Electron Ion*, kurz *OEI*) und solche, die keine ungepaarten Elektronen besitzen (engl. *Even Electron Ion*, kurz *EEI*). Für OEI ist (Gr1) erfüllt, für EEI nicht. Bedingungen (Gr2) und (Con) werden wir auch für Ionen beibehalten, wenngleich dadurch wenige mögliche Bruttoformeln von Fragmentionen ausgeschlossen werden. Dies wird gerechtfertigt durch die deutliche Beschränkung der Anzahl möglicher Bruttoformeln (vgl. Anhang D).

### Lösung des quadratischen Optimierungsproblems

Zur Berechnung von Vergleichswerten nach Gleichung 5.6 ist ein Optimierungsproblem der Form

$$\min \sum_{j \in p} \left( c_j - \sum_{i \in n} x_i d_{ij} \right)^2$$

NB:  $x_i \geq 0, \quad i \in n$

zu lösen. Es handelt sich dabei um ein *kleinstes Quadrate* (engl. *Least Squares*, kurz *LS*) Problem mit Nebenbedingungen, welche hier mit dem Algorithmus NLPQL [126] bearbeitet wird.

#### 5.4.1 Ranking von Bruttoformeln

##### Berechnung von Vergleichswerten für Bruttoformeln

Für ein gegebenes Massenspektrum  $I$  und eine Bruttoformel  $\beta$  seien  $\beta_i \subseteq \beta$ ,  $i \in n$  alle Summenformeln, die

$$\text{(Frag)} \quad I(m_{\beta_i}) > 0,$$

$$\text{(Gr2)} \quad \sum_{X \in \mathcal{E}} v_X \beta_i(X) - 2 \max_{\beta_i(X) > 0} v_X \geq 0 \text{ und}$$

$$\text{(Con)} \quad \sum_{X \in \mathcal{E}} v_X \beta_i(X) - \sum_{X \in \mathcal{E}} 2 \cdot \beta_i(X) + 2 \geq 0$$

erfüllen. Dann berechnen wir den Vergleichswert von  $\beta$  bzgl.  $I$  durch

$$\text{MV}(I, \beta) = 1 - \sqrt{\left( \sum_m (I(m))^2 \right)^{-1} \min_{\mathbf{x} \geq 0} \sum_m \left( I(m) - \sum_{i \in n} x_i I_{\beta_i}(m) \right)^2} \quad (5.7)$$

##### 5.4.5 Beispiel:

Wir berechnen den Vergleichswert von  $\beta = \text{C}_6\text{H}_{12}\text{O}_2$  für das Spektrum  $I$  aus Beispiel 5.3.2. Tabelle 5.5 listet die Summenformeln  $\beta_i \subseteq \beta$ , die (Frag), (Gr2) und (Con) erfüllen. Die letzte Spalte gibt jeweils den Wert für  $x_i$  an, mit dem das Isotopenmuster von  $\beta_i$  für eine Lösung von 5.7 gewichtet wird. Als Vergleichswert erhalten wir  $\text{MV}(I, \beta) = 0,9942863$ . Dieser Wert erscheint — ganz subjektiv betrachtet — recht gut. Spektrum  $I$  gehört übrigens tatsächlich zu einer Verbindung mit Summenformel  $\text{C}_6\text{H}_{12}\text{O}_2$ . Der Grund für die Abweichung zum Vergleichswert 1, den man für den korrekten Kandidaten erwarten würde, liegt an dem Peak bei Masse 75. Dieser hat für einen Isotopenpeak von  $\text{C}_3\text{H}_6\text{O}_2$  zu niedrige Intensität. Aufgrund des theoretischen Isotopenmusters von  $\text{C}_3\text{H}_6\text{O}_2$  müsste man bei Masse 75 ein Peak der Höhe 0,034



$\beta_i$	$m_{\beta_i}$	$x_i$
C <sub>2</sub> H <sub>2</sub>	26	0,0317
C <sub>2</sub> H <sub>3</sub>	27	0,1751
C <sub>2</sub> H <sub>4</sub>	28	0,3999
C <sub>2</sub> H <sub>5</sub>	29	0,2426
CH <sub>2</sub> O	30	0,0000
C <sub>2</sub> H <sub>6</sub>	30	0,0000
CH <sub>3</sub> O	31	0,0283
CH <sub>4</sub> O	32	0,0047
O <sub>2</sub>	32	0,0000
HO <sub>2</sub>	33	0,0049
C <sub>3</sub> H <sub>2</sub>	38	0,0062
C <sub>3</sub> H <sub>3</sub>	39	0,0970
C <sub>3</sub> H <sub>4</sub>	40	0,0081
C <sub>2</sub> O	40	0,0000
C <sub>2</sub> HO	41	0,0093
C <sub>3</sub> H <sub>5</sub>	41	0,2447
C <sub>2</sub> H <sub>2</sub> O	42	0,0593
C <sub>3</sub> H <sub>6</sub>	42	0,0000
C <sub>2</sub> H <sub>3</sub> O	43	0,0000
C <sub>3</sub> H <sub>7</sub>	43	0,4585
C <sub>2</sub> H <sub>4</sub> O	44	0,0000
C <sub>3</sub> H <sub>8</sub>	44	0,0177
CO <sub>2</sub>	44	0,0000
CHO <sub>2</sub>	45	0,0236

$\beta_i$	$m_{\beta_i}$	$x_i$
C <sub>2</sub> H <sub>5</sub> O	45	0,0000
C <sub>4</sub> H <sub>3</sub>	51	0,0052
C <sub>3</sub> HO	53	0,0068
C <sub>4</sub> H <sub>5</sub>	53	0,0067
C <sub>3</sub> H <sub>2</sub> O	54	0,0057
C <sub>4</sub> H <sub>6</sub>	54	0,0000
C <sub>3</sub> H <sub>3</sub> O	55	0,0000
C <sub>4</sub> H <sub>7</sub>	55	0,1419
C <sub>3</sub> H <sub>4</sub> O	56	0,0507
C <sub>4</sub> H <sub>8</sub>	56	0,0000
C <sub>2</sub> O <sub>2</sub>	56	0,0000
C <sub>2</sub> HO <sub>2</sub>	57	0,3064
C <sub>3</sub> H <sub>5</sub> O	57	0,0000
C <sub>4</sub> H <sub>9</sub>	57	0,0000
C <sub>2</sub> H <sub>2</sub> O <sub>2</sub>	58	0,0000
C <sub>3</sub> H <sub>6</sub> O	58	0,0000
C <sub>4</sub> H <sub>10</sub>	58	0,0021
C <sub>2</sub> H <sub>3</sub> O <sub>2</sub>	59	0,0739
C <sub>3</sub> H <sub>7</sub> O	59	0,1354
C <sub>2</sub> H <sub>4</sub> O <sub>2</sub>	60	0,0000
C <sub>3</sub> H <sub>8</sub> O	60	0,0000
C <sub>5</sub>	60	0,0000
C <sub>6</sub> H	73	0,0160
C <sub>3</sub> H <sub>5</sub> O <sub>2</sub>	73	0,0000

$\beta_i$	$m_{\beta_i}$	$x_i$
C <sub>4</sub> H <sub>9</sub> O	73	0,0000
C <sub>6</sub> H <sub>2</sub>	74	0,0000
C <sub>3</sub> H <sub>6</sub> O <sub>2</sub>	74	1,0374
C <sub>4</sub> H <sub>10</sub> O	74	0,0000
C <sub>6</sub> H <sub>3</sub>	75	0,0000
C <sub>3</sub> H <sub>7</sub> O <sub>2</sub>	75	0,0000
C <sub>4</sub> H <sub>3</sub> O <sub>2</sub>	83	0,0000
C <sub>5</sub> H <sub>7</sub> O	83	0,0000
C <sub>6</sub> H <sub>11</sub>	83	0,0096
C <sub>4</sub> H <sub>4</sub> O <sub>2</sub>	84	0,0000
C <sub>5</sub> H <sub>8</sub> O	84	0,0025
C <sub>6</sub> H <sub>12</sub>	84	0,0000
C <sub>4</sub> H <sub>5</sub> O <sub>2</sub>	85	0,0889
C <sub>5</sub> H <sub>9</sub> O	85	0,2111
C <sub>4</sub> H <sub>6</sub> O <sub>2</sub>	86	0,0000
C <sub>5</sub> H <sub>10</sub> O	86	0,0000
C <sub>4</sub> H <sub>7</sub> O <sub>2</sub>	87	0,0000
C <sub>5</sub> H <sub>11</sub> O	87	0,2637
C <sub>4</sub> H <sub>8</sub> O <sub>2</sub>	88	0,0031
C <sub>5</sub> H <sub>12</sub> O	88	0,0000
C <sub>6</sub> O	88	0,0000
C <sub>5</sub> H <sub>9</sub> O <sub>2</sub>	101	0,0138

Tabelle 5.5: Berechnung des Vergleichswertes für C<sub>6</sub>H<sub>12</sub>O<sub>2</sub> zu dem Spektrum aus Beispiel 5.3.2

vorfinden. Jedoch ist  $I(75) = 0,028$ . Die anderen Fragment-Bruttoformeln für Masse 74, C<sub>6</sub>H<sub>2</sub> und C<sub>4</sub>H<sub>10</sub>O, würden sogar zu einer noch größeren Differenz führen. Leider stößt man bei experimentellen Spektren des Öfteren auf Abweichungen bei Intensitäten von Isotopenpeaks, was die automatische Verifikation zumindest hinsichtlich der Summenformel-Bestimmung stark beeinträchtigt.

Um eine Vorstellung zu gewinnen, in wie fern unser Vergleichswert zur Ermittlung der Bruttoformel hilfreich sein kann, wollen wir alle Bruttoformeln  $\beta \in \mathcal{B}_{\mathcal{E}_{11}}^C$  mit Masse  $m_{\beta} = 116$  generieren, deren Vergleichswerte zu  $I$  berechnen und hinsichtlich dieser Werte abfallend anordnen. Wir sprechen dann von einem *Ranking*. Insgesamt gibt es 1451 Summenformeln mit Masse 116, wovon 220 aus  $\mathcal{B}_{\mathcal{E}_{11}}^C$  sind. 23 Summenformeln haben einen besseren Vergleichswert als C<sub>6</sub>H<sub>12</sub>O<sub>2</sub>. Erlaubt man nur die Elemente aus  $\mathcal{E}_4$ , dann gibt es insgesamt 162 Summenformeln zu Masse 116, 24 davon sind aus  $BfSetCon\mathcal{E}_4$  und C<sub>6</sub>H<sub>12</sub>O<sub>2</sub> belegt Rang 9 unter diesen 24 Formeln.

Tabelle enthält die besten 40 Summenformeln  $\beta \in \mathcal{B}_{\mathcal{E}_{11}}^C$  zusammen mit ihren Vergleichswerten. In [67] werden drei empirisch ermittelte *Filter* angegeben, die in der Natur selten vorkommende Summenformeln ausschließen. Nach Anwendung dieser Filter verbleiben 153 Bruttoformeln mit Elementen aus

	$\beta$	$MV(I, \beta)$	Filter
1	C <sub>3</sub> H <sub>5</sub> N <sub>2</sub> OP	0,9995987	×
2	C <sub>3</sub> H <sub>8</sub> N <sub>4</sub> O	0,9994732	
3	C <sub>4</sub> H <sub>5</sub> N <sub>2</sub> OF	0,9990449	×
4	C <sub>4</sub> H <sub>5</sub> O <sub>2</sub> P	0,9981975	×
5	C <sub>4</sub> H <sub>8</sub> N <sub>2</sub> O <sub>2</sub>	0,9978946	×
6	C <sub>5</sub> H <sub>12</sub> N <sub>2</sub> O	0,9976136	×
7	C <sub>3</sub> H <sub>8</sub> N <sub>2</sub> OSi	0,9974305	×
8	C <sub>3</sub> H <sub>4</sub> N <sub>2</sub> O <sub>3</sub>	0,9966711	
9	C <sub>4</sub> H <sub>9</sub> N <sub>2</sub> P	0,9962826	×
10	C <sub>3</sub> H <sub>5</sub> N <sub>4</sub> F	0,9962825	×
11	C <sub>4</sub> H <sub>12</sub> N <sub>4</sub>	0,9962643	×
12	C <sub>5</sub> H <sub>9</sub> OP	0,9962090	×
13	C <sub>3</sub> H <sub>5</sub> N <sub>2</sub> FSi	0,9961477	×
14	C <sub>5</sub> H <sub>9</sub> N <sub>2</sub> F	0,9960570	×
15	C <sub>2</sub> H <sub>4</sub> N <sub>4</sub> O <sub>2</sub>	0,9958909	
16	C <sub>5</sub> H <sub>8</sub> O <sub>3</sub>	0,9958316	×
17	C <sub>5</sub> H <sub>5</sub> O <sub>2</sub> F	0,9952805	×
18	C <sub>2</sub> H <sub>4</sub> N <sub>2</sub> O <sub>2</sub> Si	0,9949604	×
19	C <sub>2</sub> H <sub>5</sub> N <sub>2</sub> SiP	0,9948218	×
20	C <sub>2</sub> H <sub>8</sub> N <sub>4</sub> Si	0,9948070	×
21	C <sub>2</sub> H <sub>5</sub> N <sub>4</sub> P	0,9947663	
22	C <sub>2</sub> H <sub>8</sub> N <sub>6</sub>	0,9947546	
23	C <sub>3</sub> H <sub>5</sub> OSiP	0,9945391	×
24	C <sub>6</sub> H <sub>12</sub> O <sub>2</sub>	0,9942863	×
25	C <sub>4</sub> H <sub>8</sub> O <sub>2</sub> Si	0,9941528	×
26	C <sub>4</sub> H <sub>5</sub> OFSi	0,9940334	×
27	C <sub>6</sub> H <sub>9</sub> OF	0,9940301	×
28	C <sub>4</sub> H <sub>12</sub> N <sub>2</sub> Si	0,9898372	×
29	C <sub>6</sub> H <sub>16</sub> N <sub>2</sub>	0,9898041	×
30	C <sub>7</sub> H <sub>16</sub> O	0,9867311	×
31	C <sub>5</sub> H <sub>12</sub> OSi	0,9867233	×
32	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub> FP	0,9786070	×
33	C <sub>5</sub> H <sub>6</sub> FP	0,9786049	×
34	C <sub>4</sub> H <sub>9</sub> SiP	0,9785747	×
35	C <sub>4</sub> H <sub>2</sub> N <sub>2</sub> F <sub>2</sub>	0,9782367	×
36	C <sub>6</sub> H <sub>13</sub> P	0,9782361	×
37	C <sub>6</sub> H <sub>6</sub> F <sub>2</sub>	0,9782056	×
38	C <sub>3</sub> HN <sub>2</sub> O <sub>2</sub> F	0,9768266	×
39	C <sub>2</sub> HN <sub>2</sub> O <sub>2</sub> P	0,9765410	
40	C <sub>2</sub> HN <sub>4</sub> OF	0,9760385	

Tabelle 5.6: Ranking von Bruttoformeln mit Masse 116 zu dem Spektrum aus Beispiel 5.3.2

$\mathcal{E}_{11}$ , von denen sich C<sub>6</sub>H<sub>12</sub>O<sub>2</sub> auf Platz 19 in die Rangordnung einreicht. Eingeschränkt auf die Elemente aus  $\mathcal{E}_4$  findet man C<sub>6</sub>H<sub>12</sub>O<sub>2</sub> auf Rang 5 von 9 Formeln. Abbildung 5.8 gibt einen Überblick über die Vergleichswerte aller 220 Bruttoformeln. Solche aus  $\mathcal{E}_4$  sind dabei grau, diejenigen welche den Filter nach S. Heuerding und T. Clerc [67] erfolgreich durchlaufen mit einem Punkt markiert. Auf die korrekte Bruttoformel ist ein Pfeil gerichtet.

Wirft man einen genaueren Blick auf die Ranglisten, so stellt man fest, dass Summenformeln mit mehreren verschiedenen Elementen die vorderen Plätze belegen. Dies liegt daran, dass für diese Kandidaten die Menge der möglichen Summenformeln von Fragmentationen größer ist. Insbesondere werden für Summenformeln mit mehreren verschiedenen monoisotopischen Elementen oft besonders hohe Vergleichswerte berechnet, da sich experimentelle Peakcluster besonders gut durch theoretische Isotopenmuster mit wenig oder gar ohne Isotopenanteil erklären lassen.

Nun, da wir in der Lage sind, Vergleichswerte für Bruttoformeln bzgl. eines Massenspektrums zu berechnen und Ranglisten zu erstellen, schiebt sich eine weitere Frage in den Vordergrund: Wie viele Kandidaten aus einer Hitliste muss man eigentlich berücksichtigen, wenn man mit einer vorgegebenen Zuverlässigkeit den richtigen Kandidaten nicht ausschließen möchte?

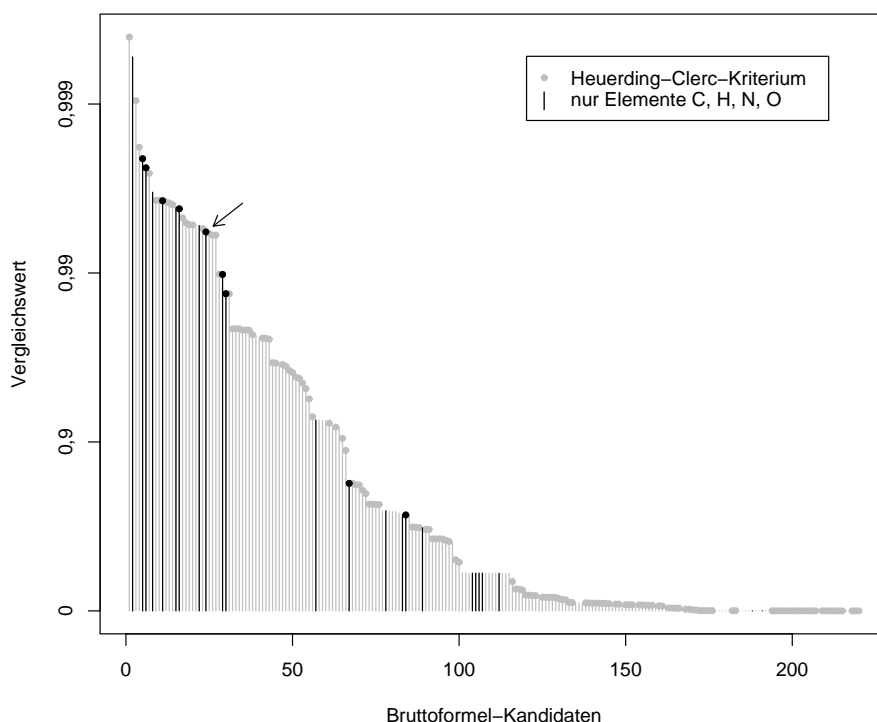


Abbildung 5.8: Vergleichswerte für die Bruttoformeln mit Masse 116

### Selektion relevanter Kandidaten aus einem Ranking

Dazu betrachten wir die Verteilung der Vergleichswerte korrekter Kandidaten für die Bruttoformel bzgl. der zugehörigen Spektren. Für eine zufällige Stichprobe von  $n = 1000$  Spektren  $I_i$  berechnen wir die jeweiligen Vergleichswerte der zugehörigen Summenformeln  $\beta_i$ .

Abbildung 5.9 zeigt die Verteilung der Vergleichswerte  $x_i := MV(I_i, \beta_i)$  als Histogramm.

Für  $p = \frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}$  bestimmen wir nun so genannte  $p$ -Quantile von  $(x_i)_{i \in n}$ . Dabei heißt eine Zahl  $q_p \in \mathbb{R}$  ein  $p$ -Quantil von  $(x_i)_{i \in n}$ , wenn

$$\frac{1}{n} |\{i \in n \mid x_i \leq q_p\}| \geq p \wedge \frac{1}{n} |\{i \in n \mid x_i \geq q_p\}| \geq 1 - p.$$

Abbildung 5.9 veranschaulicht die Bestimmung von  $p$ -Quantilen graphisch, Tabelle 5.7 enthält  $p$ -Quantile für verschiedene  $p$ .

Möchte man nun innerhalb der Stichprobe zu einem zufällig gewählten Spektrum  $I_j$ ,  $j \in n$  eine möglichst kleine Familie  $(\beta_i)_{i \in \Omega}$  von Summenformeln bestimmen, die mit einer Wahrscheinlichkeit  $\geq p$  die korrekte Formel  $\beta_j$  enthält, so genügt es alle  $\beta_i$  mit  $MV(I_j, \beta_i) \geq q_{1-p}$  zu betrachten.

Die berechneten Quantile dienen im Folgenden dazu, für einen beliebigen

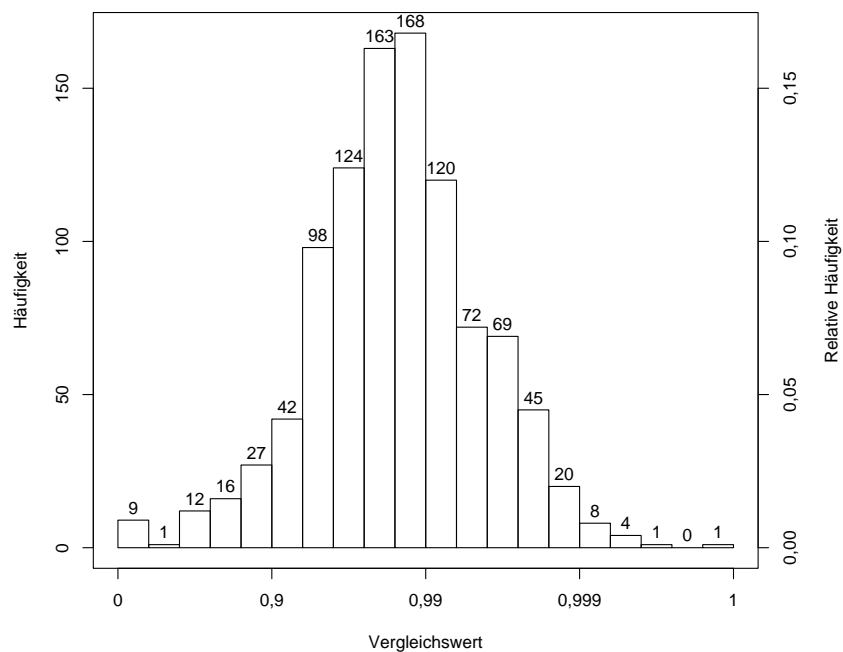


Abbildung 5.9: Histogramm der Bruttoformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren

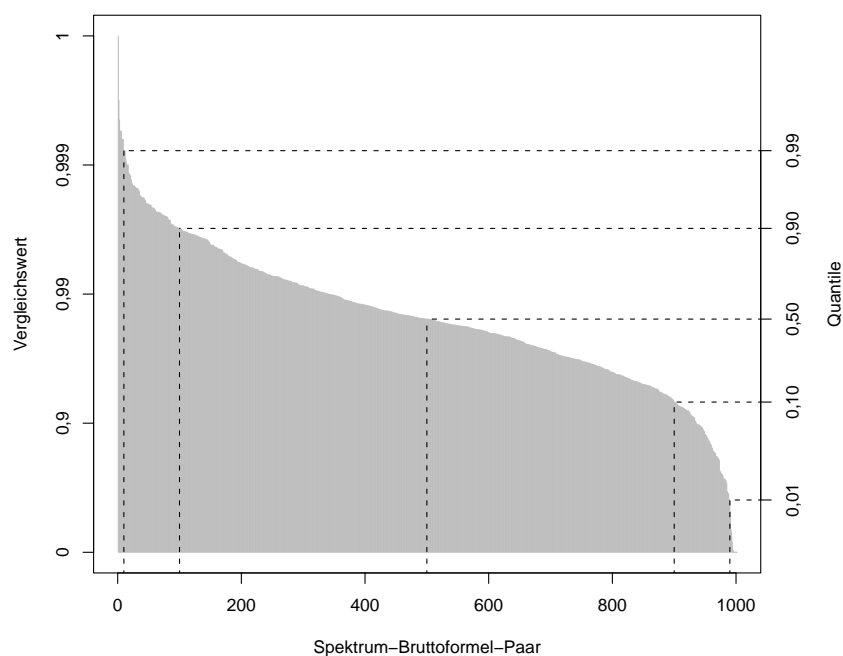


Abbildung 5.10: Verteilung der Bruttoformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren

$p$	$q_p$	$p$	$q_p$	$p$	$q_p$
0,01	0,6056967	0,10	0,9313282	0,91	0,9970228
0,02	0,7323542	0,20	0,9596450	0,92	0,9974521
0,03	0,8171856	0,30	0,9720852	0,93	0,9975946
0,04	0,8519153	0,40	0,9801180	0,94	0,9977639
0,05	0,8775297	0,50	0,9843779	0,95	0,9979752
0,06	0,8969463	0,60	0,9878099	0,96	0,9982111
0,07	0,9111062	0,70	0,9913571	0,97	0,9984834
0,08	0,9205837	0,80	0,9941656	0,98	0,9987761
0,09	0,9255477	0,90	0,9969016	0,99	0,9992254

Tabelle 5.7: Quantile  $q_p$  für Bruttoformel–Vergleichswerte zu verschiedenen Wahrscheinlichkeiten  $p$

Summenformel–Kandidaten zu entscheiden, ob er bei vorgegebener Zuverlässigkeit akzeptiert werden soll, oder nicht. Generiert man dann zu einem der Spektren unserer Stichprobe alle Summenformeln  $\beta$  mit korrekter Masse  $m = m_\beta$ , gibt eine Zuverlässigkeit  $p$  vor und wählt alle diejenigen  $\beta$  aus, deren Vergleichswerte  $MV(I, \beta) \geq q_{1-p}$  sind, so gehört mit Wahrscheinlichkeit  $p$  die korrekte Summenformel zu unserer Auswahl.

#### 5.4.6 Beispiel:

Wir entnehmen unserer Stichprobe jeweils 100 Spektren, deren zugehörige Verbindungen nur Elemente aus  $\mathcal{E}_{11}$  bzw.  $\mathcal{E}_4$  enthalten. Für jedes dieser Spektren generieren wir zunächst alle Bruttoformeln, die korrekte Molekülmasse haben, berechnen dann Vergleichswerte und nehmen ein Ranking vor. Um die Qualität der Rangfolgen für verschiedene Beispiele vergleichen zu können, definieren wir einen Kennwert, der die Position des korrekten Kandidaten in der Rangfolge zur Gesamtzahl der Kandidaten in Beziehung setzt: Die *relative Ranking-Position* (kurz *RRP*)

$$RRP := \frac{\text{Position des korrekten Kandidaten} - 1}{\text{Gesamtzahl der Kandidaten} - 1}.$$

hat den Wert 0, falls der korrekte Kandidat an erster bzw. 1, falls dieser an letzter Position platziert ist. Sie wächst linear mit steigender Position des korrekten Kandidaten im Ranking und ist sinnvoller Weise für einelementige Kandidatenmengen nicht definiert. Um die Anzahl der Kandidaten überschaubar zu halten, wurden nur Datensätze mit maximaler Mo-

lekülmasse 200 gewählt. Folgende Tabelle fasst die Ergebnisse zusammen:

$\mathcal{E}$	Min.	1.Quart.	Median	Mittel	3.Quart.	Max.
$\mathcal{E}_4$	0,0000	0,0632	0,1962	0,2463	0,4206	0,7273
$\mathcal{E}_{11}$	0,0000	0,0098	0,0685	0,1037	0,1443	0,8022

Die erste Spalte enthält die Minima der relativen Ranking-Positionen. Es gibt also für beide Grundmengen  $\mathcal{E}_4$  und  $\mathcal{E}_{11}$  von Elementen Fälle, bei denen der korrekte Kandidat an erster Stelle platziert ist. Das *erste Quartil* ist ein Synonym für das 25%-Quantil. In mindestens einem Viertel der Fälle ist die *RRP* kleiner als 0,06176 bzw. 0,00983. Der *Median* bezeichnet das 50%-Quantil, das *dritte Quartil* das 75%-Quantil. Spalte *Mittel* enthält den arithmetischen Mittelwert.

Abbildung 5.11 zeigt die Verteilung relativen Ranking-Positionen basierend auf den Elementen aus  $\mathcal{E}_4$  als Histogramm. In 36% der Fälle ist die RRP also kleiner als 0,1. Das entsprechende Diagramm zu Elementen aus  $\mathcal{E}_{11}$  enthält Abbildung 5.13. Hier sind die Rangfolgen besser als für die Beispiele zu  $\mathcal{E}_4$ . Dies liegt einerseits daran, dass Spektren zu Verbindungen mit stark isotopehaltigen Elementen wie Silizium, Schwefel, Chlor oder Brom gut erkannt werden. Andererseits erhalten falsche Bruttoformel-Kandidaten mit isotopehaltigen Elementen deutlich schlechtere Vergleichswerte und finden sich auf hinteren Plätzen der Rangfolge wieder.

Zuletzt wollen wir noch untersuchen in wie vielen Fällen sich die korrekte Bruttoformel unter den vorgeschlagenen Kandidaten befindet. Folgende Tabelle zeigt diese Anzahlen für verschiedene Verlässlichkeiten  $r$ :

$\mathcal{E}$	$r = 0,99$	$r = 0,95$	$r = 0,90$	$r = 0,75$	$r = 0,50$
$\mathcal{E}_4$	99	98	93	85	66
$\mathcal{E}_{11}$	98	96	87	76	55

Der Scatterplots von Abbildung 5.12 und 5.14 zeigen die Ranking-Position und die Kandidaten-Anzahl bei Verlässlichkeit 0,9. Punkte oberhalb der Diagonalen repräsentieren demnach Fälle, bei denen die korrekte Summenformel fälschlicherweise ausgeschlossen würde, Punkte auf und unterhalb der Diagonalen repräsentieren Beispiele, bei denen sich die korrekte Bruttoformel unter den ausgewählten Kandidaten befindet.

### Reduzierung der Dimension

Bevor wir uns der Berechnung von Vergleichswerten für Strukturformeln zuwenden, sollen an dieser Stelle noch einige Details zur Berechnung der Rankingfunktion nachgetragen werden. Ein wichtiger Schritt, um die Anwendung

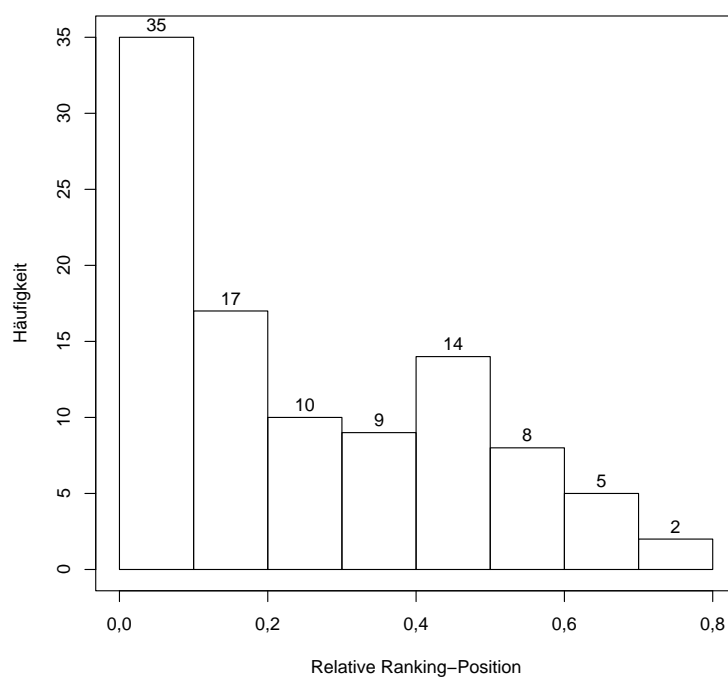


Abbildung 5.11: Histogramm der RRP für Bruttoformeln von 100 Massenspektren zu Verbindungen aus  $\mathcal{E}_4$

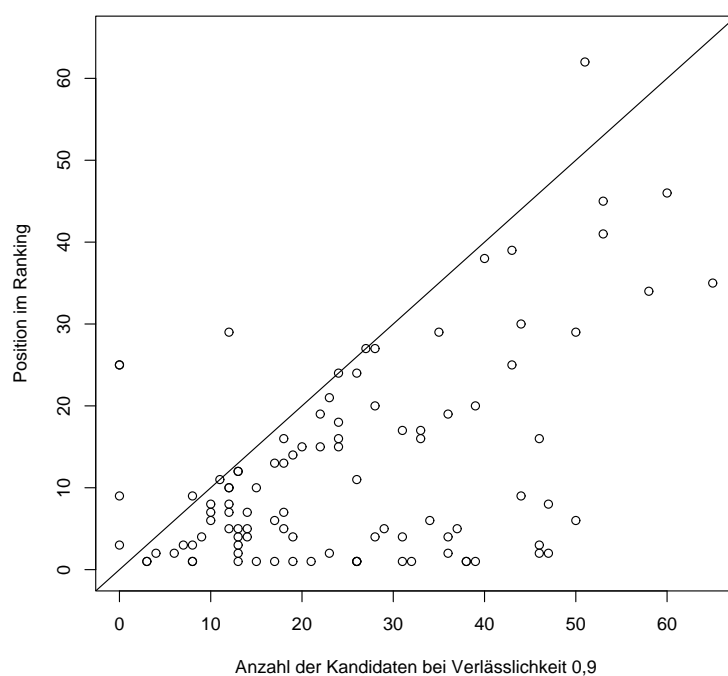


Abbildung 5.12: Ranking-Position und Kandidaten-Anzahl bei Verlässlichkeit 0,9 für Bruttoformeln zu Verbindungen aus  $\mathcal{E}_4$

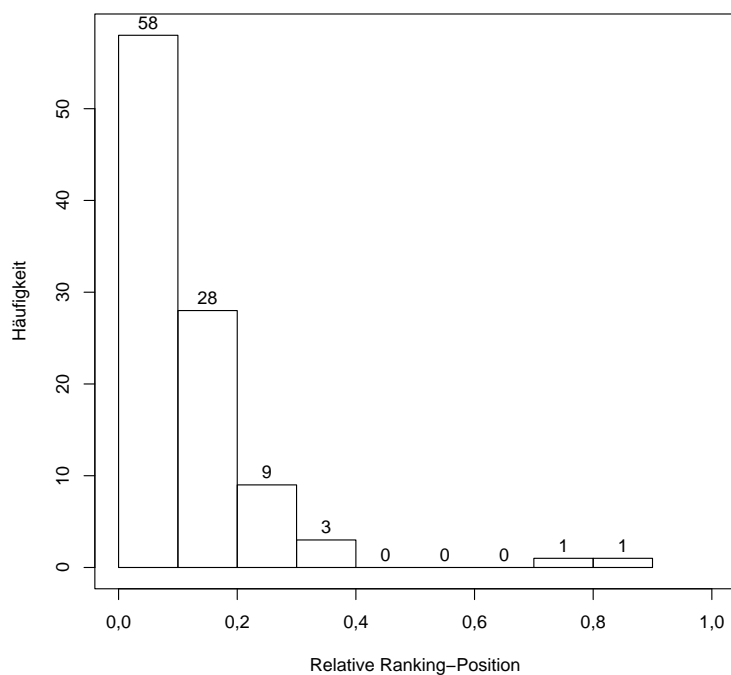


Abbildung 5.13: Histogramm der RRP für Bruttoformeln von 100 Massenspektren zu Verbindungen aus  $\mathcal{E}_{11}$

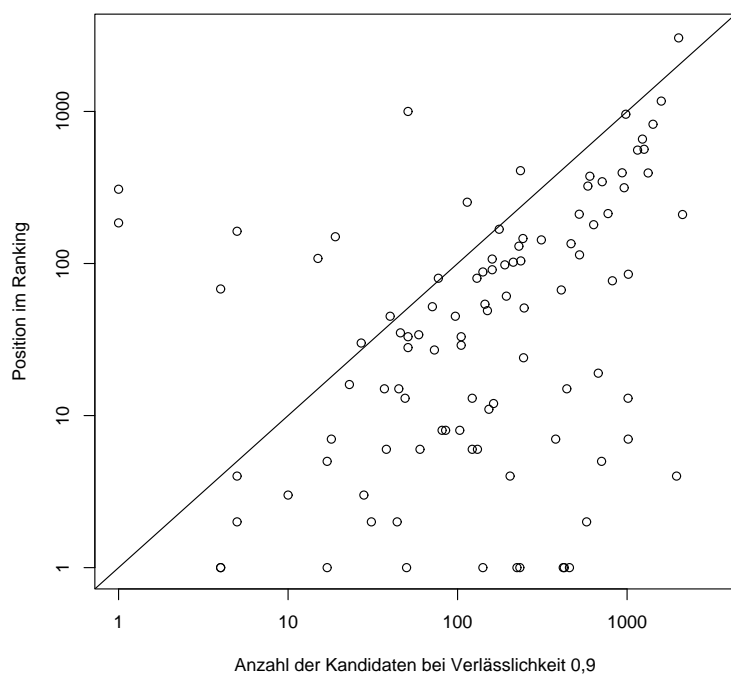


Abbildung 5.14: Ranking-Position und Kandidaten-Anzahl bei Verlässlichkeit 0,9 für Bruttoformeln zu Verbindungen aus  $\mathcal{E}_{11}$



des Verfahrens praktikabel zu machen besteht darin, das Spektrum gemäß Bemerkung 5.3.5 zu partitionieren und die Rankingfunktion zunächst lokal für die einzelnen Peakcluster zu berechnen. Somit muss man mehrere Optimierungsprobleme wesentlich kleinerer Dimension lösen. Der Vergleichswert für das gesamte Spektrum ergibt sich dann aus der Summe der Werte für die einzelnen Peakcluster: Sei  $\bigcup_{i \in n} \mathcal{P}_i$  die Zerlegung des Spektrums  $I$  in Peakcluster gemäß Bemerkung 5.3.5. Zu einem Bruttoformel-Kandidaten  $\beta$  seien  $\beta_{ij} \leq \beta$ ,  $i \in n$ ,  $j \in n_j$  diejenigen Bruttoformeln mit  $I_{\mathcal{P}_i}(m_{\beta_{ij}}) > 0$ , die (Gr2) und (Con) erfüllen. Dann ist

$$\text{MV}(I, \beta) = 1 - \sqrt{\left( \sum_m (I(m))^2 \right)^{-1} \sum_{i \in n} \min_{\mathbf{x}_i \geq 0} \sum_m \left( I_{\mathcal{P}_i}(m) - \sum_{j \in n_i} x_{ij} I_{\beta_{ij}}(m) \right)^2}$$

eine Rankingfunktion, die mit deutlich geringerem Rechenaufwand ausgewertet werden kann. Bei der Stichprobe von 1000 Datenbankspektren betrug die durchschnittliche CPU-Zeit<sup>3</sup> pro Spektrum 0,75s statt vorher 6,8s. Die so ermittelten Vergleichswerte weisen nur in wenigen Fällen (28 von 1000) Differenzen zu den über das gesamte Spektrum berechneten Werten auf. Diese Differenzen waren für unsere Stichprobe meist so klein (in 25 der 28 Fälle  $< 10^{-5}$ ), dass sie für die Kandidaten-Selektion nicht relevant wären.

Eine weitere Verringerung der Dimension der Optimierungsprobleme ließe sich erreichen, indem man nur deutlich verschiedene theoretische Isotopenmuster  $I_{\beta_{ij}}$  berücksichtigt.

### Simultane Berechnung von Bruttoformel und Isotopenmuster

Bei dem Algorithmus zur Generierung von Bruttoformeln entsteht die jeweils nächste Bruttoformel aus ihrem Vorgänger durch Inkrementierung der Anzahl von Atomen eines Elements um 1 (Algorithmus 5.4.4, Funktion Step(), Zeile 3). Es liegt nahe, an dieser Stelle auch das Isotopenmuster für die nächste Bruttoformel über 5.3.13 zu berechnen. Dies ist, falls große Anzahlen von Bruttoformeln und deren Isotopenmuster generiert werden müssen von Vorteil, da so in jedem Schritt nur eine Faltung von Isotopenmustern notwendig ist, während nach 5.3.12 ohne weitere Optimierungen für Bruttoformeln mit  $n$  Atomen  $n - 1$  Faltungen notwendig sind.

Abgesehen von der Zeitersparnis bei großen Problemen eröffnet die Berechnung der Isotopenmuster während der Bruttoformel-Generierung eine weitere Verbesserungsmöglichkeit: Wie schon in Beispiel 5.3.17 gezeigt, gibt es

<sup>3</sup>gemessen auf einem PC Pentium III mit 833MHz

Bruttoformeln  $\beta$ , bei denen die nominale Masse  $m_\beta$  und die Masse größter Intensität  $\tilde{m}_\beta$  nicht identisch sind. Um solchen Fällen besser gerecht zu werden, müsste man Bedingung (Frag) für Gleichung 5.7 ersetzen durch

$$\text{(Frag')} \quad I(\tilde{m}_{\beta_i}) > 0.$$

Sind Isotopenmuster der  $\beta_i$  während der Generierung von Fragment-Bruttoformeln bekannt, so kann diese Bedingung mit ebenso wenig Aufwand überprüft werden, wie zuvor (Frag).

## 5.4.2 Ranking von Strukturformeln

### Berechnung von Vergleichswerten für Strukturformeln

Wie bereits erwähnt, verläuft die Fragmentierung im Massenspektrometer nach größtenteils bekannten Reaktionsschemata. Wir werden dieses Wissen nutzen, um für Strukturformel-Kandidaten  $M$  einen Vergleichswert bezüglich einem experimentellen Massenspektrum  $I$  zu berechnen. Die Berechnung verläuft ähnlich wie bei Bruttoformel-Kandidaten. Allerdings können wir im Falle von Strukturformeln die Menge der möglichen Fragment-Bruttoformeln  $\beta_i$  deutlich einschränken. Wir werden nun nur noch solche Bruttoformeln berücksichtigen, zu denen es Fragmente gibt, die durch sukzessive Anwendung von Ionisierungs- und Fragmentierungsreaktionen aus  $M$  hervorgehen. Um diese zu simulieren, greifen wir auf unsere Vorarbeiten aus Abschnitt 1.5 zurück.

In Bemerkung 1.4.3 wurde darauf hingewiesen, dass für die Definition von Reaktionsschemata im Massenspektrometer ein eigener Atomtyp eingeführt wurde. Wir verwenden diesen Atomtyp, um verschiedene Mengen von Elementen zu kodieren:

A: alle Elemente

C: Kohlenstoff

H: Wasserstoff

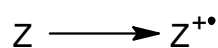
Y: alle schweren Atome (d.h. alle Elemente außer H)

Z: alle Atome mit freien Elektronenpaaren (N, O, P, S, Halogene)

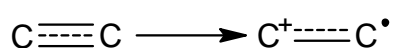
Die Reaktionen im Massenspektrometer können in drei Klassen unterteilt werden: Ionisierungs-, Fragmentierungs- und Umlagerungsreaktionen. Durch die Ionisierung im MS entsteht aus einem Molekül ohne Ladung und ungepaarten Elektronen ein einfach positiv geladenes Ion mit einem ungepaarten

Elektron. Ionisierungsreaktionen finden ausschließlich zu Beginn des Fragmentierungsweges statt. Wir werden folgende Ionisierungsreaktionen verwenden:

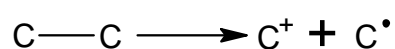
- n-Ionisation



- $\pi$ -Ionisation

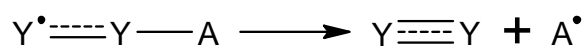


- $\sigma$ -Ionisation

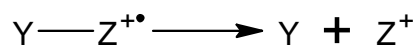


Alternativen für Bindungsvielfachheiten sind dabei als gestrichelte Linien kodiert. Nachdem eine der Ionisierungsreaktionen abgelaufen ist, können verschiedene Fragmentierungsreaktionen stattfinden. Während des Reaktionsverlaufs entstandene Neutralteilchen spielen für die weitere Fragmentierung keine Rolle. Auf positiv geladene Teilchen werden sukzessive weitere Fragmentierungsreaktionen angewandt. Wir wollen folgende Zerfalls- und Umlagerungsreaktionen berücksichtigen:

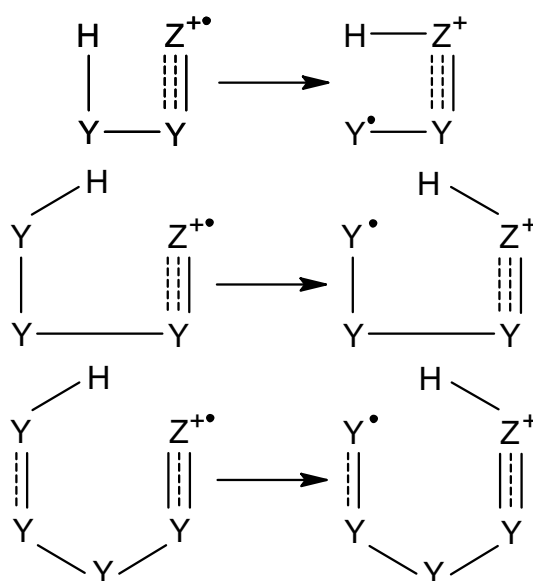
- $\alpha$ -Spaltung



- $\sigma$ -Spaltung



- H-Umlagerungen



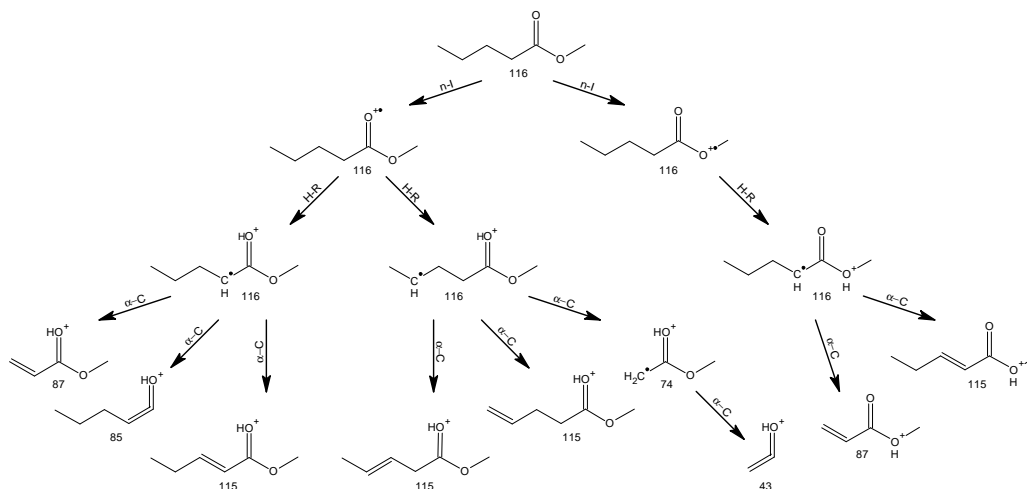


Abbildung 5.15: MS-Reaktionen von n-Pentansäuremethylester

Zur Berechnung des Vergleichswertes werden die Bruttoformeln  $\beta_i$ ,  $i \in n$  der Fragmentationen bestimmt, für die  $I(\tilde{m}_{\beta_i}) > 0$  ist, und dann

$$MV(I, M) = 1 - \sqrt{\left( \sum_m (I(m))^2 \right)^{-1} \min_{\mathbf{x} \geq 0} \sum_m \left( I(m) - \sum_{i \in n} x_i I_{\beta_i}(m) \right)^2}$$

wie zuvor im Falle von Bruttoformel-Kandidaten berechnet. Natürlich ist auch hier die Reduzierung der Dimension des Optimierungsproblems sehr ratsam. Für die folgenden Berechnungen wurden die Spektren gemäß Bemerkung 5.3.5 in Peakcluster aufgeteilt. An dieser Stelle sei darauf hingewiesen, dass die oben angegebenen Reaktionen keinerlei Anspruch auf Vollständigkeit erheben. Vielmehr bilden sie eher ein minimales System, um die Vorgänge im Massenspektrometer zu beschreiben. In den folgenden Betrachtungen wird gezeigt, in wie fern sie geeignet sind, um Vergleichswerte und Rangfolgen für Strukturkandidaten zu bestimmen.

#### 5.4.7 Beispiel:

Abbildung 5.15 zeigt die Abfolge der MS-Reaktionen für n-Pentansäuremethylester. Es finden n-Isomerisation (n-I), H-Umlagerungen mit 4 Atomen (H-R) und  $\alpha$ -Spaltungen ( $\alpha$ -C) statt. Dabei entstehen Fragmentationen der Massen 116, 115, 87, 85, 74 und 43. H-Umlagerungen mit 5 Atomen und Fragmentationen, die durch  $\sigma$ -Ionisation entstehen wurden aus Gründen der Überschaubarkeit in Abbildung 5.15 vernachlässigt. Durch H-Umlagerungen

$\beta_i$	$\tilde{m}_{\beta_i}$	$x_i$	$\beta_i$	$\tilde{m}_{\beta_i}$	$x_i$
C <sub>2</sub> H <sub>5</sub>	29	0,2515	C <sub>3</sub> H <sub>5</sub> O <sub>2</sub>	73	0,0156
C <sub>2</sub> H <sub>3</sub> O	43	0,0000	C <sub>3</sub> H <sub>6</sub> O <sub>2</sub>	74	1,0379
C <sub>3</sub> H <sub>7</sub>	43	0,4606	C <sub>5</sub> H <sub>9</sub> O	85	0,3008
CHO <sub>2</sub>	45	0,0242	C <sub>5</sub> H <sub>10</sub> O	86	0,0000
C <sub>4</sub> H <sub>9</sub>	57	0,3134	C <sub>4</sub> H <sub>7</sub> O <sub>2</sub>	87	0,2619
C <sub>2</sub> H <sub>3</sub> O <sub>2</sub>	59	0,2093	C <sub>5</sub> H <sub>9</sub> O <sub>2</sub>	101	0,0138
C <sub>2</sub> H <sub>4</sub> O <sub>2</sub>	60	0,0013			

Tabelle 5.8: Berechnung des Vergleichswertes für n-Pentansäuremethylester zu dem Spektrum aus Beispiel 5.3.2

mit 5 Atomen entstehen keine weiteren Ionen, die für die Erklärung des Spektrums relevant sind.

Abbildung 5.2 zeigt ein Massenspektrum  $I$  von n-Pentansäuremethylester. Beim Vergleich des Spektrums und der Massen von Fragmentionen, die über n-Ionisation entstehen, stellt man fest, dass einige Peaks nicht erklärt werden können. Dazu gehören die Intensitäten bei Masse 57, 55, 41, 39, 29, 28 und 27. Einige dieser Peaks können durch  $\sigma$ -Ionisation erklärt werden. Dabei entstehen 8 weitere Fragmentionen der Massen 101, 87, 73, 59, 57, 43, 29 und 15.

In Abbildung 5.16 werden alle 32 Fragmentionen aufgeführt, die über obige Reaktionsschemata aus n-Pentansäuremethylester hervorgehen. Die Strukturen sind aufsteigend nach ihrer Masse sortiert. Die Masse ist jeweils auf der rechten Seite der Kopfzeile abzulesen. Wie wir sehen, befinden sich unter unseren virtuellen Fragmenten keine mit Masse 27, 28, 39, 41 oder 55.

Als nächster Schritt zur Berechnung des Vergleichswertes werden die Summenformeln der Fragmentionen bestimmt. Tabelle 5.8 listet alle diese Bruttoformeln  $\beta_i$  auf, bei deren Masse größter Intensität  $I(\tilde{m}_{\beta_i}) > 0$  ist. In der letzten Spalte ist wiederum die Lösung  $x_i$  des Optimierungsproblems angegeben. Wir erhalten einen Vergleichswert  $MV(I, M) = 0,6052978$ .

Man kann diese Werte verwenden, um den Anteil erklärbarer Intensität seinerseits als Spektrum darzustellen, indem man  $I' = \sum_i x_i \beta_i$  berechnet. Abbildung 5.17 zeigt im oberen Bereich nochmals das gemessene Spektrum  $I$ , unten den Anteil erklärbarer Intensität  $I'$  und in der Mitte die absolute Differenz  $|I - I'|$ .

Als nächsten Schritt wollen wir untersuchen, ob unser Vergleichswert geeignet ist, verschiedene Strukturformeln hinsichtlich ihrer Relevanz bezüglich des experimentellen Spektrums zu unterscheiden. Dazu generieren wir alle Konstitutionsisomere zur Summenformel C<sub>6</sub>H<sub>12</sub>O<sub>2</sub>. Wir erhalten insgesamt

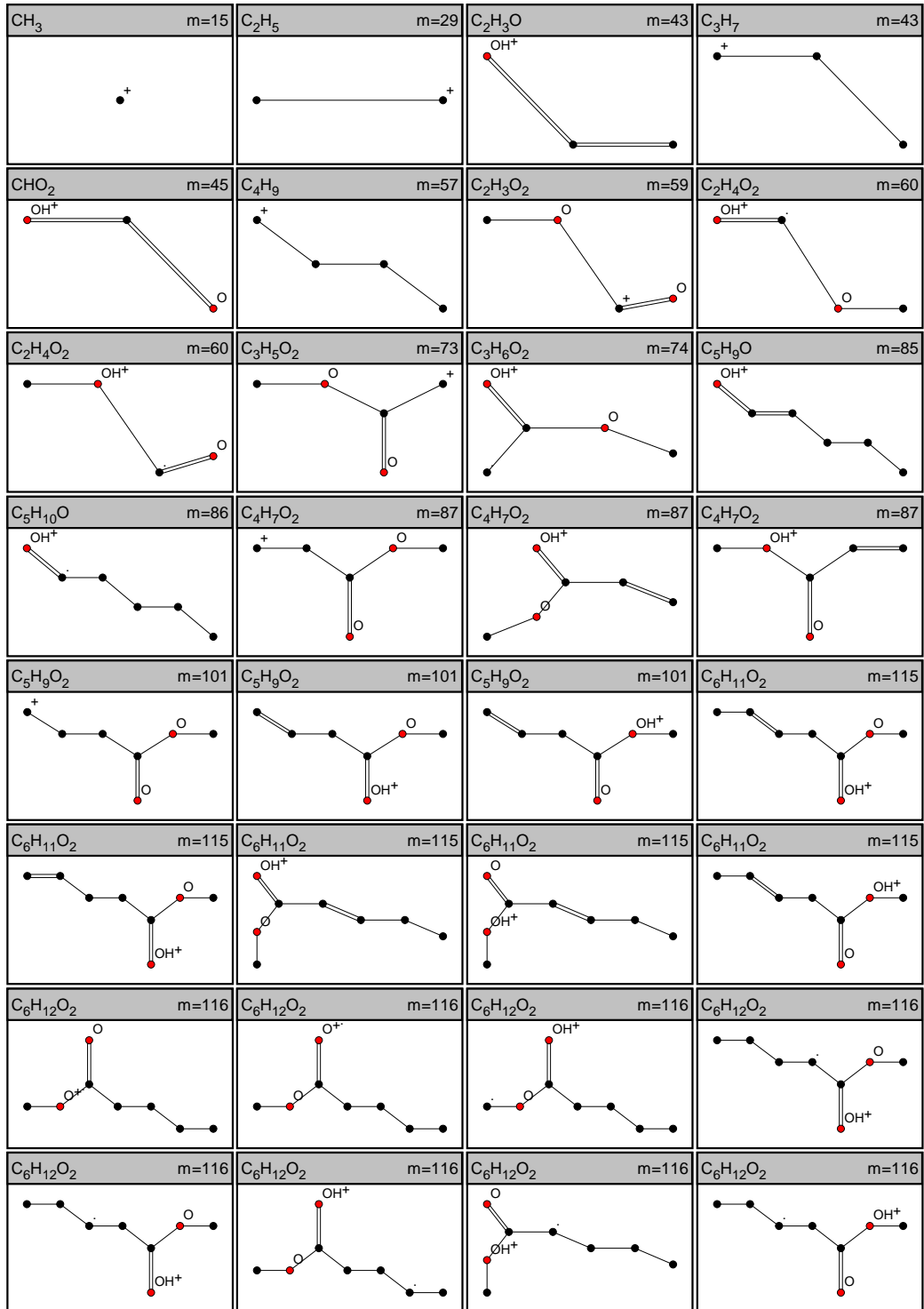


Abbildung 5.16: Mögliche Fragmentationen von n-Pentansäuremethylester

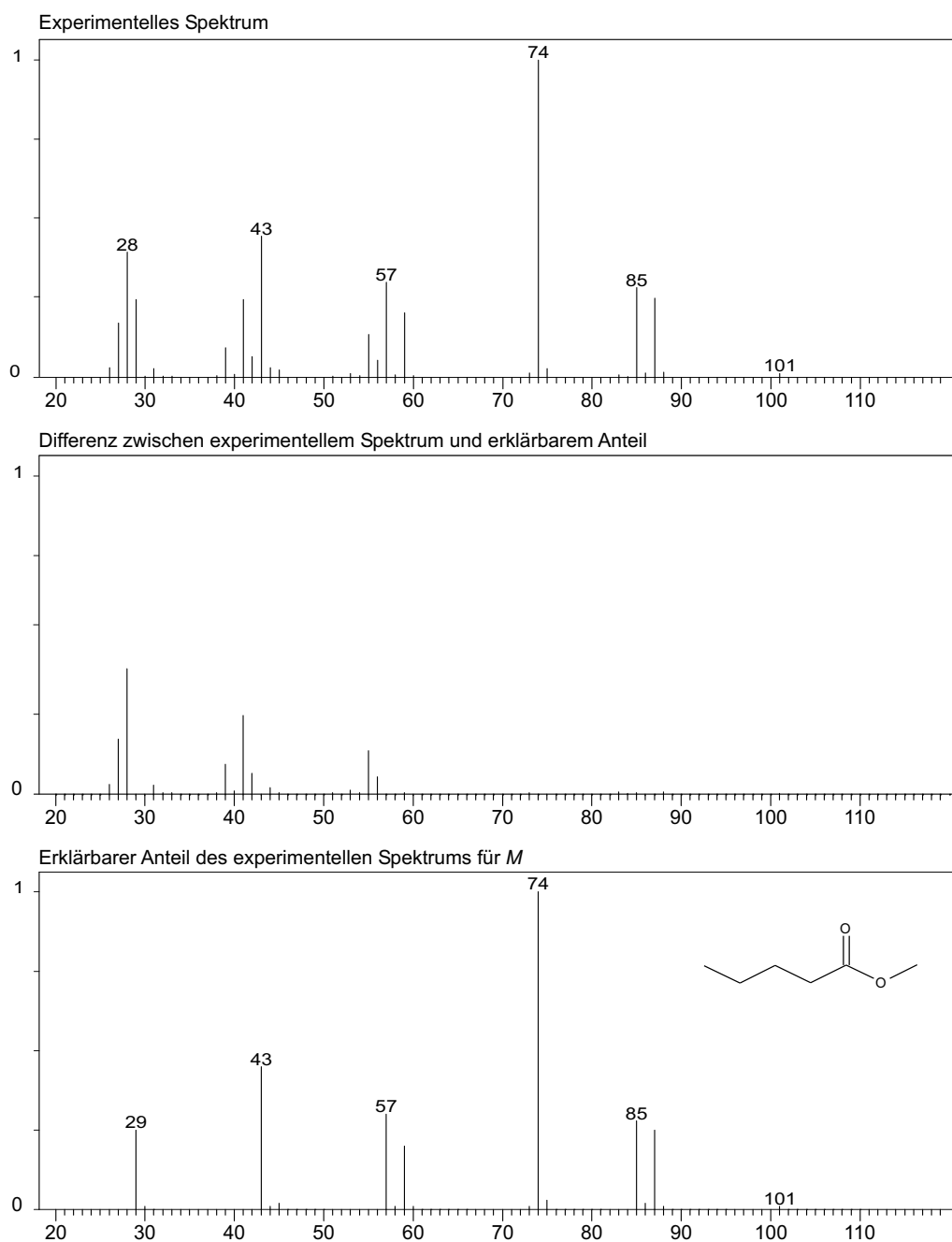


Abbildung 5.17: Gemessenes Spektrum und Anteil erklärbarer Intensität im Vergleich

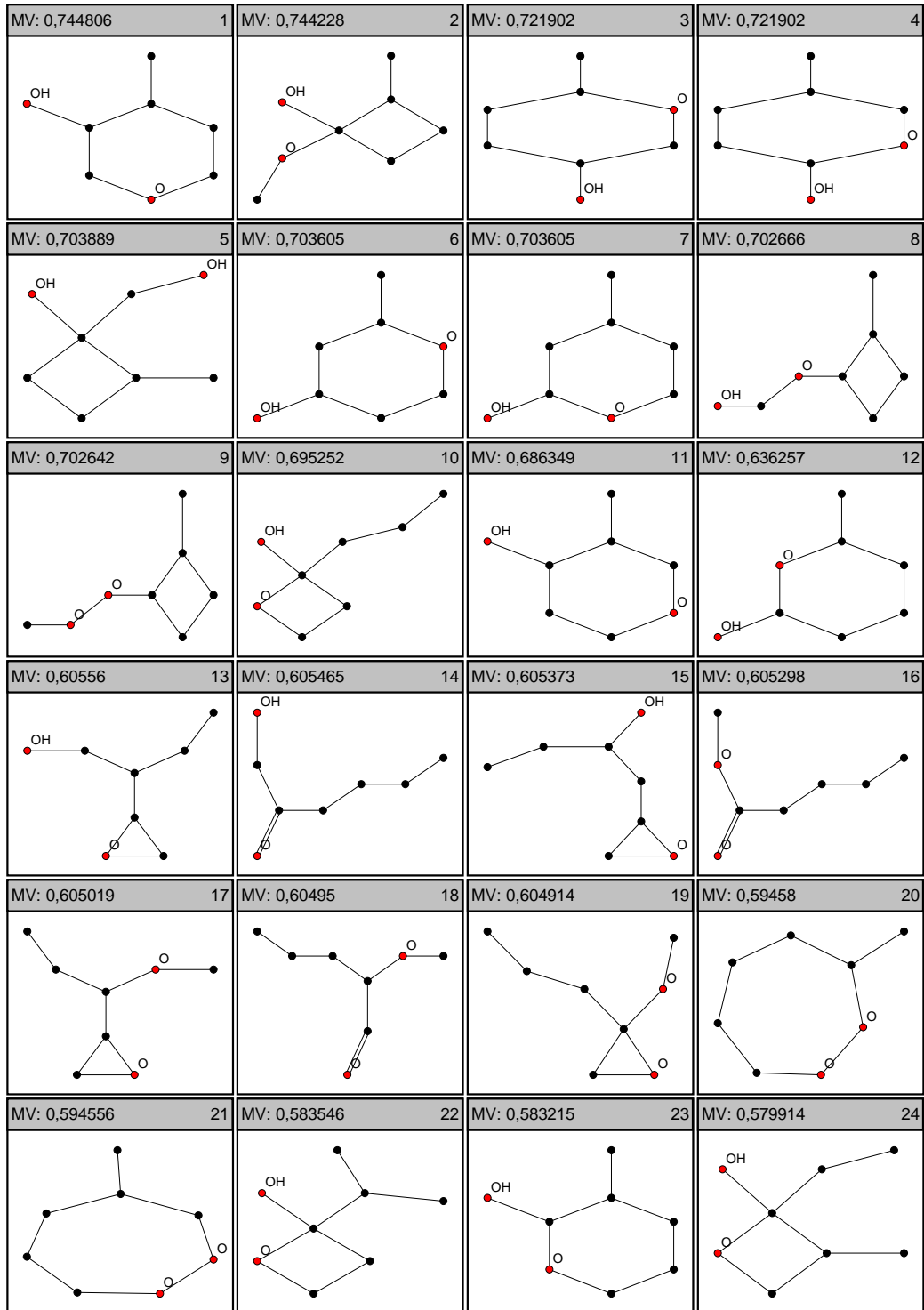


Abbildung 5.18: Ranking von  $C_6H_{12}O_2$  Isomeren bzgl. des Spektrums aus Beispiel 5.3.2



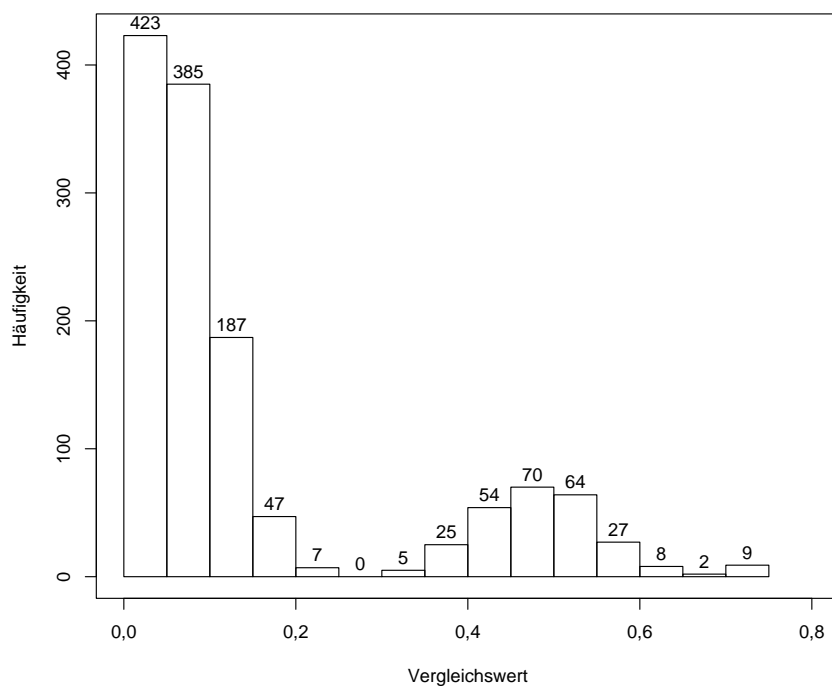


Abbildung 5.19: Histogramm der Strukturformel-Vergleichswerte für die Konstitutionsisomere von  $C_6H_{12}O_2$

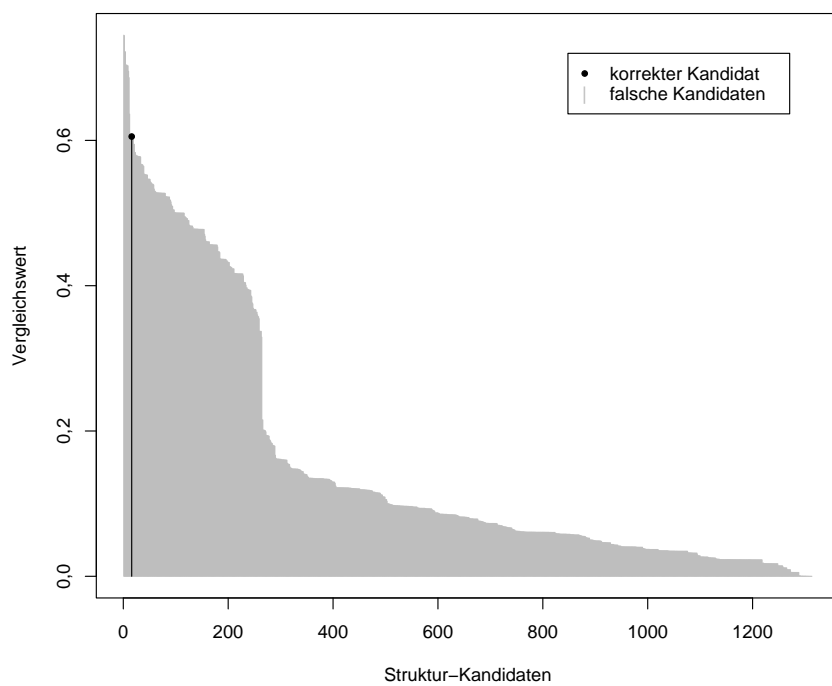


Abbildung 5.20: Verteilung der Strukturformel-Vergleichswerte für die Konstitutionsisomere von  $C_6H_{12}O_2$

1313 Kandidaten. Ordnen wir diese nach abfallenden Vergleichswerten an, so finden wir den korrekten Kandidaten *n*-Pentansäuremethylester auf Rang 16. Abbildung 5.18 zeigt die 24 bestplatzierten Kandidaten zusammen mit ihren Vergleichswerten. Auf den ersten 13 Plätzen sehen wir zyklische Strukturen. Dabei ist das Verhältnis von zyklischen und azyklischen Strukturen bei den Isomeren von  $C_6H_{12}O_2$  in etwa ausgeglichen (641 azyklische, 672 zyklische). Gäbe es eine Möglichkeit, anhand des Spektrums festzustellen, dass es sich um eine azyklische Struktur handelt, würde sich der korrekte Kandidat auf Platz 2 in dem Ranking verbessern. In Abschnitt 5.5 werden wir versuchen, derartige Kriterien für strukturelle Eigenschaften auf empirische Weise zu finden.

Abbildungen 5.19 und 5.20 zeigen ein Histogramm und die Verteilung der Vergleichswerte aller Konstitutionsisomere von  $C_6H_{12}O_2$ . Wir sehen, dass für dieses Beispiel unsere Vergleichswerte zumindest gut geeignet ist, um einen Großteil der Kandidaten auszuschließen. Man könnte hier eine Kandidaten-Selektion anhand der Verteilung der Vergleichswerte vornehmen und etwa alle Isomere mit Vergleichswerte kleiner 0,3 als irrelevant betrachten. Eine andere Vorgehensweise richtet sich an Erfahrungswerte, die wir aus einem Datensatz aufgeklärter Spektren gewinnen.

### Selektion relevanter Strukturformel-Kandidaten

Wie schon im Fall von Bruttoformeln berechnen wir zunächst wieder für eine zufällige Stichprobe von  $n = 1000$  Spektren  $I_i$  die Vergleichswerte der zugehörigen Strukturen  $M_i$ . Abbildungen 5.21 und 5.22 zeigen ein Histogramm und die Verteilung der Vergleichswerte  $x_i := MV(I_i, M_i)$ .

Erwartungsgemäß sind die Vergleichswerte generell deutlich kleiner als bei Bruttoformeln. Zumindest in den Fällen mit Werten nahe Null liegt aber der Verdacht nahe, dass die Datenbank-Spektren entweder von schlechter Qualität, oder die angegebenen Strukturen schlichtweg falsch sind.

Wie gehabt bestimmen wir für  $p = \frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}$   $p$ -Quantile von  $(x_i)_{i \in n}$  (s. Tabelle 5.9).

#### 5.4.8 Beispiel:

Wir wollen unser Verfahren zum Ranking und zur Kandidaten-Selektion wiederum an einem größeren, per Zufall ausgewählten Datensatz von 100 MS testen. Dabei berücksichtigen wir jedoch nur solche Spektren, deren zugehörige Molekülmasse höchstens 200 ist und zu deren korrekten Bruttoformel maximal 10000 Konstitutionsisomere existieren. Wir generieren alle Konstitutionsisomere, ermitteln deren Vergleichswerte und nehmen ein Ranking vor. Folgende Tabelle und Abbildung 5.23 fassen die Ergebnisse für die relativen

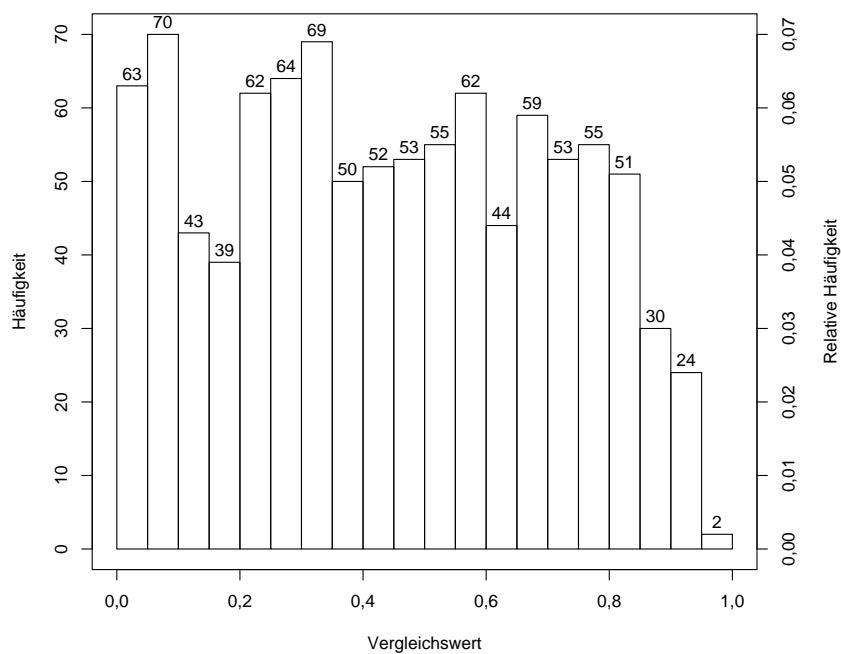


Abbildung 5.21: Histogramm der Strukturformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren

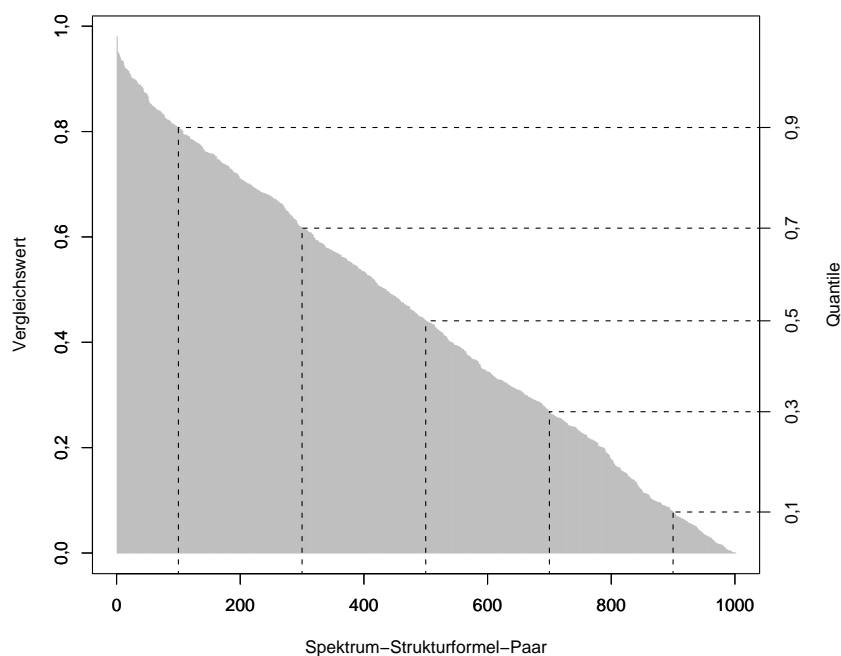


Abbildung 5.22: Verteilung der Strukturformel-Vergleichswerte für eine Stichprobe von 1000 Massenspektren

$p$	$q_p$	$p$	$q_p$	$p$	$q_p$
0,01	0,0045723	0,10	0,0777678	0,91	0,8142182
0,02	0,0142139	0,20	0,1777486	0,92	0,8224115
0,03	0,0182292	0,30	0,2680536	0,93	0,8381855
0,04	0,0288057	0,40	0,3435098	0,94	0,8462994
0,05	0,0348364	0,50	0,4405285	0,95	0,8664180
0,06	0,0464190	0,60	0,5335016	0,96	0,8845612
0,07	0,0545087	0,70	0,6163113	0,97	0,8975282
0,08	0,0615875	0,80	0,7099822	0,98	0,9104603
0,09	0,0679465	0,90	0,8073853	0,99	0,9326578

Tabelle 5.9: Quantile  $q_p$  für Strukturformel–Vergleichswerte zu verschiedenen Wahrscheinlichkeiten  $p$

Ranking–Positionen zusammen:

Min.	1.Quart.	Median	Mittel	3.Quart.	Max.
0,00000	0,07438	0,19210	0,29910	0,50000	1,00000

In Abbildung 5.24 sieht man die Anzahl ausgewählter Kandidaten bei Verlässlichkeit 90% im Bezug zur Position des korrekten Kandidaten im Ranking. Punkte oberhalb der Diagonalen repräsentieren Fälle, bei denen der korrekte Kandidat nicht ausgewählt würde. Für andere Verlässlichkeiten  $r$  zeigt folgende Tabelle die Anzahl der Fälle, bei denen sich die korrekte Strukturformel in der Menge der ausgewählten Kandidaten befindet.

$r = 0,99$	$r = 0,95$	$r = 0,90$	$r = 0,75$	$r = 0,50$
99	96	91	75	54

Im obigen Beispiel haben wir nur solche Fälle betrachtet, bei denen die Anzahl möglicher Konstitutionsisomere höchstens 10000 ist. In der Regel bilden solche Fälle eher die Ausnahme (vgl. Anhang E). Schon für kleinen Molekülmassen gibt es Summenformeln, zu denen deutlich mehr Isomere existieren. Bereits bei Massenzahlen um 200 gibt es Summenformeln, für die Anzahl von Isomeren mehrere Milliarden übersteigt, und somit selbst für äußerst effiziente Algorithmen eine vollständige Strukturgenerierung in vertretbarer Zeit nicht mehr möglich ist. Erschwerend kommt hinzu, dass Molekülmasse 200 eher die untere Grenze typischer Analyten für die Massenspektrometrie darstellt (vgl. Abbildung 5.6 und 5.7). Es ist deshalb unbedingt notwendig, bereits vor der Strukturgenerierung den Strukturraum sinnvoll einzugrenzen. Mit dieser Problematik wollen wir uns im nächsten Abschnitt auseinandersetzen.

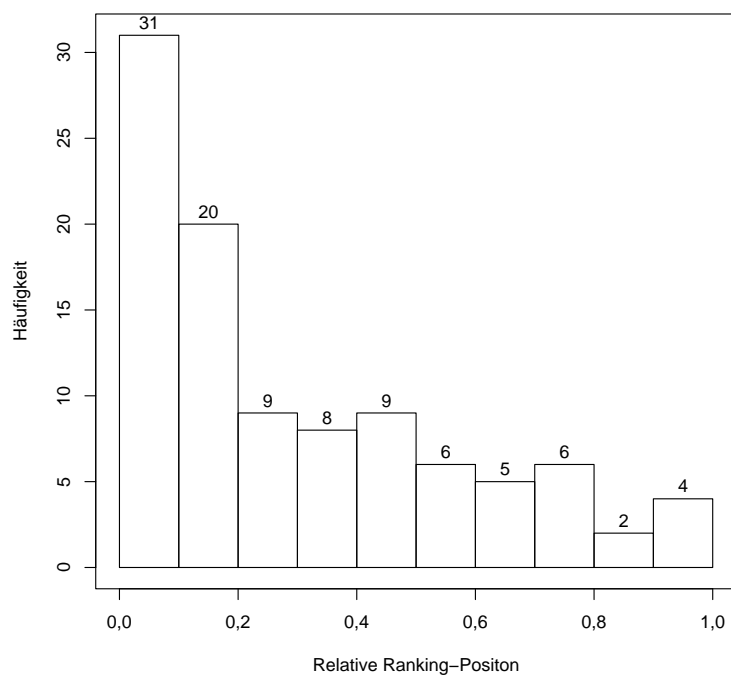


Abbildung 5.23: Histogramm der RRP für Strukturformeln von 100 Massenspektren

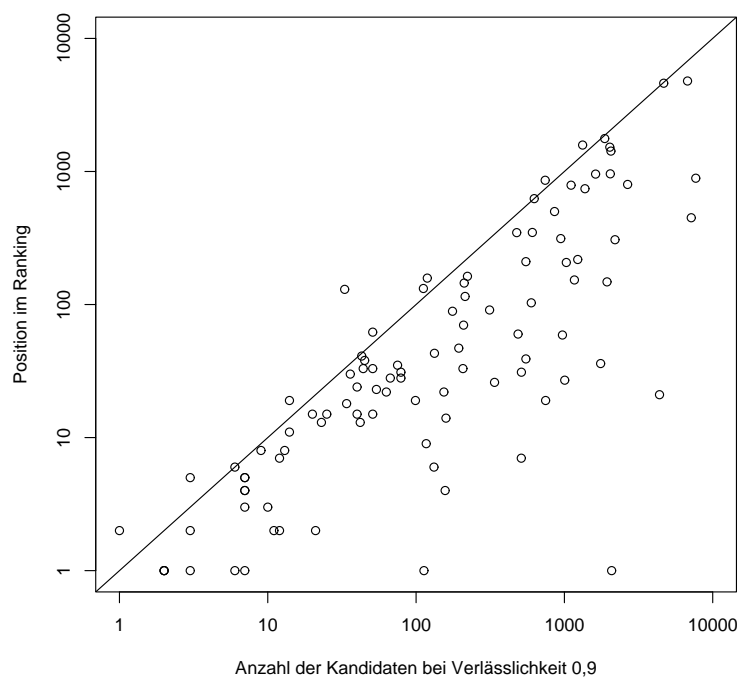


Abbildung 5.24: Ranking-Position und Kandidaten-Anzahl bei Verlässlichkeit 0,9 für Strukturformeln

## 5.5 Klassifikation von Massenspektren

Für die statistischen Untersuchungen im letzten Abschnitt haben wir uns auf Massenspektren beschränkt, zu denen die Strukturräume bei vorgegebener korrekter Bruttoformel eine Größe von 10000 Konstitutionsisomeren nicht übersteigen. Tatsächlich werden diese Fälle eher die Ausnahme bilden (vgl. Anhang E). Um die Strukturräume einzuschränken ist es wichtig, schon *vor* der Strukturgenerierung *strukturelle Eigenschaften* (engl. *Structural Properties*, kurz *SP*) des Analyten bestimmen zu können, die im günstigsten Fall bereits *während* der Strukturgenerierung berücksichtigt werden. Eine Möglichkeit, um aus Massenspektren Informationen über vorhandene und abwesende SP zu gewinnen, bieten *MS-Klassifikatoren*.

Ein MS-Klassifikator  $\Phi_S$  zu der binären strukturellen Eigenschaft  $S$  ist eine Abbildung

$$\Phi_S : \mathcal{I} \longrightarrow \mathbb{B},$$

die einem Massenspektrum  $I$  eine Klasse  $b \in \mathbb{B}$  zuweist. In der Regel wird  $S$  durch eine molekulare Substruktur definiert, und ist *wahr*, falls ein molekularer Graph diese Substruktur besitzt, anderenfalls *falsch*. Allgemeiner kann  $S$  aber ein beliebiger binärer molekularer Deskriptor sein.

Abbildung 5.25 zeigt die Vorgehensweise zur Berechnung und Anwendung eines MS-Klassifikators. Dabei sei bemerkt, dass dieses Prinzip nicht auf MS beschränkt ist. So wird beispielsweise in [105] die Konstruktion von IR-Klassifikatoren nach demselben Schema beschrieben.

Voraussetzung für die Konstruktion eines Spektren-Klassifikators ist eine Datenbasis aufgeklärter Spektren, in der hinreichend viele Strukturen mit und ohne der untersuchten strukturellen Eigenschaft  $S$  vorliegen. Das Vorhandensein von  $S$  bildet die Zielvariable für ein statistisches Lernverfahren zur Klassifikation.

Nun läge im Fall von Massenspektren der Gedanke nahe, die Intensitäten der Peaks als Vorhersagevariablen zu verwenden. Allerdings sind die Intensitäten selbst kaum mit strukturellen Eigenschaften in Beziehungen zu bringen. Stattdessen verwendet man *MS-Deskriptoren*, die besser geeignet sind, um MS-Struktur-Beziehungen zu modellieren.

Das Klassifikationsverfahren liefert eine Vorhersagefunktion, die verwendet werden kann, um für unbekannte Spektren zu entscheiden, ob  $S$  für die weitere Strukturaufklärung als *gegeben* oder *verboten* zu betrachten ist.

### 5.5.1 MS-Deskriptoren

Analog zur Vorgehensweise bei der Bestimmung von Struktur-Eigenschafts-Beziehungen werden Massenspektren durch MS-Deskriptoren auf reelle Zah-

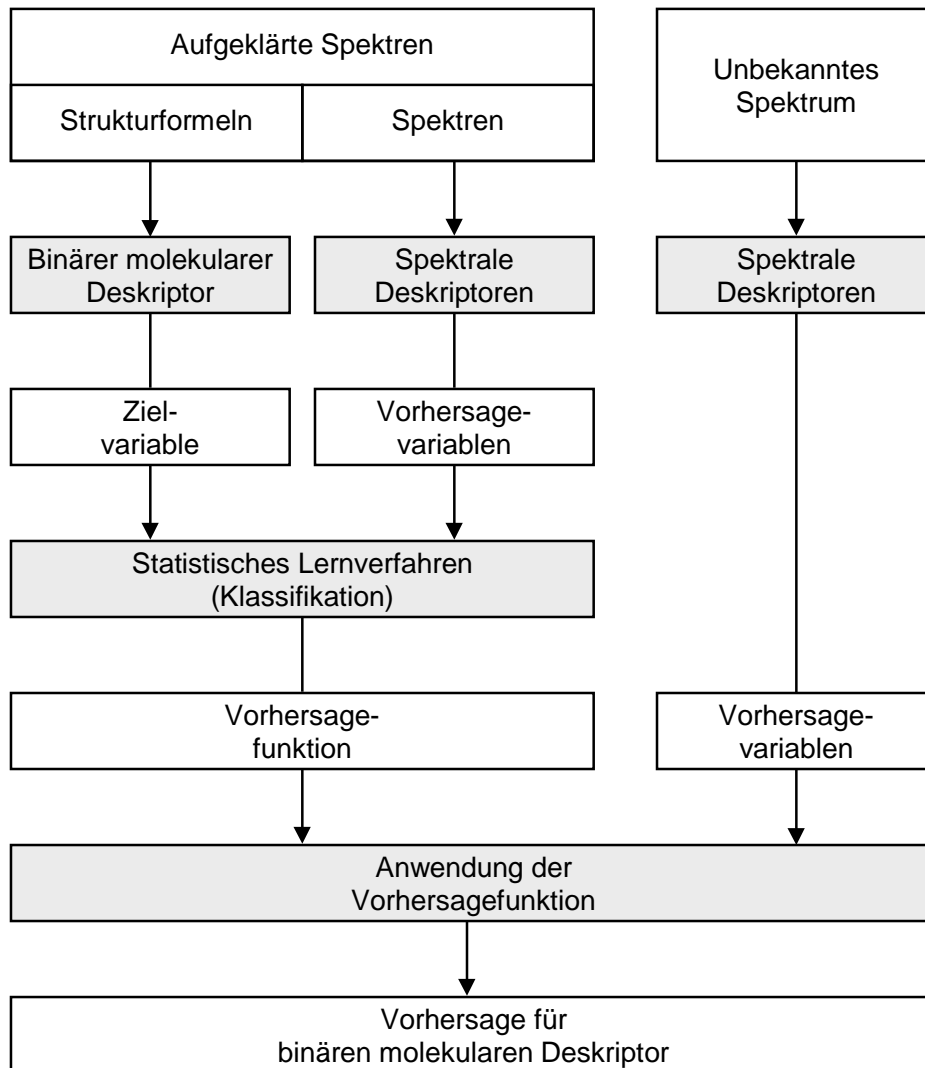


Abbildung 5.25: Vorgehensweise bei der Vorhersage struktureller Eigenschaften durch Spektren-Klassifikation

len abgebildet. Die so gewonnenen Werte sind geeignet, um Beziehungen zwischen Massenspektren und strukturellen Eigenschaften zu finden.

### 5.5.1 Definition:

Ein *MS-Deskriptor* ist eine Abbildung

$$D : \mathcal{I} \longrightarrow \mathbb{R}, \quad I \longmapsto D(I).$$

In der Literatur [154, 163] werden solche Abbildung auch *MS-Features* oder *MS-Invarianten* genannt. Wir wählen hier der einheitlichen Nomenklatur wegen die Bezeichnung *MS-Deskriptor*. Im Folgenden werden einige wichtige MS-Deskriptoren vorgestellt.

### 5.5.2 Beispiele:

*Ionenserien-Deskriptoren* summieren Intensitäten in Abständen von je 14 Masseneinheiten. Dieser Abstand entspricht gerade der Massendifferenz, die durch Abspaltung einer  $\text{CH}_2$ -Gruppe auftritt. Für  $r \in \underline{14}$  sind die Modulo-14-Deskriptoren definiert durch

$$\text{MD14}_r(I) = \sum_i I'(14i + r),$$

wobei nur Massen über 38 berücksichtigt werden:

$$I'(m) := \begin{cases} I(m) & \text{falls } m \geq 39, \\ 0 & \text{sonst.} \end{cases}$$

*Autokorrelations-Deskriptoren* beschreiben die Massenunterschiede in einem Spektrum. Für Massendifferenzen  $d > 0$  definiert man

$$\text{AUCO}_d(I) = \frac{\sum_m I(m)I(m+d)}{\sum_m I(m)^2}$$

sowie eingeschränkt auf die untere bzw. obere Hälfte des Spektrums

$$\begin{aligned} \text{ACLH}_d(I) &= \frac{\sum_{m \leq \hat{m}/2} I(m)I(m+d)}{\sum_m I(m)^2}, \\ \text{ACUH}_d(I) &= \frac{\sum_{m \geq \hat{m}/2} I(m)I(m+d)}{\sum_m I(m)^2}. \end{aligned}$$

*Logarithmische Intensitätsverhältnisse* werden für Paare aus Massen  $m$  und Massendifferenzen  $d > 0$  berechnet:

$$\text{LIQN}_{m,d}(I) = \ln \frac{I'(m)}{I'(m+d)},$$



wobei Intensitäten, die kleiner als 0,01 sind, angehoben werden:

$$I'(m) := \begin{cases} I(m) & \text{falls } I(m) \geq 0,01, \\ 0,01 & \text{sonst.} \end{cases}$$

*Spektrrentyp-Deskriptoren* beschreiben die Form des Massenspektrums, also beispielsweise die Verteilung der Peaks oder die Symmetrie. Der *Schwerpunkt* des Massenspektrums ist definiert als

$$\text{CENT}(I) = \frac{10^5}{\hat{m}} \sum_m m I(m).$$

Die Symmetrie des Spektrums bezüglich der Masse  $m$  lässt sich durch die *Symmetriefunktion*

$$\text{sym}_I(m) = \sum_{d=0}^{\hat{m}-m} I(m-d)I(m+d)$$

bewerten. Die kleinste Masse, bei der die Symmetriefunktion ihr Maximum annimmt, findet Verwendung in

$$\text{SYMX}(I) = \frac{1}{\hat{m}} \min\{m' \mid \forall m : \text{sym}_I(m') \geq \text{sym}_I(m)\}.$$

Weitere MS-Deskriptoren werden über den Basispeak definiert:

$$\begin{aligned} \text{MBAS}(I) &= 100 \cdot \frac{\tilde{m}}{\hat{m}}, \\ \text{BASE}(I) &= 100 \cdot \frac{I(\tilde{m})}{\sum_m I(m)}. \end{aligned}$$

Der Anteil *kleiner* Fragmente wird beschrieben durch

$$\text{DUST}(I) = 100 \cdot \frac{\sum_{m=1}^{78} I(m)}{\sum_{m=1}^{\hat{m}} I(m)},$$

den Intensitätsanteil von Peaks *geradzahlig*er Massen an der Gesamtintensität berechnet

$$\text{EVEN}(I) = 100 \cdot \frac{\sum_i I(2i)}{\sum_m I(m)}.$$

Ein weiterer Deskriptor PN10 liefert die Anzahl der *wichtigen* Peaks im Spektrum. Dies sind Peaks, deren Intensität größer als 10% der Basisintensität beträgt. Ist ein solcher Peak bei Masse  $m$  gefunden, so werden die Peaks bei

$m + 1$  und  $m + 2$  nicht gezählt. Dies verfolgt den Zweck, Isotopenpeaks hoher Intensität zu vernachlässigen. Damit ist dieser Deskriptor jedoch abhängig von der Richtung, in der das Spektrum durchlaufen wird, und ist somit keine Spektren-Invariante im engeren Sinn. Gleiches trifft auf SYMX zu.

Eine Vielzahl weiterer Deskriptoren wurden bei der Entwicklung der Software *MSclass* [154] zur Klassifikation von Massenspektren getestet. Schließlich fanden in den 160 Klassifikatoren aus *MSclass* 32 Deskriptoren mit insgesamt 431 Parameter-Kombinationen Verwendung.

In den nächsten beiden Abschnitten werden wir verschiedene Verfahren zur Klassifikation für bereits früher behandelte sowie zuvor nicht untersuchte strukturelle Eigenschaften vergleichen.

### 5.5.2 MS-Klassifikatoren

Gegeben sei eine Bibliothek von Massenspektren mit den zugehörigen Verbindungen, dargestellt durch ihre molekularen Graphen. Zur Konstruktion eines MS-Klassifikators  $\Phi_S$  für die strukturelle Eigenschaft  $S$  wählt man aus dieser Bibliothek  $m^W$  MS-Struktur-Paare, die diese Eigenschaft aufweisen und  $m^F$  Paare, bei denen  $S$  nicht vorliegt. Ausgangspunkt für die Konstruktion unseres MS-Klassifikators sind also Tupel

$$(I_i, y_i) \in \mathcal{I} \times \mathbb{B}, \quad i \in m := m^W + m^F,$$

wobei  $y_i = \text{wahr}$ , falls  $S$  bei der zu  $I_i$  gehörigen Struktur vorliegt,  $y_i = \text{falsch}$  anderenfalls. Gesucht wird eine Funktion

$$\Phi_S : \mathcal{I} \longrightarrow \mathbb{B},$$

die unsere MS-Struktur-Beziehung mathematisch beschreibt. Wie man  $\Phi_S$  bestimmt, wurde bereits zu Beginn von Abschnitt 5.5 und in Kapitel 3 erläutert. Typischerweise ist  $\Phi_S$  aus mehreren, nacheinander auszuführenden Abbildungen zusammengesetzt:

- Zuerst werden Massenspektren durch MS-Deskriptoren  $\mathcal{D} = (D_i)_{i \in n}$  auf reelle Zahlen abgebildet:

$$\mathcal{D} : \mathcal{I} \longrightarrow \mathbb{R}^n, \quad I \longmapsto (D_i(I))_{i \in n}.$$

- Transformationen der Deskriptorenwerte

$$\tau = (\tau_i)_{i \in n} : \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

die zum Trainieren der Vorhersagefunktion notwendig oder hilfreich waren, müssen durchgeführt werden.

- Die eigentliche Vorhersagefunktion  $f : \mathbb{R}^n \rightarrow \mathbb{B}$ , die aus einem statistischen Lernverfahren gewonnen wurde, wird angewendet.

Zusammenfassend können wir den MS-Klassifikator schreiben als Komposition

$$\Phi_S = f \circ \tau \circ \mathcal{D}.$$

Als Klassifikationsverfahren wurden in einer früheren Arbeit [163] LDA, KNN, SIMCA und ANN getestet und verglichen. Dabei erwiesen sich ANN und LDA als besser geeignete Verfahren. Wir werden im Folgenden zunächst Klassifikatoren durch CART und LDA berechnen, und dann mit SVM und ANN vergleichen.

### Klassifikation durch Entscheidungsbäume

In einer Grundmenge von 86052 Paaren aus Strukturen und zugehörigen Massenspektren der *NIST*-Massenspektren-Bibliothek (Abschnitt 5.3.5) wurde zunächst nach strukturellen Eigenschaften gemäß Anhang C gesucht. Für insgesamt 77 solcher Eigenschaften wurden je mindestens 300 Datensätze gefunden, die diese Eigenschaft besitzen und mindestens 300, bei denen sie nicht vorliegt. Per Zufall wurden nun disjunkte Lern- und Testsätze ausgewählt. Beide enthalten je 150 Spektren, deren Strukturen die untersuchte strukturelle Eigenschaft aufweisen sowie 150, bei denen die Eigenschaft nicht vorliegt.

Für jedes dieser so ausgewählten Spektren wurden 445 MS-Deskriptoren gemäß [163] berechnet:

- MD14<sub>r</sub>,  $r = 1, \dots, 14$ ,
- AUCO<sub>d</sub>, ACLH<sub>d</sub>, ACUH<sub>d</sub>,  $d = 1, \dots, 50$ ,
- LIQN<sub>m,d</sub>,  $m = 39, \dots, 175$ ,  $d = 1, 2$ ,
- CENT, SYMX, MBAS, BASE, DUST, EVEN, PN10.

Sie sind die Vorhersagevariablen für unser Klassifikationsverfahren. Zielvariable ist die Klassenzugehörigkeit, also *wahr*, falls *S* vorhanden, *falsch* anderenfalls.

#### 5.5.3 Beispiel:

Zunächst konstruieren wir einen Klassifikationsbaum zur Erkennung der Substruktur Methylester (s. Anhang C.5). Wir verwenden dabei die in der *R*-Schnittstelle festgelegten Standardparameter (`mincut = 5`, `minsize = 10`, `mindev = 0,01`).

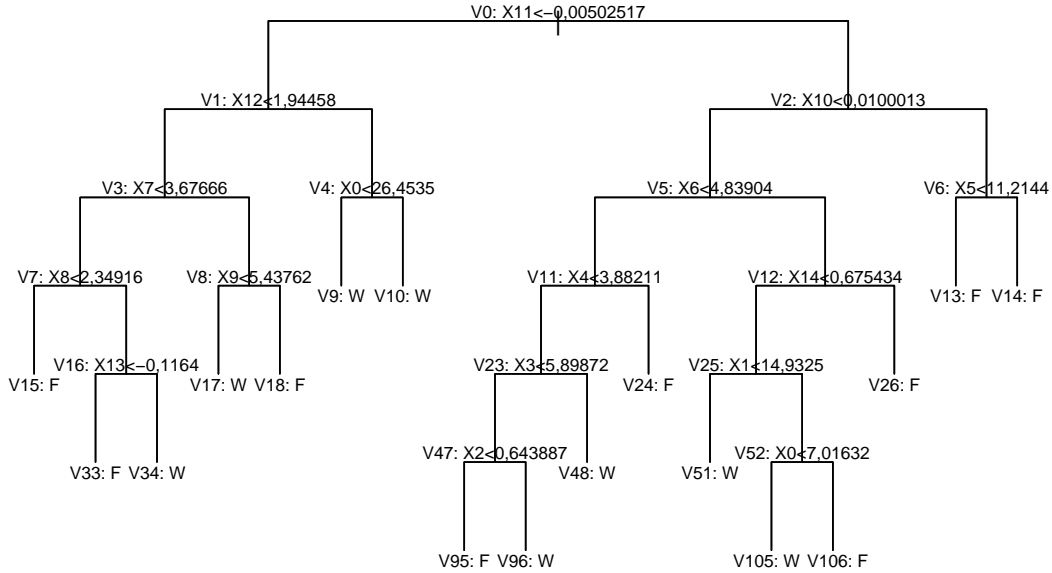


Abbildung 5.26: Klassifikationsbaum für Methylester

Der resultierende CT (Abbildung 5.26) verwendet 15 Deskriptoren:

$$\begin{aligned}
 X_0 &= \text{MD14}_1, & X_1 &= \text{AC LH}_3, & X_2 &= \text{ACUH}_3, \\
 X_3 &= \text{AC LH}_{26}, & X_4 &= \text{AUCO}_{29}, & X_5 &= \text{AUCO}_{31}, \\
 X_6 &= \text{AUCO}_{32}, & X_7 &= \text{ACUH}_{32}, & X_8 &= \text{AC LH}_{39}, \\
 X_9 &= \text{ACUH}_{46}, & X_{10} &= \text{LIQN}_{51,1}, & X_{11} &= \text{LIQN}_{58,1}, \\
 X_{12} &= \text{LIQN}_{59,2}, & X_{13} &= \text{LIQN}_{74,2}, & X_{14} &= \text{LIQN}_{99,1}.
 \end{aligned}$$

Er hat 16 innere und 17 terminale Knoten. Die Knoten sind gemäß Abschnitt 3.2.4 nummeriert. Tabelle 5.10 zeigt die Belegung der inneren Knoten  $V_i$  mit Entscheidungsregeln  $X_{j_i} < a_i$  sowie die Rückgabewerte der terminalen Knoten. Zudem ist in den Spalten  $m_i$ ,  $m_i^W$ ,  $m_i^F$  angegeben, wieviele Beobachtungen insgesamt und aus den einzelnen Klassen bei der Konstruktion in den jeweiligen Knoten verarbeitet werden.

Dieser Tabelle kann man die Fehlklassifikationen für den Lernsatz entnehmen: Für den Fehler erster Art (als *falsch* klassifizierte *wahre* Beobachtungen) summieren wir  $1 + 1 + 2 + 1 + 2 + 1 + 3 = 11$  Beobachtungen, für den Fehler zweiter Art  $1 + 3 = 4$  Beobachtungen. Dies ergibt eine Missklassifikationsrate von  $MCE_{LS} = \frac{11}{300} + \frac{4}{300} = 0,05$ . Erwartungsgemäß liegt die Missklassifikationsrate für den Testsatz  $MCE_{TS} = \frac{35}{300} + \frac{38}{300} = \frac{77}{300} = 0,25667$  deutlich höher.

Bei einem genaueren Blick auf den CT erkennen wir zwei Knoten,  $V_4$  und  $V_6$ , deren Nachfolger terminale Knoten sind, und jeweils die gleiche Antwort

$i$	$j_i$	$X_{j_i}$	$a_i$	$m_i$	$m_i^W$	$m_i^F$
0	11	LIQN <sub>58,1</sub>	-0,00502517	300	150	150
1	12	LIQN <sub>59,2</sub>	1,94458	126	105	21
3	7	ACUH <sub>32</sub>	3,67666	59	39	20
7	8	AC LH <sub>39</sub>	2,34916	29	12	17
15		terminal	<i>falsch</i>	13	1	12
16	13	LIQN <sub>74,2</sub>	-0,1164	16	11	5
33		terminal	<i>falsch</i>	6	1	5
34		terminal	<i>wahr</i>	10	10	0
8	9	ACUH <sub>46</sub>	5,43762	30	27	3
17		terminal	<i>wahr</i>	25	25	0
18		terminal	<i>falsch</i>	5	2	3
4	0	MD14 <sub>1</sub>	26,4535	67	66	1
9		terminal	<i>wahr</i>	62	62	0
10		terminal	<i>wahr</i>	5	4	1
2	10	LIQN <sub>51,1</sub>	0,0100013	174	45	129
5	6	AUCO <sub>32</sub>	4,83904	97	42	55
11	4	AUCO <sub>29</sub>	3,88211	59	15	44
23	3	AC LH <sub>26</sub>	5,89872	27	14	13
47	2	ACUH <sub>3</sub>	0,643887	18	5	13
95		terminal	<i>falsch</i>	10	0	10
96		terminal	<i>wahr</i>	8	5	3
48		terminal	<i>wahr</i>	9	9	0
24		terminal	<i>falsch</i>	32	1	31
12	14	LIQN <sub>99,1</sub>	0,675434	38	27	11
25	1	AC LH <sub>3</sub>	14,9325	29	26	3
51		terminal	<i>wahr</i>	18	18	0
52	0	MD14 <sub>1</sub>	7,01632	11	8	3
105		terminal	<i>wahr</i>	6	6	0
106		terminal	<i>falsch</i>	5	2	3
26		terminal	<i>falsch</i>	9	1	8
6	5	AUCO <sub>31</sub>	11,2144	77	3	74
13		terminal	<i>falsch</i>	69	0	69
14		terminal	<i>falsch</i>	8	3	5

Tabelle 5.10: Details der Knoten des Klassifikationsbaums für Methylester

$j_i$	$X_{j_i}$	$X_{j_i}(I)$	$j_i$	$X_{j_i}$	$X_{j_i}(I)$	$j_i$	$X_{j_i}$	$X_{j_i}(I)$
0	MD14 <sub>1</sub>	30,50969	5	AUCO <sub>31</sub>	30,90479	10	LIQN <sub>51,1</sub>	0,000000
1	AC LH <sub>3</sub>	8,214255	6	AUCO <sub>32</sub>	8,717663	11	LIQN <sub>58,1</sub>	-3,015535
2	ACUH <sub>3</sub>	1,733027	7	ACUH <sub>32</sub>	2,985316	12	LIQN <sub>59,2</sub>	3,015535
3	AC LH <sub>26</sub>	0,000000	8	AC LH <sub>39</sub>	0,000000	13	LIQN <sub>74,2</sub>	4,605170
4	AUCO <sub>29</sub>	10,24026	9	ACUH <sub>46</sub>	0,1361653	14	LIQN <sub>99,1</sub>	0,000000

Tabelle 5.11: Deskriptorenwerte für das Spektrum aus Beispiel 5.3.2

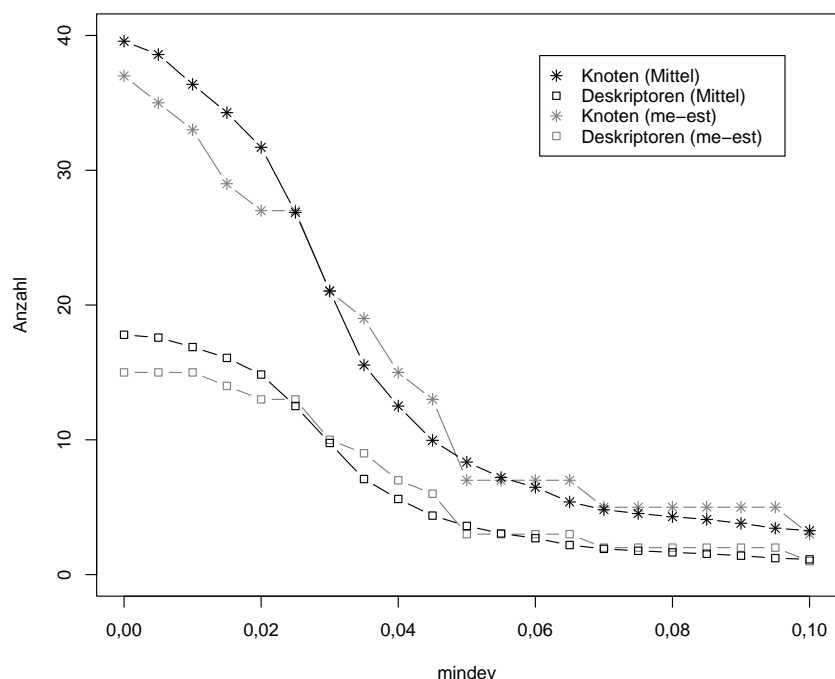


Abbildung 5.27: Komplexität der Klassifikationsbäume

liefern. Man könnte demnach  $V_4$  durch einen terminalen Knoten ersetzen, der *wahr* ausgibt, ebenso  $V_6$  durch einen terminalen Knoten, der *falsch* liefert. Dies kann im Zuge des *Lichtens* eines Entscheidungsbaumes durchgeführt werden. Andererseits enthält aber der ungerichtete Entscheidungsbaum Informationen, die für die Verlässlichkeit einer Vorhersage herangezogen werden können. So würde man einer Vorhersage durch  $V_{13}$  aufgrund der Fehlklassifikationen des Lernsatzes wohl eine höhere Verlässlichkeit einräumen als einer Vorhersage durch  $V_{14}$ .

Schließlich wollen wir den CT zur Vorhersage über das Vorhandensein von Methylester für das Spektrum aus Beispiel 5.3.2 heranziehen. Tabelle 5.11 enthält die benötigten Deskriptorenwerte. Bei der Anwendung des CT werden Knoten  $V_0$ ,  $V_1$  und  $V_4$  durchlaufen. Die Vorhersage terminiert in  $V_{10}$ . Wir erhalten *wahr* als Prognose. Das untersuchte Spektrum gehört tatsächlich zu einer Verbindung, die Methylesterals Substruktur enthält (vgl. Abschnitte 5.4.2 und 5.6.1).

Wir werden unsere weiteren Untersuchungen auf die bereits angesprochenen 77 strukturellen Eigenschaften gemäß Anhang C ausdehnen. Besonders interessiert uns dabei, inwiefern die Komplexität der Klassifikationsbäume ihre Vorhersagefähigkeit beeinflusst. Wir messen die Komplexität von CT in der Anzahl verwendeter Deskriptoren und der Anzahl von Knoten. Der CART-

Algorithmus kann über mehrere Parameter gesteuert werden, die sich auf die Komplexität der CT auswirken.

In der hier verwendeten Implementierung gibt es einen Parameter `mindev`, der vorschreibt, wie groß im Vergleich zur Wurzel die Abweichung innerhalb eines Knotens mindestens sein muss, damit eine weitere Aufspaltung erfolgt. Kleinere Werte für `mindev` werden also zu komplexeren CT führen. Wir konstruieren zu jeder der 77 strukturellen Eigenschaften Klassifikationsbäume mit  $\text{mindev} = k \cdot 0,005$  für  $k = 0, \dots, 20$ .

Abbildung 5.27 zeigt die arithmetischen Mittel der Anzahlen von Knoten und Deskriptoren in Abhängigkeit von `mindev`. Zudem sind die Deskriptoren- und Knotenanzahlen für CT zu Methylester in die Graphik aufgenommen. Es ist deutlich erkennbar, dass die Komplexität der Klassifikationsbäume mit wachsendem `mindev` abnimmt.

Abbildung 5.28 gibt Aufschluss über die Missklassifikationsraten der CT zu verschiedenen Werten für `mindev`. Dabei wurden arithmetische Mittel der MCE hinsichtlich Lern- und Testsatz für die 77 strukturellen Eigenschaften gebildet. Erwartungsgemäß nehmen die Missklassifikationsraten  $MCE_{LS}$  für den Lernsatz mit abnehmender Komplexität der CT zu. Die Missklassifikationsraten für den Testsatz sind deutlich niedriger als für den Lernsatz. Für  $MCE_{TS}$  erhält man Minima bei Klassifikationsbäumen mittlerer Komplexität. So ist für Klassifikationsbäume zu Methylester  $MCE_{TS}$  minimal bei  $\text{mindev} = 0,45$ . Das arithmetische Mittel der  $MCE_{TS}$  über alle 77 strukturellen Eigenschaften ist am kleinsten für  $\text{mindev} = 0,4$ .

Abbildung 5.29 zeigt nochmals das arithmetische Mittel der MCE sowie die Missklassifikationsraten beschränkt auf die beiden Klassen. Überraschend groß ist der Unterschied für die arithmetischen Mittel der Missklassifikationsraten  $MCE_{TS}^F$  und  $MCE_{TS}^W$ . In Tabelle 5.12 sind  $MCE_{TS}^F$ ,  $MCE_{TS}^W$  und  $MCE_{TS}$  für CT mit  $\text{mindev} = 0,04$  zu den 77 strukturellen Eigenschaften im Einzelnen aufgeführt.

### Klassifikation durch lineare Verfahren

Zum Vergleich wollen wir unter denselben Voraussetzungen (gleiche strukturelle Eigenschaften, gleiche Lern- und Testsätze, gleiche Grundmenge von Deskriptoren) MS-Klassifikatoren durch lineare Verfahren bestimmen.

Die Deskriptoren-Selektion nehmen wir mit einem  $l$ -fachen schrittweisen Verfahren vor. Wir führen das schrittweise Verfahren bis zu  $n = 30$  Deskriptoren durch und testen dabei verschiedene Parameter  $l \in \{5, 20, 50\}$ . Die Bestimmung der Modelle erfolgt mit Klassifikation durch (OLS-)Regression (Abschnitte 3.1.1 und 3.2.1). Zur Bewertung der Modelle in den einzelnen Schritten ziehen wir  $MCE_{LS}$  heran (vgl. Abschnitt 4.4.3).

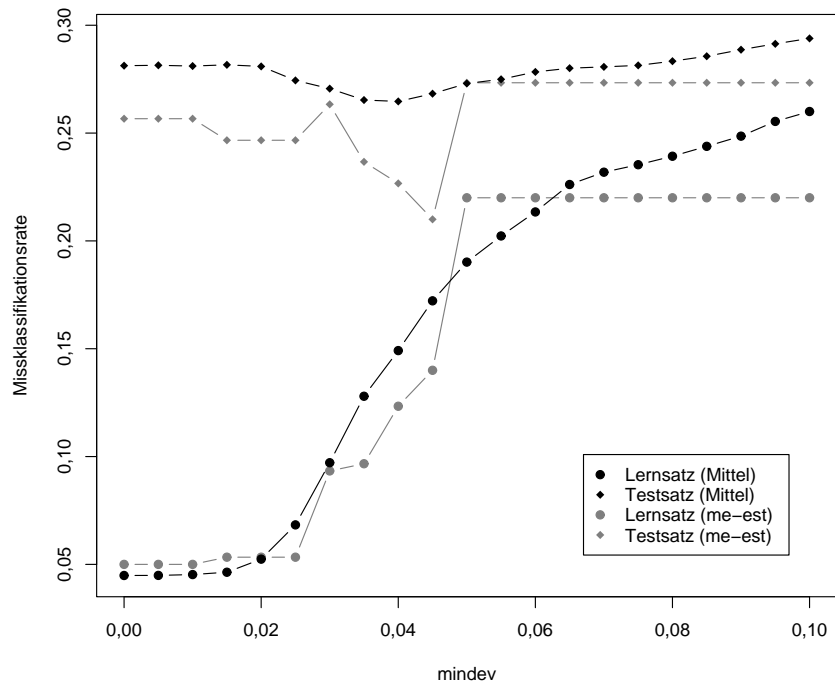


Abbildung 5.28: Mittlere Missklassifikationsraten für Lern- und Testsatz bei Klassifikation durch CT

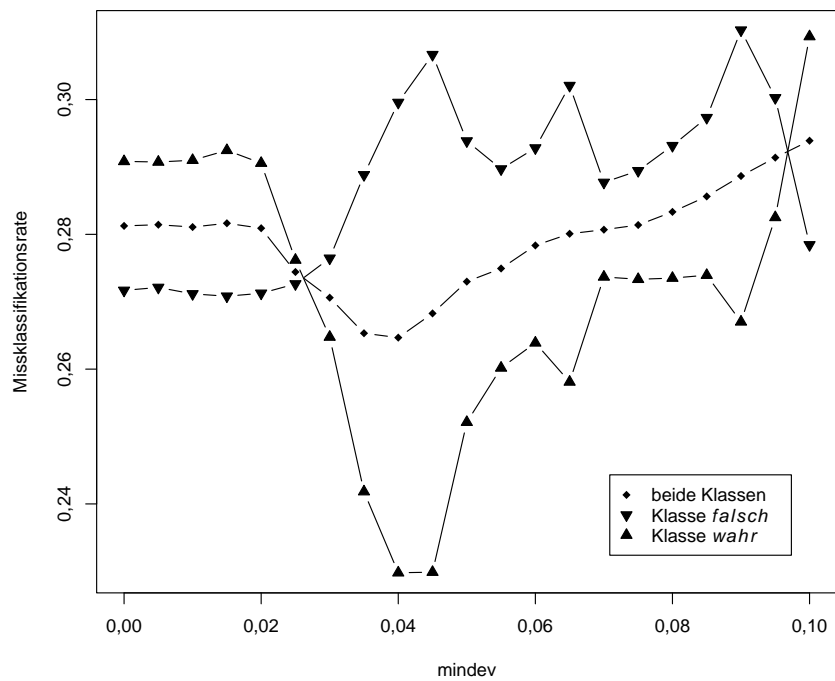


Abbildung 5.29: Mittlere Missklassifikationsraten für den Testsatz bzgl. der beiden Klassen bei Klassifikation durch CT



Name	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$	Name	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$
C quart ch	0,2733	0,4867	0,3800	benz-0	0,4400	0,1467	0,2933
C4 H9	0,3467	0,1467	0,2467	ar-COOCH2*	0,1733	0,3600	0,2667
C5 H11	0,3133	0,1800	0,2467	naph	0,3800	0,1133	0,2467
C5 H11 *	0,2533	0,5600	0,4067	phen-1-OH	0,3800	0,4000	0,3900
C6 H13	0,2200	0,2200	0,2200	phen	0,2867	0,4200	0,3533
C6 H13 n	0,1667	0,2867	0,2267	phen-2-OH	0,3467	0,3400	0,3433
C7 H15	0,1533	0,1333	0,1433	CO-C6H3-0	0,5133	0,1067	0,3100
C8 H17	0,1067	0,1000	0,1033	r>C=C<r	0,5067	0,2000	0,3533
C9 H19	0,2267	0,0800	0,1533	(CH3)2>C=C	0,3800	0,1400	0,2600
C10 H21	0,2067	0,1067	0,1567	Br	0,1800	0,1867	0,1833
C11 H23	0,1133	0,1533	0,1333	Cl	0,1200	0,2067	0,1633
hydr carb	0,3000	0,1533	0,2267	N	0,3800	0,2267	0,3033
(CH3)3-C	0,2400	0,1533	0,1967	N 2	0,5000	0,1667	0,3333
ar-CHO	0,2400	0,2067	0,2233	P	0,4000	0,2533	0,3267
ar-CO,N2	0,2400	0,5000	0,3700	Si	0,3600	0,2200	0,2900
ar-CH2	0,2400	0,4067	0,3233	Si 1	0,2800	0,2400	0,2600
ar-N,NHN	0,4200	0,3800	0,4000	Si >=2	0,1933	0,1667	0,1800
ar-C ch	0,4000	0,1067	0,2533	S	0,2733	0,2133	0,2433
ar-C r	0,1800	0,1867	0,1833	CH3-COO	0,3600	0,1400	0,2500
ar-CO	0,3467	0,2800	0,3133	CH3-CO	0,2400	0,1267	0,1833
ar-CH	0,4800	0,1600	0,3200	alc tert	0,4333	0,1467	0,2900
ar-CH2CH2	0,3667	0,3467	0,3567	am tert	0,5400	0,1867	0,3633
ar-Cl	0,0933	0,1333	0,1133	n-C4H9-0	0,1200	0,2067	0,1633
ar-CO-CH2	0,5533	0,1400	0,3467	C2H5-CO	0,2333	0,3400	0,2867
ar-COO	0,3000	0,3400	0,3200	CF3	0,1867	0,1667	0,1767
ar-F	0,1933	0,2467	0,2200	CF3-CO	0,0933	0,4133	0,2533
ar-N	0,4200	0,2067	0,3133	NH-CH2-CH2	0,3000	0,2600	0,2800
ar-N ch	0,4467	0,2733	0,3600	CH3-0-CH2	0,1467	0,0933	0,1200
ar-N r	0,4933	0,2067	0,3500	N(CH3)2	0,2200	0,1867	0,2033
ar-0	0,3467	0,3733	0,3600	CH3-COOCH	0,1667	0,2067	0,1867
ar-0-CH2	0,6000	0,0800	0,3400	et-est	0,4333	0,2400	0,3367
ar-0-CH3	0,2600	0,2733	0,2667	me-est	0,2467	0,2067	0,2267
ar-S r	0,3333	0,2600	0,2967	C2H5-0	0,7333	0,1667	0,4500
biphenyl	0,2067	0,2267	0,2167	(CH2)6-CO	0,2467	0,2467	0,2467
C6H4-Br	0,1467	0,1867	0,1667	NO	0,3933	0,3533	0,3733
C6H4 omp	0,2600	0,3067	0,2833	S-CH2	0,4200	0,2467	0,3333
ph-C	0,2933	0,1733	0,2333	(CH3)3 Si	0,0400	0,0667	0,0533
ph-CH2-0	0,0467	0,0933	0,0700	r 5+6	0,2933	0,4600	0,3767
ph	0,3000	0,2733	0,2867				

Tabelle 5.12: Missklassifikationsraten von MS-Klassifikatoren (CT) für 77 strukturelle Eigenschaften

Dies wirft weitere Probleme auf. Die Effizienz des  $l$ -fachen schrittweisen Verfahrens beruht darauf, dass in jedem Schritt nur die besten  $l$  Modelle ausgewählt und weiter verwendet werden. Treten mehrere Modelle mit gleichem MCE auf, so gibt es Situationen, in denen die besten  $l$  Modelle nicht eindeutig bestimmt werden können. Zunächst kann man sich behelfen, indem man die Selektion in dem betroffenen Schritt um all die Modelle vergrößert, welche gleiche MCE wie das  $l$ -te Modell haben. Da die MCE aufgrund der Größe des Lernsatzes nur 300 verschiedene Werte annehmen können, kommt es zu Fällen, wo diese Situation in mehreren aufeinander folgenden Schritten eintritt, und somit sehr große Mengen gleich guter Modelle berücksichtigt werden müssen. Wir schaffen dem Abhilfe, indem wir bei gleicher Missklassifikationsrate RSS als zusätzliches Kriterium heranziehen. Kleine RSS bedeuten, dass die Klassen gut separiert werden. Wir wählen deshalb  $MCE_{LS} + |LS|^{-1}RSS_{LS}$  als Auswahlkriterium.

Für die so gewonnen Deskriptorensätze verwenden wir LDA zur Klassifikation<sup>4</sup>. Abbildung 5.30 zeigt die arithmetischen Mittel der Missklassifikationsraten für Lern und Testsatz bei verschiedenen  $l$  in Abhängigkeit von der Anzahl verwendeter Deskriptoren. Wir sehen, dass für den Lernsatz die Werte der MCE mit zunehmender Deskriptorenanzahl streng monoton abnehmen, bei größerem  $l$  sind die arithmetischen Mittel für gleiche Anzahl von Deskriptoren kleiner. Letzteres gilt auch für den Testsatz. In Abhängigkeit der Deskriptorenanzahl haben wir auf dem Testsatz jedoch Minima im Bereich zwischen 10 und 15 Deskriptoren. Für größere Anzahlen von Deskriptoren wachsen die Werte tendenziell wieder an.

Genaueren Aufschluss über das Verhalten der Mittelwerte von  $MCE_{TS}$  für  $l = 50$  liefert Abbildung 5.31. Wir erkennen ein globales Minimum bei  $n = 13$  Deskriptoren. Zusätzlich sind die mittleren Missklassifikationsraten für die einzelnen Klassen eingezeichnet. Zwar sind die Werte für Klasse *falsch* meist höher als für Klasse *wahr*, jedoch sind die Abweichungen nicht so deutlich wie bei Klassifikationsbäumen.

Tabelle 5.12 zeigt  $MCE_{TS}^F$ ,  $MCE_{TS}^W$  und  $MCE_{TS}$  für die Modelle mit  $n = 13$  Deskriptoren im Einzelnen. In Abbildung 5.32 sind die  $MCE_{TS}$  der Klassifikationsbäume mit  $\text{mindev} = 0,04$  und der linearen Modelle mit  $n = 13$  Deskriptoren gegeneinander abgetragen. Datenpunkte oberhalb der Diagonalen stehen dabei für strukturelle Eigenschaften, die durch LM besser vorhergesagt werden als durch CT. Man sieht deutlich die Vorteile zugunsten der linearen Methode. Bei strukturellen Eigenschaften, für die entweder Mo-

---

<sup>4</sup>Unter den hier bestehenden Voraussetzungen — binäre Klassifikation mit gleich vielen Beobachtungen beider Klassen im Lernsatz — ist LDA identisch mit Klassifikation durch MLR.

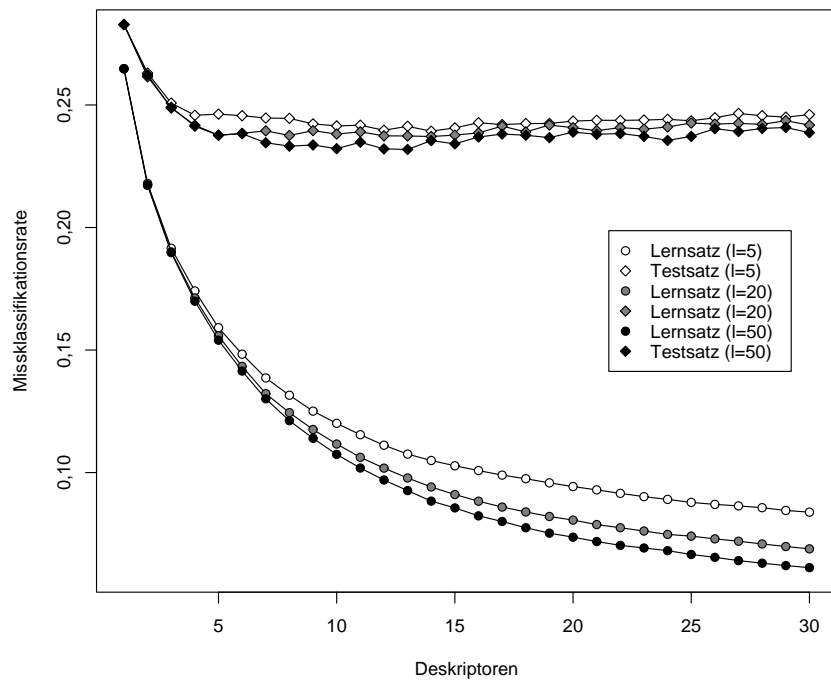


Abbildung 5.30: Mittlere Missklassifikationsraten für Lern- und Testsatz bei Klassifikation durch LDA

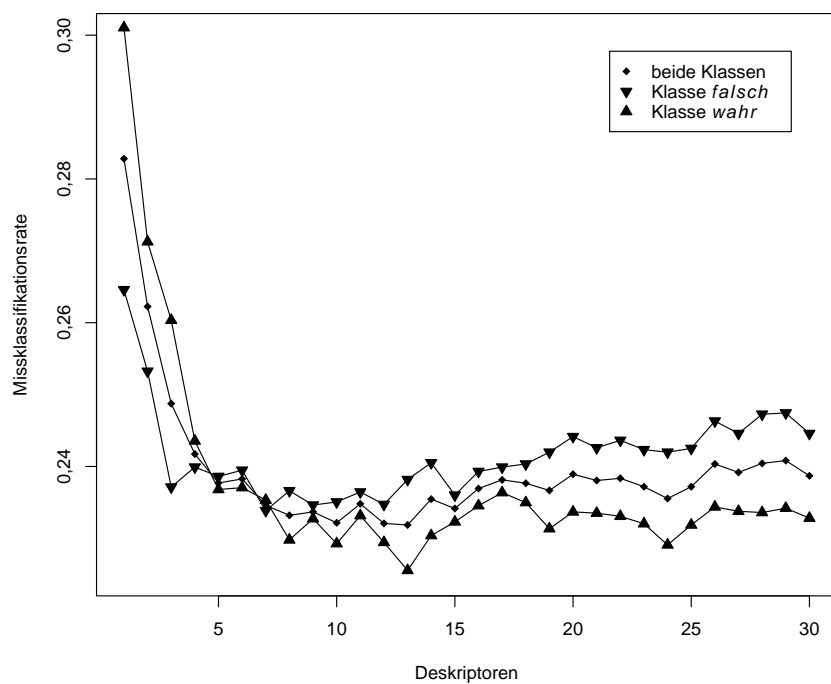


Abbildung 5.31: Mittlere Missklassifikationsraten für den Testsatz bzgl. der beiden Klassen bei Klassifikation durch LDA

Name	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$	Name	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$
C quart ch	0,3400	0,3933	0,3667	benz-O	0,2200	0,1933	0,2067
C4 H9	0,2467	0,2333	0,2400	ar-COOCH2*	0,1400	0,2067	0,1733
C5 H11	0,2933	0,1533	0,2233	naph	0,2933	0,0933	0,1933
C5 H11 *	0,4000	0,4067	0,4033	phen-1-OH	0,3200	0,3000	0,3100
C6 H13	0,2267	0,2800	0,2533	phen	0,2867	0,3333	0,3100
C6 H13 n	0,2333	0,1467	0,1900	phen-2-OH	0,2867	0,2867	0,2867
C7 H15	0,1533	0,1533	0,1533	CO-C6H3-O	0,3467	0,2533	0,3000
C8 H17	0,2067	0,0933	0,1500	r>C=C<r	0,1533	0,2867	0,2200
C9 H19	0,1800	0,0600	0,1200	(CH3)2>C=C	0,2733	0,2600	0,2667
C10 H21	0,2000	0,1133	0,1567	Br	0,1067	0,1867	0,1467
C11 H23	0,1600	0,0533	0,1067	Cl	0,0600	0,2400	0,1500
hydr carb	0,2600	0,1067	0,1833	N	0,3000	0,2333	0,2667
(CH3)3-C	0,1267	0,1533	0,1400	N 2	0,3533	0,3333	0,3433
ar-CHO	0,2533	0,2067	0,2300	P	0,2267	0,3133	0,2700
ar-CO,N2	0,3600	0,2267	0,2933	Si	0,1467	0,1200	0,1333
ar-CH2	0,2867	0,3467	0,3167	Si 1	0,1933	0,2600	0,2267
ar-N,NHN	0,3467	0,2267	0,2867	Si >=2	0,0933	0,0933	0,0933
ar-C ch	0,2600	0,1467	0,2033	S	0,2933	0,3133	0,3033
ar-C r	0,1667	0,1733	0,1700	CH3-COO	0,2733	0,2133	0,2433
ar-CO	0,3600	0,1600	0,2600	CH3-CO	0,1600	0,1933	0,1767
ar-CH	0,2867	0,2600	0,2733	alc tert	0,2533	0,3533	0,3033
ar-CH2CH2	0,2867	0,1933	0,2400	am tert	0,2800	0,3533	0,3167
ar-Cl	0,0867	0,1067	0,0967	n-C4H9-O	0,1533	0,2533	0,2033
ar-CO-CH2	0,3800	0,3267	0,3533	C2H5-CO	0,1733	0,2267	0,2000
ar-COO	0,3533	0,2333	0,2933	CF3	0,1400	0,1667	0,1533
ar-F	0,2200	0,1800	0,2000	CF3-CO	0,1267	0,2267	0,1767
ar-N	0,2000	0,2867	0,2433	NH-CH2-CH2	0,2200	0,3200	0,2700
ar-N ch	0,3067	0,2133	0,2600	CH3-O-CH2	0,1067	0,1467	0,1267
ar-N r	0,3200	0,2600	0,2900	N(CH3)2	0,2133	0,1733	0,1933
ar-O	0,3533	0,2933	0,3233	CH3-COOCH	0,2400	0,2200	0,2300
ar-O-CH2	0,3533	0,3467	0,3500	et-est	0,3467	0,3400	0,3433
ar-O-CH3	0,3267	0,2400	0,2833	me-est	0,2000	0,3400	0,2700
ar-S r	0,2867	0,2133	0,2500	C2H5-O	0,2600	0,3267	0,2933
biphenyl	0,2933	0,1400	0,2167	(CH2)6-CO	0,2000	0,2133	0,2067
C6H4-Br	0,0800	0,1400	0,1100	NO	0,3667	0,3733	0,3700
C6H4 omp	0,2600	0,1933	0,2267	S-CH2	0,1867	0,3800	0,2833
ph-C	0,2467	0,1467	0,1967	(CH3)3 Si	0,0533	0,0667	0,0600
ph-CH2-O	0,0733	0,0467	0,0600	r 5+6	0,3600	0,3267	0,3433
ph	0,1600	0,1933	0,1767				

Tabelle 5.13: Missklassifikationsraten von MS-Klassifikatoren (LDA) für 77 strukturelle Eigenschaften

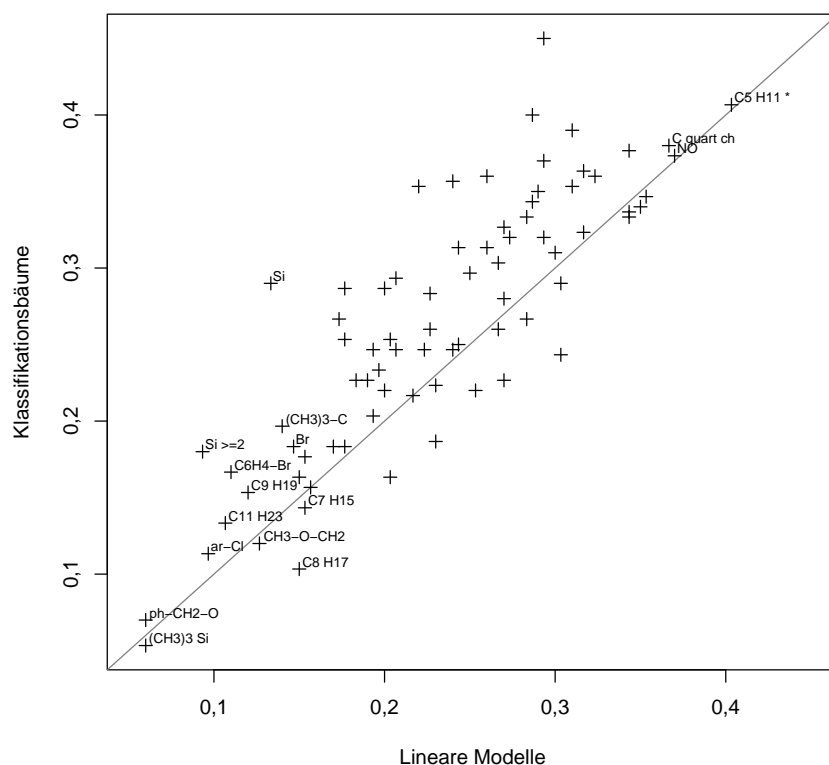


Abbildung 5.32: Missklassifikationsraten für CT mit  $\text{mindev} = 0,04$  und LM mit 13 Deskriptoren

delle mit  $MCE_{TS} < 0,15$  oder keine Modelle mit  $MCE_{TS} < 0,35$  bestimmt wurden, sind zudem die Namen der strukturellen Eigenschaften eingetragen.

### Klassifikation durch ANN und SVM

Neben CT und LDA wollen wir auch ANN und SVM zur Klassifikation von Massenspektren testen. Beim CART-Algorithmus erfolgt die Auswahl der Deskriptoren während der Konstruktion des Entscheidungsbaumes. MLR kann mit sehr geringem Zeitaufwand durchgeführt werden. Deshalb ist es möglich, eine Vielzahl verschiedener Teilmengen innerhalb vertretbarer Zeit zu testen, und gut geeignete Teilmengen von Deskriptoren auszuwählen.

Für ANN und SVM scheiden diese Möglichkeiten aus. Um die verschiedenen Verfahren trotzdem vergleichen zu können, berechnen wir ANN und SVM für die durch CT und MLR bestimmten Teilmengen von Deskriptoren. Wir testen dabei für die 77 strukturellen Eigenschaften neben CT ( $\text{mindev} = 0,04$ ) und LDA auch ANN mit einem, zwei und drei HN sowie SVM mit linearem, radialem, polynomialem ( $\text{degree} = 2$ ) und sigmoid Kernel.

	Min.	1.Quart.	Median	Mittel	3.Quart.	Max.
CT	0,05333	0,20333	0,26667	0,26468	0,33333	0,45000
LDA	0,07667	0,19333	0,24000	0,24762	0,31333	0,46000
ANN, 1HN	0,07000	0,19000	0,24333	0,24498	0,30333	0,48667
ANN, 2HN	0,08000	0,19333	0,25333	0,25260	0,31333	0,45333
ANN, 3HN	0,06000	0,20333	0,26333	0,26048	0,32667	0,44667
SVM, lin	0,08333	0,19667	0,24000	0,24749	0,31667	0,47000
SVM, rad	0,06667	0,19000	0,23333	0,24082	0,30333	0,45667
SVM, pol	0,06667	0,19333	0,24000	0,24368	0,31333	0,44667
SVM, sig	0,08667	0,22000	0,28000	0,28831	0,36333	0,47333

Tabelle 5.14: Missklassifikationsraten verschiedener Klassifikationsverfahren bei Deskriptoren–Selektion durch CT

	Min.	1.Quart.	Median	Mittel	3.Quart.	Max.
CT	0,07000	0,21000	0,25667	0,26485	0,33667	0,45667
LDA	0,06000	0,17667	0,23000	0,23186	0,29000	0,40333
ANN, 1HN	0,08000	0,18000	0,24333	0,23740	0,29667	0,40667
ANN, 2HN	0,09000	0,19000	0,25000	0,24823	0,29667	0,41333
ANN, 3HN	0,08667	0,18000	0,25000	0,24848	0,30000	0,44000
SVM, lin	0,06000	0,17333	0,23333	0,23329	0,28667	0,42000
SVM, rad	0,06333	0,17667	0,22333	0,23160	0,29333	0,42333
SVM, pol	0,05333	0,18000	0,23333	0,23247	0,29333	0,43333
SVM, sig	0,09333	0,18667	0,24667	0,24831	0,30667	0,41000

Tabelle 5.15: Missklassifikationsraten verschiedener Klassifikationsverfahren bei schrittweiser Deskriptoren–Selektion durch MLR

Deskriptoren wurden zum einen durch den CART–Algorithmus (`mindev = 0,04`) gewählt. Hierbei kann die Anzahl von Deskriptoren für die einzelnen SP variieren. Als zweites Verfahren zur Variablen–Auswahl wurde das bereits zuvor verwendete schrittweise Verfahren in Verbindung mit MLR herangezogen. Dabei wurden Teilmengen von 13 Deskriptoren durch 50–fache schrittweise Variablen–Selektion ausgewählt.

Tabellen 5.14 und 5.15 fassen die Missklassifikationsraten auf den Testsätzen zusammen. Tabelle 5.14 enthält Ergebnisse für die Variablen–Selektion durch CART. Man sieht, dass LDA für diese Deskriptoren–Sätze bessere Vorhersagefähigkeit hat als Klassifikation durch CT. Gemessen am arithmetischen Mittel und am Median über alle 77 SP liefern SVM mit radialem Kernel die kleinsten Missklassifikationsraten.

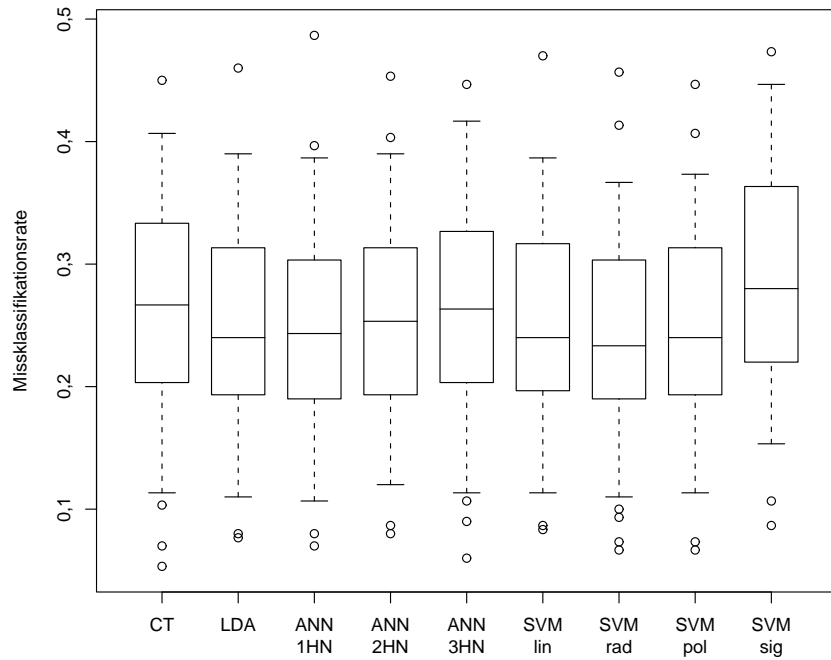


Abbildung 5.33: Missklassifikationsraten verschiedener Klassifikationsverfahren bei Deskriptoren-Selektion durch CT

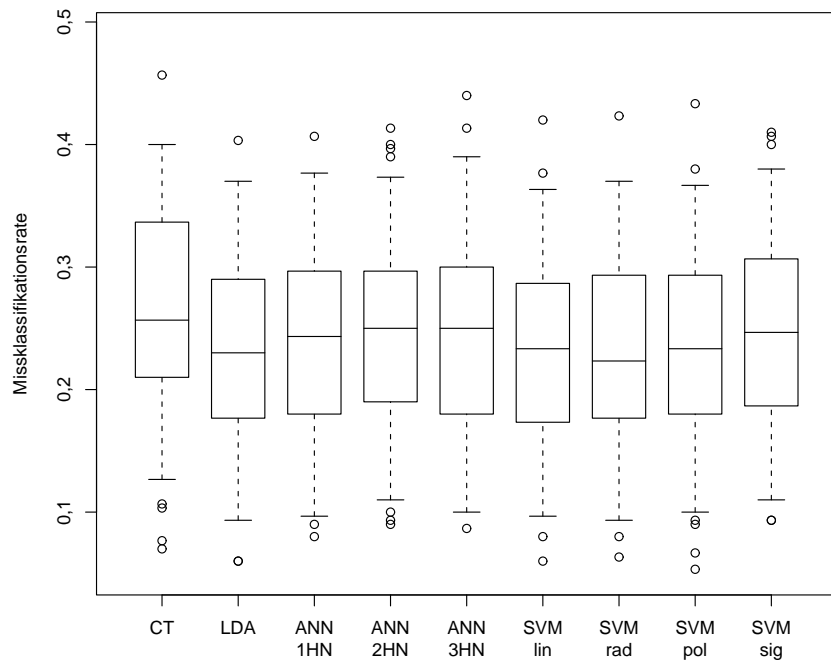


Abbildung 5.34: Missklassifikationsraten verschiedener Klassifikationsverfahren bei schrittweiser Deskriptoren-Selektion durch MLR

Abbildung 5.33 veranschaulicht die Ergebnisse durch einen *Boxplot*. Die Rechtecke in einem Boxplot werden begrenzt durch das erste und dritte Quartil, die vertikale Linie innerhalb der Box repräsentiert den Median. Der durch gestrichelte Linien angedeutete Bereich beschreibt den größten und den kleinsten Wert, der weniger als die **range**-fache Höhe der Box von deren Grenzen entfernt ist. Hier wurde **range** = 0,75 gewählt. Werte, die außerhalb dieses Bereichs liegen, werden durch Kreise repräsentiert.

Tabelle 5.15 fasst die Missklassifikationsraten auf dem Testsatz für Variablen-Selektion durch das schrittweise Verfahren mit MLR zusammen. Hier sind die linearen Modelle deutlich besser als CT, und werden nur von SVM mit radialem Kernel geringfügig übertroffen. Im Vergleich zur Deskriptoren-Selektion durch CT schneiden bei der schrittweisen Selektion LDA, ANN und SVM besser ab. Abbildung 5.34 illustriert die Ergebnisse wiederum als Boxplot. Man sieht, dass die Boxen tiefer liegen als in Abbildung 5.33.

Unter den hier getesteten Verfahren zeigen SVM mit radialem Kernel in Verbindung mit Variablen-Selektion durch MLR die beste Vorhersagefähigkeit. Tabelle 5.12 zeigt zu diesem Verfahren  $MCE_{TS}^F$ ,  $MCE_{TS}^W$  und  $MCE_{TS}$  für die 77 SP im Einzelnen.

Leider können die Ergebnisse mit anderen Arbeiten auf diesem Gebiet nur eingeschränkt verglichen werden. Zunächst unterscheiden sich die zufällig ausgewählten Lern- und Testsätze, weiterhin wurden unterschiedliche Verfahren zur Deskriptoren-Selektion verwendet. Jedoch wird beim Vergleich mit den Ergebnissen aus [163] deutlich, dass übereinstimmend manche SP gut zur Klassifikation geeignet sind, andere schlechter.

Jüngst konnten durch Verwendung von *Boosted Trees* [149] deutliche Verbesserungen in der Vorhersagefähigkeit gegenüber Entscheidungsbäumen und linearen Verfahren erzielt werden. Allerdings unterscheiden sich diese Berechnungen in den verwendeten Deskriptoren und den Größen von Lern- und Testsatz von den hier dargestellten Ergebnissen.

Um die Forschung auf diesem entscheidenden Teilgebiet der CASE mit MS voranzutreiben, wäre es wichtig, Referenz-Datensätze anzulegen und öffentlich verfügbar zu machen, damit verschiedene Ansätze vergleichbar werden, und die rasanten Entwicklungen auf dem Gebiet der maschinellen Lernverfahren rasch für die konkrete Anwendung genutzt werden können.

### 5.5.3 Systematische Suche nach neuen Substrukturen für MS-Klassifikatoren

Wir fragen uns, ob es jenseits der in Anhang C vorgestellten strukturellen Eigenschaften kleine Substrukturen gibt, die gut durch MS-Klassifikatoren



Name	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$	Name	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$
C quart ch	0,4000	0,3000	0,3500	benz-0	0,1867	0,1667	0,1767
C4 H9	0,2333	0,2667	0,2500	ar-COOCH2*	0,1667	0,1733	0,1700
C5 H11	0,2600	0,1733	0,2167	naph	0,1800	0,1533	0,1667
C5 H11 *	0,4200	0,4267	0,4233	phen-1-OH	0,2933	0,3400	0,3167
C6 H13	0,1733	0,3067	0,2400	phen	0,3200	0,3400	0,3300
C6 H13 n	0,1933	0,2133	0,2033	phen-2-OH	0,2600	0,2600	0,2600
C7 H15	0,1200	0,2000	0,1600	CO-C6H3-0	0,3333	0,2800	0,3067
C8 H17	0,1600	0,1267	0,1433	r>C=C<r	0,1200	0,3000	0,2100
C9 H19	0,1800	0,0800	0,1300	(CH3)2>C=C	0,2867	0,2133	0,2500
C10 H21	0,1533	0,0800	0,1167	Br	0,1667	0,2333	0,2000
C11 H23	0,1133	0,1133	0,1133	Cl	0,1200	0,1867	0,1533
hydr carb	0,1867	0,1067	0,1467	N	0,3333	0,2267	0,2800
(CH3)3-C	0,1400	0,1600	0,1500	N 2	0,3400	0,3667	0,3533
ar-CHO	0,2067	0,2000	0,2033	P	0,3067	0,3200	0,3133
ar-CO,N2	0,3533	0,2267	0,2900	Si	0,1000	0,1400	0,1200
ar-CH2	0,3533	0,2867	0,3200	Si 1	0,2067	0,2533	0,2300
ar-N,NHN	0,3800	0,2067	0,2933	Si >=2	0,0800	0,1067	0,0933
ar-C ch	0,2533	0,1533	0,2033	S	0,3067	0,2667	0,2867
ar-C r	0,2067	0,1800	0,1933	CH3-COO	0,2400	0,1933	0,2167
ar-CO	0,4067	0,1733	0,2900	CH3-CO	0,1600	0,1800	0,1700
ar-CH	0,3200	0,2867	0,3033	alc tert	0,2800	0,3867	0,3333
ar-CH2CH2	0,3200	0,1600	0,2400	am tert	0,2800	0,2667	0,2733
ar-Cl	0,0933	0,1000	0,0967	n-C4H9-0	0,1467	0,2133	0,1800
ar-CO-CH2	0,3667	0,3133	0,3400	C2H5-CO	0,1533	0,2267	0,1900
ar-COO	0,4000	0,1800	0,2900	CF3	0,1600	0,1733	0,1667
ar-F	0,1933	0,2533	0,2233	CF3-CO	0,1267	0,2667	0,1967
ar-N	0,2133	0,3000	0,2567	NH-CH2-CH2	0,2467	0,3133	0,2800
ar-N ch	0,3000	0,2333	0,2667	CH3-0-CH2	0,0867	0,1467	0,1167
ar-N r	0,2933	0,3067	0,3000	N(CH3)2	0,2267	0,1800	0,2033
ar-0	0,3867	0,2867	0,3367	CH3-COOCH	0,2200	0,1867	0,2033
ar-0-CH2	0,3267	0,3200	0,3233	et-est	0,4333	0,3067	0,3700
ar-0-CH3	0,2667	0,2267	0,2467	me-est	0,1733	0,3400	0,2567
ar-S r	0,2533	0,2067	0,2300	C2H5-0	0,2867	0,3067	0,2967
biphenyl	0,2733	0,1533	0,2133	(CH2)6-CO	0,2200	0,1667	0,1933
C6H4-Br	0,0733	0,1467	0,1100	NO	0,3267	0,3867	0,3567
C6H4 omp	0,2133	0,1800	0,1967	S-CH2	0,2667	0,3267	0,2967
ph-C	0,1867	0,2000	0,1933	(CH3)3 Si	0,0333	0,0933	0,0633
ph-CH2-0	0,0733	0,0867	0,0800	r 5+6	0,3333	0,3933	0,3633
ph	0,1733	0,2400	0,2067				

Tabelle 5.16: Missklassifikationsraten von MS-Klassifikatoren (SVM mit radialem Kernel) für 77 strukturelle Eigenschaften

Substruktur	Verfahren	$MCE_{TS}^F$	$MCE_{TS}^W$	$MCE_{TS}$
C–Si–C	SVM, rad	0,0467	0,1133	0,0800
C–Si–N	LDA	0,0800	0,0667	0,0733
C=C–Cl	ANN, 1HN	0,0800	0,1067	0,0933
Cl–C=C–Cl	LDA	0,0467	0,0733	0,0600
O=C–O–Si	SVM, rad	0,0733	0,0600	0,0667
C–C–N–Si	CT	0,0267	0,0867	0,0567
C=C–N–Si	CT	0,0400	0,0800	0,0600
C <sub>2</sub> >N–Si	LDA	0,1000	0,0733	0,0867
C–N–Si–C	LDA	0,0733	0,0800	0,0767
C <sub>2</sub> >Si–N	SVM, rad	0,0667	0,0867	0,0767
C=C–O–Si	LDA	0,1067	0,0533	0,0800
C–O–Si–C	SVM, pol	0,1067	0,0533	0,0800
C <sub>2</sub> >Si–O	SVM, pol	0,0667	0,0867	0,0767
C <sub>2</sub> >Si–C	SVM, pol	0,0600	0,1333	0,0967
C–C=C–Cl	ANN, 3HN	0,0733	0,1000	0,0867

Tabelle 5.17: Missklassifikationsraten für gut klassifizierbare Substrukturen

erkennbar sind. Wir wollen eine systematische Suche nach Substrukturen durchführen, für die gute MS-Klassifikatoren bestimmt werden können.

Wir verwenden dafür Algorithmus 4.3.7, der zu einer gegebenen Menge molekularer Graphen systematisch alle molekularen Teilgraphen findet. In unserer MS-Struktur-Datenbasis werden auf diese Weise alle Substrukturen mit bis zu 3 Kanten ermittelt. Davon existieren 1790. Diese werden als strukturelle Eigenschaften zur Klassifikation von Massenspektren herangezogen. Wir verfahren wie unter Abschnitt 5.5.2 beschrieben, und bilden zunächst Lern- und Testsätze mit je 150 Datensätzen pro Klasse.

Von den 1790 Substrukturen verbleiben 301, für die die geforderten Lern- und Testsätze gebildet werden können. Für jede dieser 301 Substrukturen berechnen wir MS-Klassifikatoren. Zur Deskriptoren-Selektion verwenden wir wie zuvor beschrieben das 50-fache schrittweise Verfahren mit MLR. Auf diese Weise werden für jede Substruktur 13 MS-Deskriptoren ermittelt, die für die Modellierung relevant sind. Zur Klassifikation verwenden wir CT, LDA, ANN mit einem, zwei und drei HN sowie SVM mit linearem, radialem, polynomialem (`degree = 2`) und sigmoid Kernel.

Wir sind nur an Substrukturen interessiert, für die gute Klassifikatoren gefunden werden können. Es gibt 15 Substrukturen, für die insgesamt 80 Klassifikatoren mit  $MCE_{TS} < 0,1$  ermittelt werden. Diese 80 Klassifikatoren

verteilen sich wie folgt auf die einzelnen Verfahren:

CT	LDA	ANN 1HN	ANN 2HN	ANN 3HN	SVM lin	SVM rad	SVM pol	SVM sig
4	13	10	4	6	12	10	14	7

Tabelle 5.17 gibt für jede der 15 Substrukturen  $MCE_{TS}^F$ ,  $MCE_{TS}^W$ ,  $MCE_{TS}$  und das Klassifikationsverfahren des Klassifikators mit bester Vorhersagefähigkeit an.

## 5.6 Automatisierte Strukturaufklärung mit Massenspektrometrie

Die beschriebenen Verfahren zur MS-Klassifikation und zum Ranking von Brutto- und Strukturformeln legen es nahe, diese einzelnen Prozeduren in Verbindung mit einem Strukturgenerator zur automatisierten Strukturaufklärung nach Schema 5.1 und 5.4 zu koppeln. Nach den Ergebnissen über die Güte von Rankingfunktionen (Abschnitt 5.4) und die Missklassifikationsraten von MS-Klassifikatoren (Abschnitt 5.5) dürfte jedoch auch dem optimistischen Leser bewusst sein, dass die Verwendung solchen Systems im automatischen Modus nach derzeitigem Stand der Forschung riskant ist. Im interaktiven Betrieb kann das Zusammenspiel der einzelnen Komponenten dagegen durchaus zu einem Erkenntnisgewinn für konkrete Anwendungen reichen. In *MOLGEN-MS* [58, 76] wurde ein solcher Prototyp realisiert. Im Folgenden wird die automatische Strukturaufklärung mit MS anhand zweier Beispiele durchgeführt.

### 5.6.1 Beispiel: n-Pentansäuremethylester

Für die Extraktion struktureller Eigenschaften verwenden wir MS-Klassifikatoren aus *MSclass* [154]. Diese Klassifikatoren wurden über Klassifikation durch Regression (Abschnitt 3.1.1) ermittelt. Beim Trainieren der Klassifikatoren wurden für die Werte der Diskriminanzfunktion Prozent-Quantile berechnet. Somit ist es bei der Vorhersage möglich, eine Genauigkeit für die Prognose anzugeben. Klassifikator-Antworten mit geringer Genauigkeit können auf diese Weise unterdrückt werden. Allerdings verringert sich damit auch die Anzahl struktureller Eigenschaften für die Bruttoformel- und Strukturgenerierung.

Insgesamt stehen 160 Klassifikatoren zu 85 strukturellen Eigenschaften (Anhang C) zur Verfügung. Für manche Eigenschaften gibt es bis zu 4 verschiedene Klassifikatoren. Solche Klassifikatoren zur gleichen strukturellen Eigenschaft unterscheiden sich in den verwendeten Deskriptoren und/oder dem Klassifikationsverfahren.

#### MS-Klassifikation

Angewandt auf das Spektrum *I* aus Beispiel 5.3.2 erhalten wir 68 Klassifikator-Antworten mit einer Genauigkeit von mindestens 95%, wovon 5 positiv und 63 negativ ausfallen. Die Ausgabe von *MSclass* für die positiven Klassifikations-Ergebnisse lautet:

Name	Class	Prec	Description	Type
non ar /2	true	99.03	aroma: non aromatic	RBF
me-est /1	true	98.16	func: ester: methyl	LDA
non ar /1	true	97.20	aroma: non aromatic	LDA
me-est /2	true	97.14	func: ester: methyl	LDA
me-est /4	true	95.07	func: ester: methyl	RBF

Dabei wird pro Zeile ein Klassifikator-Ergebnis angegeben. In der ersten Spalte befindet sich der Name des Klassifikators. Dieser ist zusammengesetzt aus der Abkürzung für die strukturelle Eigenschaft und der Nummer des Klassifikators. Wir haben hier 3 positive Antworten zur strukturellen Eigenschaft *Methylester*. Des Weiteren ist in Spalte *Prec* die Genauigkeit der Vorhersage und in der letzten Spalte das verwendete Klassifikationsverfahren aufgeführt. Als Klassifikationsverfahren werden in *MSclass* LDA und ANN mit *radialen Basis-Funktionen* (kurz *RBF*) verwendet.

Wichtigstes Resultat in diesem Beispiel ist die Erkennung der funktionellen Gruppe Methylester, die 3 von insgesamt 4 Klassifikatoren für diese Substruktur als vorhanden prognostizieren. Die Aussage, dass der Analyt nicht aromatisch ist, erweist sich für den weiteren Strukturaufklärungsprozess als wenig ergiebig.

Zunächst muss man Klassifikationsergebnisse für gleiche strukturelle Eigenschaften zusammenfassen. Hierfür bieten sich verschiedene, auf Mehrheitsbeschluss basierende Verfahren an: *relative*, *absolute*, *qualifizierte*, *konsistente* oder *einstimmige* Mehrheit. Wir verwenden das Prinzip der konsistenten Mehrheit, d.h. es muss mindestens eine Klassifikator-Antwort mit der geforderten Mindestgenauigkeit vorliegen und es dürfen keine gegenteiligen Klassifikationsergebnisse eintreten. Auf diese Weise werden 2 strukturelle Eigenschaften als anwesend und 40 als abwesend vorhergesagt.

### Bestimmung der Bruttoformel

Wegen der Anwesenheit von Methylester muss für die Bruttoformel  $\beta$  des Analyten gelten

$$\beta(\text{C}) \geq 2, \beta(\text{H}) \geq 3, \beta(\text{O}) \geq 2.$$

Aus demselben Grund muss  $\beta$  mindestens ein Doppelbindungsäquivalent aufweisen:  $\text{DBE}(\beta) \geq 1$ . Negative Klassifikator-Antworten liefern Informationen über die Abwesenheit der Elemente P und Si:

$$\beta(\text{P}) = \beta(\text{Si}) = 0.$$

Für die Bestimmung von Bruttoformel-Kandidaten wollen wir voraussetzen, dass die Molekülmasse von 116 amu bekannt ist. In Beispiel 5.4.5 haben wir

bereits Bruttoformeln zu dieser Masse unter verschiedenen Nebenbedingungen generiert. In  $\mathcal{B}_{\varepsilon_{11}}^C$  gibt es 9 Bruttoformeln, welche die oben genannten Anforderungen für die Vielfachheiten der Elemente erfüllen. Eine davon verstößt gegen die DBE-Bedingung :  $\text{DBE}(\text{C}_2\text{H}_3\text{F}_3\text{O}_2) = 0$ . Für die verbleibenden 8 Bruttoformeln werden folgende Vergleichswerte berechnet:

1.  $\text{MV}(I, \text{C}_4\text{H}_8\text{N}_2\text{O}_2) = 0,9978946$
2.  $\text{MV}(I, \text{C}_3\text{H}_4\text{N}_2\text{O}_3) = 0,9966711$
3.  $\text{MV}(I, \text{C}_2\text{H}_4\text{N}_4\text{O}_2) = 0,9958909$
4.  $\text{MV}(I, \text{C}_5\text{H}_8\text{O}_3) = 0,9958316$
5.  $\text{MV}(I, \text{C}_5\text{H}_5\text{FO}_2) = 0,9952805$
6.  $\text{MV}(I, \text{C}_6\text{H}_{12}\text{O}_2) = 0,9942863$
7.  $\text{MV}(I, \text{C}_4\text{H}_4\text{O}_4) = 0,8240869$
8.  $\text{MV}(I, \text{C}_4\text{H}_4\text{O}_2\text{S}) = 0,8205838$

Basierend auf den Quantilen für Bruttoformel-Vergleichswerte aus Tabelle 5.7 kann man nun eine Auswahl relevanter Bruttoformel-Kandidaten treffen. Möchte man mit einer Zuverlässigkeit von 95% die korrekte Bruttoformel in der Selektion vorfinden, so muss man alle Kandidaten  $\beta$  mit Vergleichswert  $\text{MV}(I, \beta) \geq 0,8775297$  berücksichtigen. Auf diese Weise werden  $\text{C}_4\text{H}_4\text{O}_4$  und  $\text{C}_4\text{H}_4\text{O}_2\text{S}$  ausgeschlossen.

### Bestimmung der Strukturformel

Zu den verbleibenden 6 Summenformeln gibt es insgesamt 118669 Strukturformeln. Aus der strukturellen Eigenschaft Methylester gewinnt man eine vorgeschriebene Substruktur und aus den 40 abwesenden strukturellen Eigenschaften erhält man 52 verbotene Substrukturen. Mit diesen Restriktionen verbleiben 123 Strukturformeln. Unter Verwendung der *permanenten Badlist* nach [150] mit 32 Substrukturen kann man den Strukturraum um weitere 4 Kandidaten verkleinern.

Abbildung 5.35 zeigt die Strukturformel-Vergleichswerte zu den 123 Kandidaten. Vergleichswerte der 4 Strukturen, die durch die permanente Badlist ausgeschlossen werden, sind grau markiert. Auf der rechten Seite sind entlang der  $y$ -Achse Prozent-Quantile für Strukturformel-Vergleichswert nach Tabelle 5.9 aufgetragen. Möchte man nach diesen Werten eine Menge von Kandidaten auswählen, die mit einer Genauigkeit von 80% die korrekte Struktur umfasst, so sind die 13 besten Kandidaten zu betrachten. In diesem Beispiel heben sich die Vergleichswerte der 13 besten Kandidaten deutlich von den übrigen Werten ab. In Abbildung 5.36 sind die 28 bestplatzierten Strukturkandidaten sortiert nach abfallenden Vergleichswerten dargestellt. Die kor-

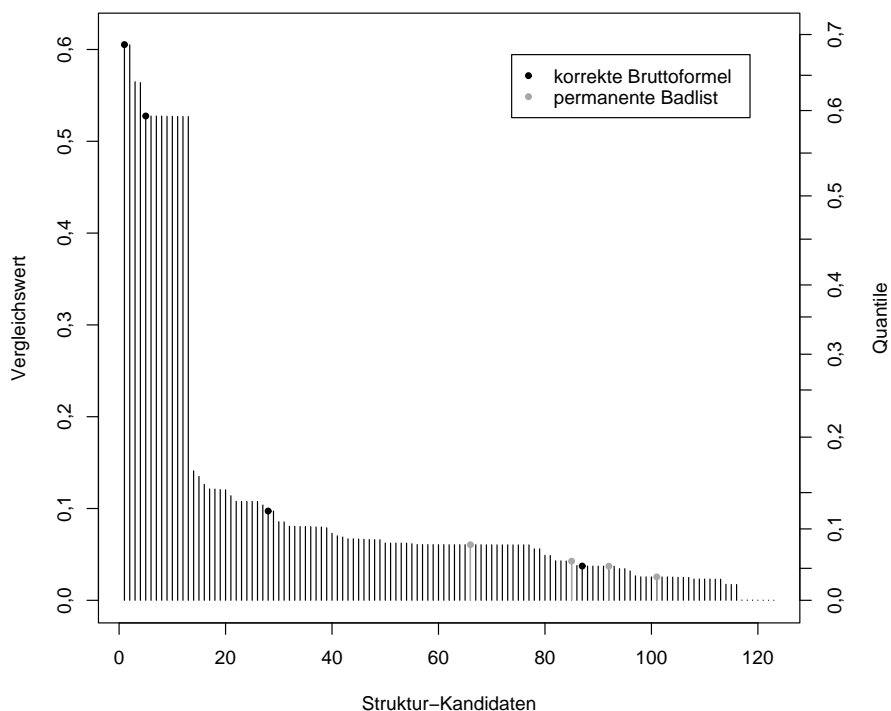
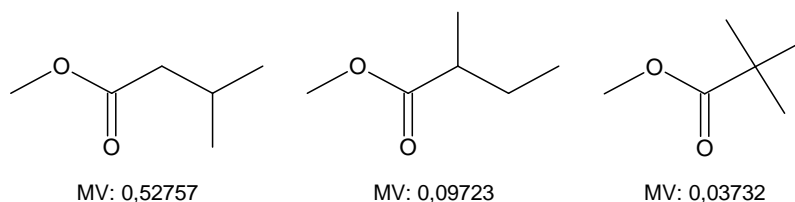


Abbildung 5.35: Vergleichswerte der Strukturkandidaten für das Spektrum aus Beispiel 5.3.2

rekte Struktur, *n*-Pentansäuremethylester, befindet sich in diesem Ranking tatsächlich an erster Position.

Ein besonders deutliches Ergebnis erhält man, wenn man sich auf Strukturformeln zur korrekten Bruttoformel  $C_6H_{12}O_2$  beschränkt. Man erhält dann 4 Strukturkandidaten. Die korrekte Strukturformel hat einen Vergleichswert von 0,60530. Die weiteren Strukturkandidaten zu  $C_6H_{12}O_2$  und ihre Vergleichswerte sind:



In Abbildung 5.35 sind die Vergleichswerte von Strukturen zur korrekten Bruttoformel durch eine schwarze Markierung hervorgehoben.

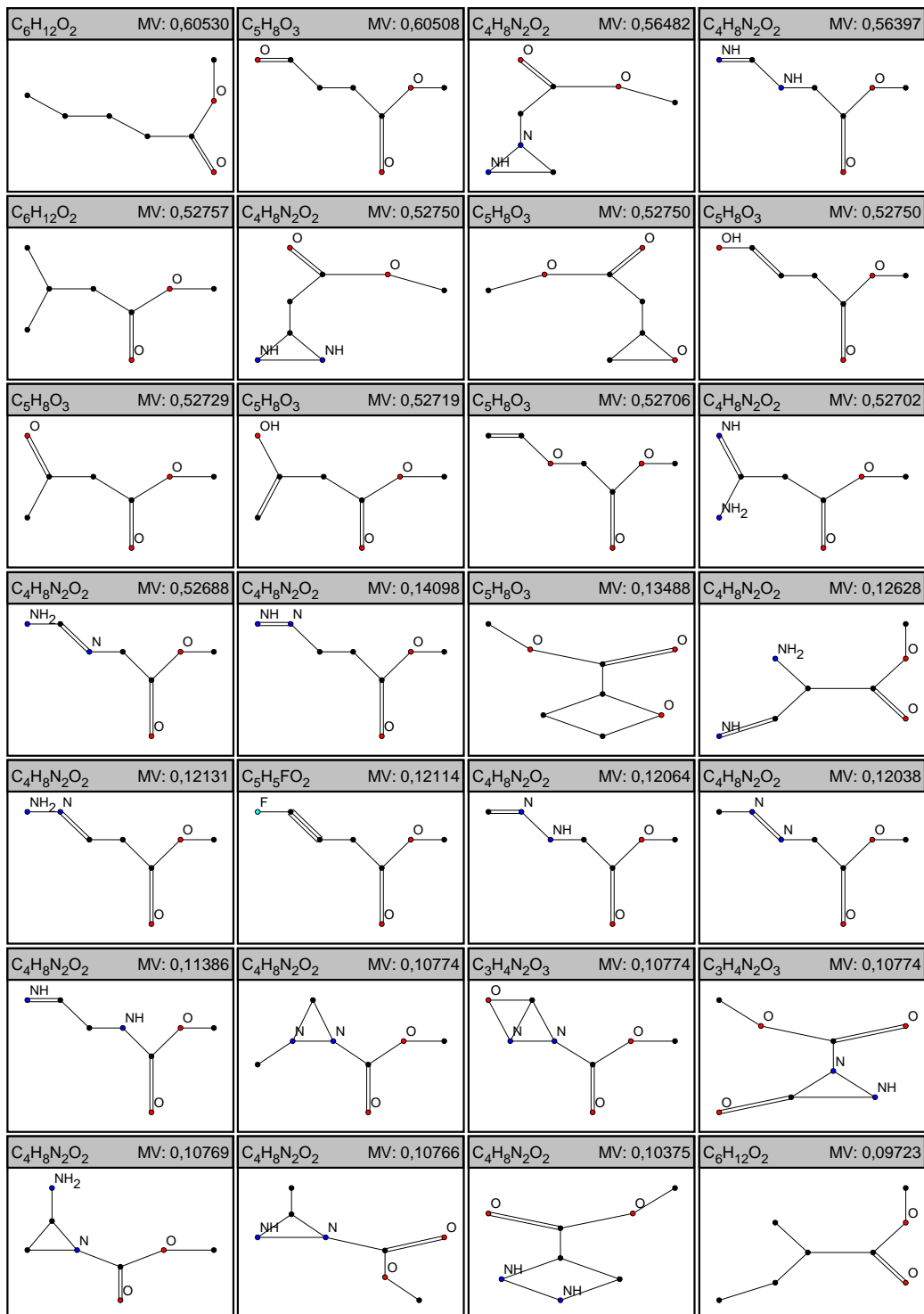


Abbildung 5.36: Ranking von Strukturkandidaten für das Spektrum aus Beispiel 5.3.2



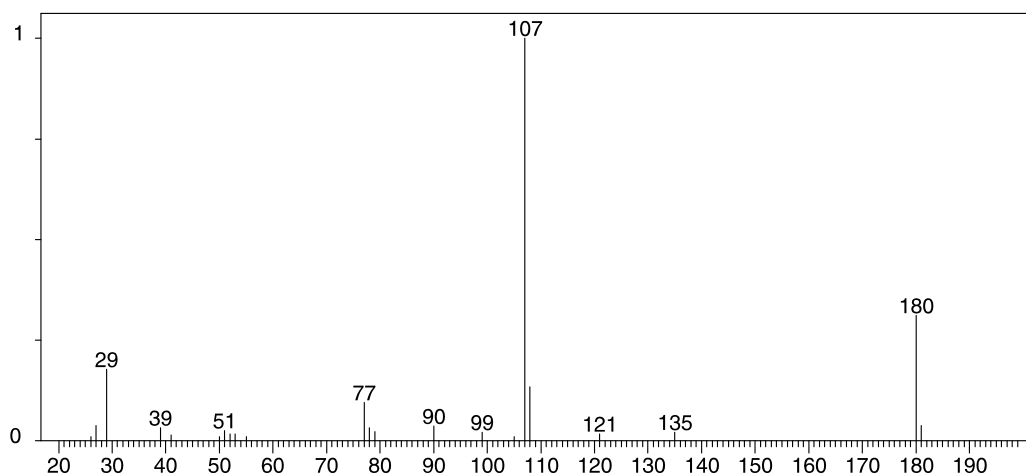


Abbildung 5.37: Massenspektrum von 3-Hydroxyphenylethylacetat

### 5.6.2 Beispiel: 3-Hydroxyphenylethylacetat

Abbildung 5.37 zeigt das Massenspektrum unseres zweiten Beispiels. Wir beginnen unsere Strukturaufklärung wie zuvor mit MS-Klassifikation.

#### MS-Klassifikation

Die 160 Klassifikatoren aus *MSClass* liefern bei 95% Mindestgenauigkeit 6 positive und 42 negative Antworten. Die positiven Klassifikator-Ergebnisse sind:

Name	Prec	Description	Type
phen	/1 99.17	aroma: phenol (1-3 OH), alkyl-subst.	LDA
phen	/2 99.15	aroma: phenol (1-3 OH), alkyl-subst.	LDA
benz-0	/1 99.03	aroma: CH <sub>2</sub> - C <sub>6</sub> H <sub>4</sub> - O - (o,m,p)	LDA
phen-1-OH	/2 99.01	aroma: phenol (1 OH), alkyl-subst.	LDA
phen-1-OH	/1 99.00	aroma: phenol (1 OH), alkyl-subst.	LDA
et-est	/1 96.11	func: ester: ethyl	RBF

Wir fassen die Klassifikator-Ergebnisse für gleiche strukturelle Eigenschaften wieder durch Mehrheitsprinzip zusammen. Mit konsistenter Mehrheit erhalten wir 4 vorhandene und 26 verbotene strukturelle Eigenschaften.

### Bestimmung der Bruttoformel

Für die Bruttoformel  $\beta$  der unbekanntenen Verbindung erhalten wir daraus folgende Einschränkungen

$$\beta(\text{C}) \geq 7, \beta(\text{H}) \geq 6, \beta(\text{O}) \geq 2, \beta(\text{Si}) = 0 \text{ und } \text{DBE}(\beta) \geq 4.$$

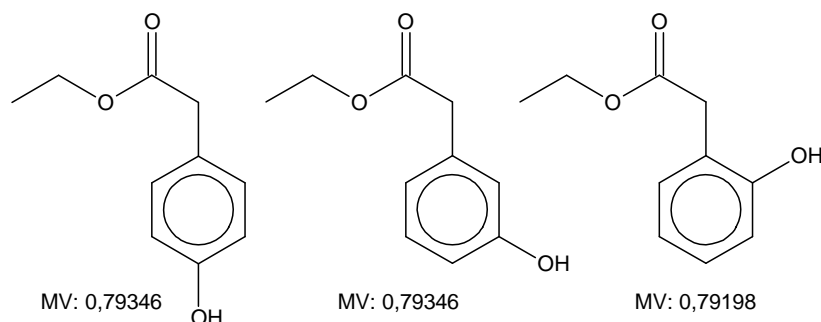
Die Molekülmasse von 180 amu setzen wir wiederum als bekannt voraus. Basierend auf den Elementen aus  $\mathcal{E}_{11}$  gibt es zu dieser Masse insgesamt 14353 Bruttoformeln. Davon sind 2064 aus  $\mathcal{B}_{\mathcal{E}_{11}}^{\text{C}}$ . Unter Berücksichtigung der Beschränkungen für die Vielfachheiten der Elemente verbleiben 23 Bruttoformeln, 9 genügen zusätzlich der DBE-Einschränkung. Für diese berechnen wir folgende Vergleichswerte:

1.  $\text{MV}(I, \text{C}_9\text{H}_{12}\text{N}_2\text{O}_2) = 0,9923722$
2.  $\text{MV}(I, \text{C}_{10}\text{H}_9\text{O}_2\text{F}) = 0,9921804$
3.  $\text{MV}(I, \text{C}_{11}\text{H}_{16}\text{O}_2) = 0,9915165$
4.  $\text{MV}(I, \text{C}_{10}\text{H}_{12}\text{O}_3) = 0,9914339$
5.  $\text{MV}(I, \text{C}_9\text{H}_9\text{O}_2\text{P}) = 0,9911746$
6.  $\text{MV}(I, \text{C}_8\text{H}_8\text{N}_2\text{O}_3) = 0,9909998$
7.  $\text{MV}(I, \text{C}_7\text{H}_8\text{N}_4\text{O}_2) = 0,9905219$
8.  $\text{MV}(I, \text{C}_9\text{H}_8\text{O}_4) = 0,9896856$
9.  $\text{MV}(I, \text{C}_9\text{H}_8\text{O}_2\text{S}) = 0,9818188$

Bei einer Genauigkeit von 95% für die Bruttoformel-Selektion werden alle 9 Kandidaten ausgewählt.

### Bestimmung der Strukturformel

Als Eingabe für die Strukturgenerierung erhält man neben den Summenformeln 3 Einträge für die Goodlist und 26 Einträge für die Badlist. Der Goodlist-Eintrag für *benz-O* besteht seinerseits aus 5 Substrukturen, die mit logischem *oder* verknüpft sind (vgl. Anhang C.2). Die für *MOLGEN* verwertbaren strukturellen Informationen, welche *phen* und *phen-1-OH* liefern, beschränken sich auf einen, einfach mit OH substituierten Benzolring. Die Strukturgenerierung liefert folgende 3 Strukturen:



Die 3 Strukturkandidaten unterscheiden sich also nur in der Position der, an den Benzolring substituierten OH-Gruppe. Die große Ähnlichkeit dieser Strukturen drückt sich auch in ihren Vergleichswerten bzgl.  $I$  aus. Die korrekte Struktur ist in der Mitte abgebildet, *3-Hydroxyphenyllessigsäureethylester*.

## Fazit und Ausblick

Wie wir in den beiden Beispielen gesehen haben, können die vorgestellten Verfahren grundsätzlich zur automatisierten Strukturaufklärung herangezogen werden. Es ist aber notwendig alle Verbesserungsmöglichkeiten, insbesondere für die Interpretation und Verifikation auszuschöpfen. Dazu zählen im Problembereich der

- MS-Klassifikation
  - die Entwicklung weiterer Deskriptoren, die für die einzelnen Klassifikationsprobleme speziell angepasst sind,
  - die Entwicklung von Deskriptoren, die weitere verfügbare Informationen, wie etwa die Molekülmasse oder exakte Fragmentmassen einbeziehen können,
  - die Testung weiterer Verfahren zur Deskriptoren-Selektion,
  - die Testung weiterer Klassifikationsverfahren, wie SVM, sowie die Parameter-Optimierung für diese Verfahren,
  - die Berücksichtigung weiterer struktureller Eigenschaften, auch wenn für diese nur Klassifikatoren mit geringer Vorhersagefähigkeit gefunden werden können. Dies kann dann kompensiert werden durch
  - die Filterung der Klassifikationsergebnisse unter Einbeziehung der logischen Implikationen zwischen den einzelnen strukturellen Eigenschaften [153].

- Strukturgenerierung
  - die direkte Verarbeitung aromatischer Substrukturen.
- MS-Verifikation
  - die Testung weiterer Rankingfunktionen, etwa aus [132],
  - die Testung verschiedener Parameter für Rankingfunktionen, wie eine untere Grenze für die DBE von Fragment-Bruttoformeln bei der Berechnung von Vergleichswerten für die Summenformel, oder die Verwendung verschiedener Sätze von Fragmentierungsreaktionen bei der Bestimmung von Strukturformel-Vergleichswerten,
  - die Entwicklung und Testung von Kriterien zur Plausibilität von Brutto- und Strukturformeln,
  - die Berücksichtigung reaktionsdynamischer Aspekte.

## 5.7 Quantitative MS-Eigenschafts- Beziehungen

Wie wir gesehen haben bereitet sowohl die Strukturaufklärung als auch die Suche nach Struktur-Eigenschafts-Beziehung mitunter erhebliche Probleme. Natürlich ist die topologische, und noch mehr die geometrische Struktur einer chemischen Verbindung der Schlüssel zur Berechnung und Vorhersage ihrer Eigenschaften. In den meisten Fällen ist man schon allein aus Gesichtspunkten der Datenverarbeitung und Dokumentation darauf angewiesen, die Struktur einer untersuchten Verbindung möglichst exakt zu bestimmen.

Es gibt allerdings auch Situationen, in denen man bestrebt ist, den erheblichen Aufwand einer Strukturbestimmung zu vermeiden. Man stelle sich ein kombinatorisch-chemisches Experiment vor, bei dem sowohl die Synthese als auch die spektroskopische Vermessung der untersuchten Verbindungen zu geringen Kosten durchgeführt werden können, das Screening jedoch mit verhältnismäßig hohen Kosten zu Buche schlägt. Falls die Strukturaufklärung wegen der beschriebenen Probleme ein unüberwindbares Hindernis darstellt, kann man versuchen, die vorhandenen spektroskopischen Daten mit der betrachteten Eigenschaft direkt in Beziehung zu setzen.

Abbildung 5.38 skizziert die Vorgehensweise bei der Suche und Anwendung *quantitativer Spektrum-Eigenschafts-Beziehungen*. Die untersuchte Eigenschaft ist dabei Zielvariable für ein statistisches Lernverfahren. Vorhersagevariablen werden über spektrale Deskriptoren gewonnen.

### Beispiel: Siedepunkte von Decanen

Wir wollen diese Vorgehensweise an dem einfachen Beispiel von Decanen und ihren Siedepunkten demonstrieren: In unserer Spektrendatenbank sind Massenspektren zu 37 Decanen abgelegt, der *Beilstein*-Datenbank haben wir bereits in Abschnitt 4.4.1 Siedepunkte für 50 Decane entnommen. Zunächst stellen wir fest, welche Strukturen in beiden Datenbeständen vorhanden sind. Dazu bringen wir die Strukturen beider Anfragen auf kanonische Form, bestimmen den Durchschnitt, und ordnen die Siedepunkte den Massenspektren gleicher Struktur zu. Auf diese Weise erhalten wir 29 Paare von Massenspektren und Siedepunkten (Abbildung 5.39).

### Deskriptoren

Wir verwenden wieder die 445 MS-Deskriptoren aus Abschnitt 5.5. Zunächst entfernen wir alle Deskriptoren, die auf den 29 MS konstant sind. Es handelt sich dabei um 167 logarithmische Intensitätsverhältnisse, die durchwegs den

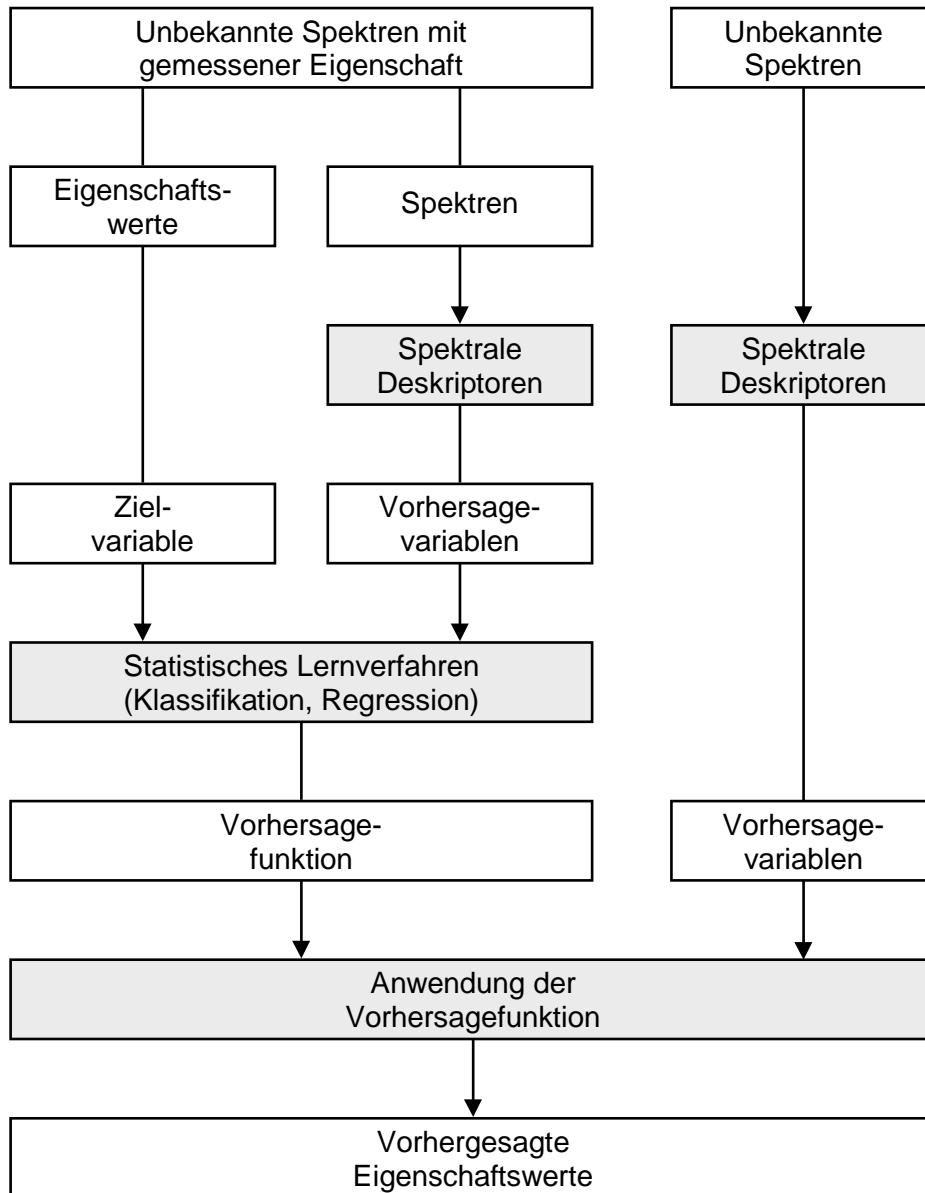


Abbildung 5.38: Vorgehensweise bei der Vorhersage von Eigenschaften durch Spektren–Eigenschafts–Beziehungen

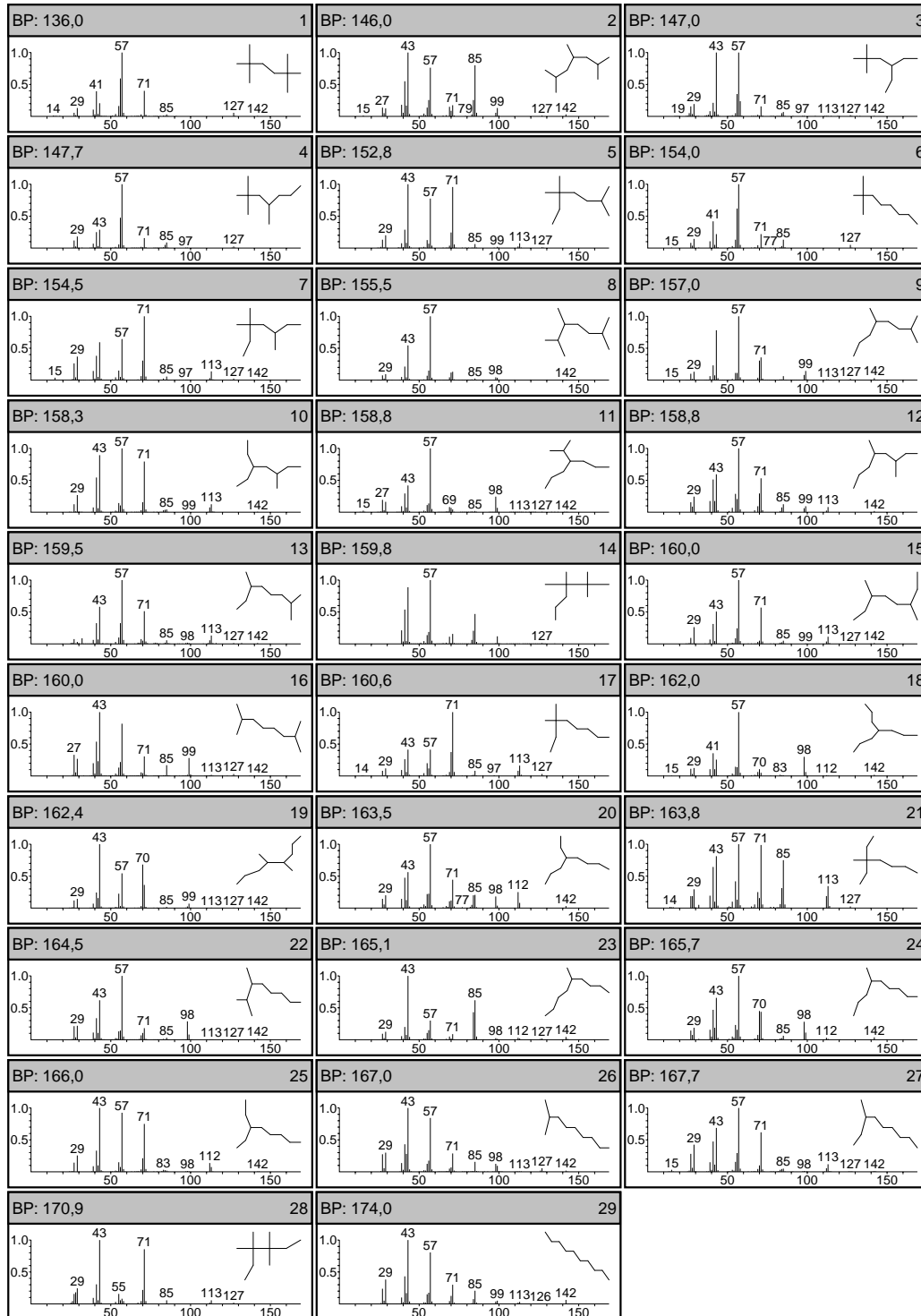


Abbildung 5.39: Massenspektren, Strukturen und BP von 29 Decanen

Wert Null annehmen. Weitere 15 Deskriptoren sind nur für jeweils ein Spektrum ungleich Null. Dies sind logarithmische Intensitätsverhältnisse  $LIQN_{m,d}$  mit folgenden Parametern  $(m, d)$ :

$$(45, 1), (45, 2), (50, 2), (76, 2), (78, 2), (80, 2), (81, 1), (109, 2), (110, 1), (115, 1), (116, 1), (116, 2), (124, 2), (125, 1), (126, 2).$$

Diese werden wir ebenfalls vernachlässigen. Unter den verbleibenden 263 Deskriptoren suchen wir nach vollständigen Korrelationen. Wir finden solche Korrelationen wiederum nur unter logarithmischen Intensitätsverhältnissen bei folgenden Parametern  $(m, d)$ :

$$(49, 2) \sim (50, 1), (58, 1) \sim (58, 2), (63, 2) \sim (64, 1) \sim (65, 1), (72, 1) \sim (72, 2), (75, 2) \sim (76, 1), (79, 1) \sim (79, 2), (86, 1) \sim (86, 2), (96, 2) \sim (97, 1), (97, 2) \sim (99, 2), (100, 1) \sim (100, 2), (110, 2) \sim (111, 1), (111, 2) \sim (113, 2), (125, 2) \sim (127, 1) \sim (127, 2), (140, 2) \sim (141, 1) \sim (142, 1) \sim (142, 2).$$

Auf diese Weise entfernen wir weitere 18 Deskriptoren.

### Lineare Modelle

Unter den übrigen 245 Deskriptoren suchen wir  $n$ -Teilmengen von  $n = 1, \dots, 5$  Deskriptoren<sup>5</sup>, die beste lineare Modelle bzgl.  $R^2$  liefern. Wir erhalten mit

$n = 1$  Deskriptor: ACLH<sub>49</sub>

$$f = -6628,0X_0 + 161,17$$

$$R^2 = 0,32093, S = 6,7378, F = 12,760, R_{CV}^2 = 0,16159, S_{CV} = 7,4866$$

$n = 2$  Deskriptoren: MD14<sub>9</sub>, AUCCO<sub>12</sub>

$$f = -22,767X_0 + 1,1299X_1 - 152,18$$

$$R^2 = 0,55538, S = 5,5558, F = 16,238, R_{CV}^2 = 0,43511, S_{CV} = 6,2623$$

$n = 3$  Deskriptoren: MD14<sub>9</sub>, ACLH<sub>12</sub>, ACUH<sub>31</sub>

$$f = -25,070X_0 - 0,77719X_1 + 24,750X_2 + 153,69$$

$$R^2 = 0,70976, S = 4,5777, F = 20,379, R_{CV}^2 = 0,61497, S_{CV} = 5,2725$$

$n = 4$  Deskriptoren: MD14<sub>9</sub>, ACLH<sub>12</sub>, ACUH<sub>21</sub>, ACUH<sub>31</sub>

$$f = -27,862X_0 + 0,82634X_1 - 6,1882X_2 + 25,663X_3 + 152,45$$

$$R^2 = 0,78011, S = 4,0666, F = 21,287, R_{CV}^2 = 0,65533, S_{CV} = 5,0913$$

$n = 5$  Deskriptoren: AUCCO<sub>12</sub>, AUCCO<sub>22</sub>, ACUH<sub>32</sub>, AUCCO<sub>34</sub>, AUCCO<sub>36</sub>

$$f = 1,2582X_0 - 90,607X_1 + 42,643X_2 - 38,316X_3 + 138,14X_4 + 153,99$$

$$R^2 = 0,90551, S = 2,7231, F = 44,085, R_{CV}^2 = 0,64012, S_{CV} = 5,3144$$

---

<sup>5</sup>Der Fall für  $n = 5$  wurde dabei mit der Funktion `regsubsets` aus dem  $R$ -Paket `leaps` berechnet.



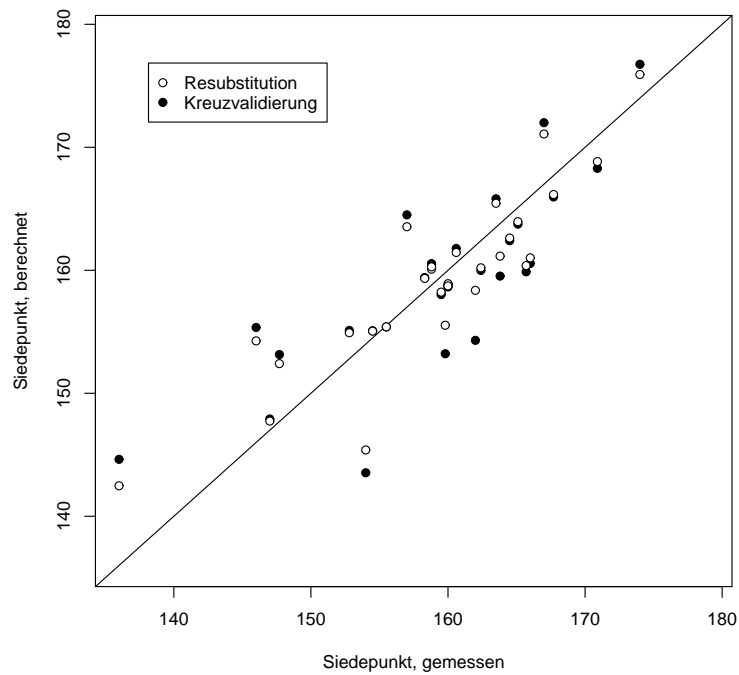


Abbildung 5.40: Scatterplot gemessener und vorhergesagter BP für das beste LM mit 4 MS-Deskriptoren

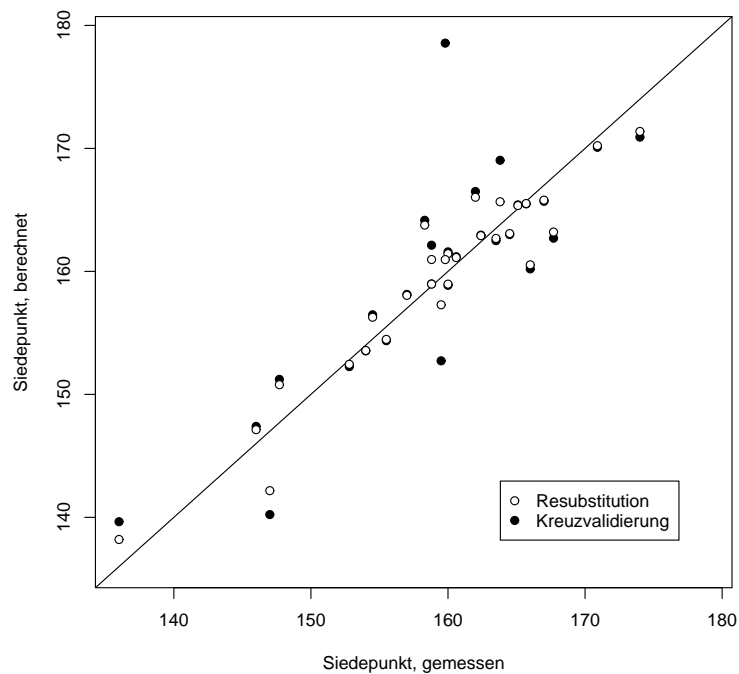


Abbildung 5.41: Scatterplot gemessener und vorhergesagter BP für das beste LM mit 5 MS-Deskriptoren

$n$	ANN, 1HN		ANN, 2HN		ANN, 3HN	
	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$
1	0,31739	0,13562	0,36684	0,16172	0,37585	0,14647
2	0,55609	0,15128	0,55655	0,31550	0,55656	0,34845
3	0,71538	0,57385	0,71531	0,57192	0,71520	0,56007
4	0,79181	0,60404	0,79213	0,60005	0,79240	0,60473
5	0,90266	0,62013	0,90463	0,62322	0,90565	0,62194

$n$	SVM, lin		SVM, pol		SVM, rad	
	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$
1	0,28778	0,00864	0,37342	0,21192	0,28385	-0,0452
2	0,54330	0,50699	0,56767	0,47356	0,59001	0,33982
3	0,70816	0,58911	0,73234	0,34815	0,59760	0,31635
4	0,77338	0,66210	0,85029	0,67688	0,76924	0,45853
5	0,76256	0,09732	0,69839	-15,669	0,70315	0,43496

Tabelle 5.18:  $R^2$  und  $R_{CV}^2$  verschiedener MS–BP–Beziehungen

Wir sehen, dass hauptsächlich Autokorrelations–Deskriptoren für die besten linearen Modelle ausgewählt werden, während logarithmische Intensitätsverhältnisse keinen Einfluss nehmen. Dies erklärt sich aus der Tatsache, dass es sich bei den Strukturen um Isomere handelt, und zudem in der Summenformel abgesehen von Kohlenstoff keine stark isotopenhaltigen Elemente auftreten.

Auffällig ist, dass  $R_{CV}^2$  für  $n = 5$  Deskriptoren kleiner ist als für  $n = 4$  Deskriptoren. Dies resultiert im Wesentlichen aus dem durch LOO–CV vorhergesagten Wert von 178,6 für Spektrum 14 in Abbildung 5.39, der somit deutlich von dem experimentellen Wert 159,8 abweicht. Abbildungen 5.40 und 5.41 zeigen Scatterplots gemessener und berechneter BP für die besten LM mit 4 und 5 Deskriptoren sowie die durch LOO–CV berechneten Werte.

### Nichtlineare Modelle

Wir wollen die durch BSS ermittelten Teilmengen von Deskriptoren zur Bestimmung weiterer, nichtlinearer Modelle heranziehen: Neuronale Netze mit einem, zwei und drei versteckten Neuronen sowie Support–Vektor–Maschinen mit linearem, polynomialem und radialem Kernel.

Zunächst berechnen wir ANN mit Startgewichten 0 (Abschnitt 3.2.2). Diese haben den Vorteil der Reproduzierbarkeit, führen aber mitunter zu schlechteren Vorhersagefunktionen als ANN mit zufällig gewählten Startgewichten. Als vorverarbeitende Maßnahme wurde eine Bereichsskalierung von Deskrip-

$n$	ANN, 1HN		ANN, 2HN		ANN, 3HN	
	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$
1	0,32699	0,29209	0,37655	0,24402	0,37698	0,08144
2	0,55610	0,62940	0,64321	0,46891	0,62491	0,51761
3	0,71375	0,67758	0,73173	0,67563	0,74387	0,55154
4	0,79057	0,84058	0,83141	0,84403	0,86065	0,75792
5	0,88069	0,93096	0,91965	0,93035	0,92811	0,91350

$n$	SVM, lin		SVM, pol		SVM, rad	
	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$	$R^2$	$R_{CV}^2$
1	0,28778	0,04655	0,37342	0,21635	0,20122	0,18359
2	0,54331	0,51416	0,60301	0,52595	0,59853	0,51139
3	0,70846	0,59146	0,73063	0,60683	0,72864	0,71497
4	0,77534	0,66717	0,83984	0,71080	0,82497	0,60157
5	0,90162	0,68102	0,33187	0,12913	0,92570	0,66596

Tabelle 5.19:  $R^2$  und  $R_{CV}^2$  verschiedener optimierter MS-BP-Beziehungen

toren- und Eigenschaftswerten vorgenommen. Tabelle 5.18 enthält  $R^2$  und  $R_{CV}^2$  für die so ermittelten Modelle bei verschiedenen Anzahlen von Deskriptoren  $n$ .

Weiterhin werden entsprechende Werte für SVM angegeben. Zur Berechnung der SVM wurde eine Autoskalierung der Variablen durchgeführt, für die Parameter `cost` und `gamma` übernehmen wir zunächst die Default-Einstellungen des  $R$ -Pakets *e1071*. Diese Einstellungen sind für manche Problemstellungen nur schlecht geeignet. Dies manifestiert sich in sehr niedrigen, teilweise sogar negativen Werten für  $R_{CV}^2$ .

Für diese Fälle bietet *e1071* die Funktion `tune` zur Suche nach besseren Parameter-Sätzen. Wir testen Parameter `cost` =  $2^i$ ,  $i \in \{-1, \dots, 7\}$  und `gamma` =  $2^j$ ,  $j \in \{-7, \dots, 3\}$  und wählen diejenigen SVM zu den Kombinationen mit kleinster  $RSS_{CV}$  aus. Tabelle 5.19 zeigt die zugehörigen Werte für  $R^2$  und  $R_{CV}^2$ .

Für ANN testen wir jeweils 10 verschiedene zufällig gewählte Startgewichtungen. Wiederum wählen wir Modelle mit kleinster  $RSS_{CV}$  aus. Die resultierenden Werte für  $R^2$  und  $R_{CV}^2$  sind ebenfalls in Tabelle 5.19 zusammengefasst. Man beachte jedoch, dass bei dieser Vorgehensweise  $R_{CV}^2$  mitunter größer als  $R^2$  wird, was hauptsächlich durch Zufallseinflüsse bedingt ist, und somit  $R_{CV}^2$  nur eingeschränkt zur Modell-Selektion herangezogen werden kann.

### Ausblick

Wir haben gesehen, dass sich über MS-Deskriptoren und statistische Lernverfahren Beziehungen zwischen Massenspektren und Siedepunkten von Decanen finden lassen. Bei anderen spektroskopischen Methoden, insbesondere NMR, ist es möglich, strukturelle Eigenschaften über molekulare Deskriptoren und statistische Lernverfahren mit spektroskopischen Daten in Beziehung zu setzen [75, 92, 93, 94], und somit vorhergesagte Peaklagen für das Ranking von Strukturkandidaten heranzuziehen.

Bei MS sind Strukturen nur schwer mit den spektralen Daten in Beziehung zu bringen. Zumindest für die Lage der Peaks kann man dies über virtuelle Fragmentierung (Abschnitt 5.4.2) erreichen. Allerdings ist allein die Lage der Peaks bei großen Strukturräumen ähnlicher Strukturen oft nicht ausreichend, um aussagekräftige Rankings zu erzielen. Zur Verbesserung muss man nach Wegen suchen, die Intensitäten der Peaks einzubeziehen.

Die Vorhersage von Peakintensitäten aus Strukturen ist, wie bereits mehrfach angesprochen, nach derzeitigem Kenntnisstand nur für kleine Substanzklassen unter Berücksichtigung ausgewählter Fragmentierungsreaktionen möglich [11, 46, 47, 62, 68, 130]. MS-Deskriptoren zeigen einen Weg, um massenspektrometrische Daten unter Berücksichtigung der Peakintensitäten für statistische Lernverfahren aufzubereiten. Im Falle von MS-Klassifikatoren (Abschnitt 5.5.2) und quantitativen MS-Eigenschafts-Beziehungen wurde dies bereits gezeigt.

Im Folgenden wollen wir drei Alternativen vorschlagen, die die Verwendung von MS-Deskriptoren zur Spektren-Verifikation aufzeigen:

- Spektren-Verifikation durch *quantitative Spektren-Struktur Beziehungen* (Abbildung 5.42): Hier werden ausgehend von einer Datenbasis aufgeklärter Spektren Modelle zur Vorhersage molekularer Deskriptoren aus Spektren über spektrale Deskriptoren und Regression ermittelt. Sind solche Modelle gefunden, können für ein unbekanntes Spektrum Werte molekularer Deskriptoren der gesuchten Struktur vorhergesagt, und mit den Deskriptorenwerten von Strukturkandidaten verglichen werden. Bei  $n$  verwendeten molekularen Deskriptoren erhält derjenige Strukturkandidat die höchste Ranking-Position, der den kürzesten Abstand zu den vorhergesagten Werten molekularer Deskriptoren im  $R^n$  aufweist.
- Spektren-Verifikation durch *quantitative Struktur-Spektren-Beziehungen* (Abbildung 5.43): Dabei ermittelt man ausgehend von einer Datenbasis aufgeklärter Spektren Modelle zur Vorhersage spektraler Deskriptoren aus Strukturformeln über molekulare Deskriptoren und Regressi-

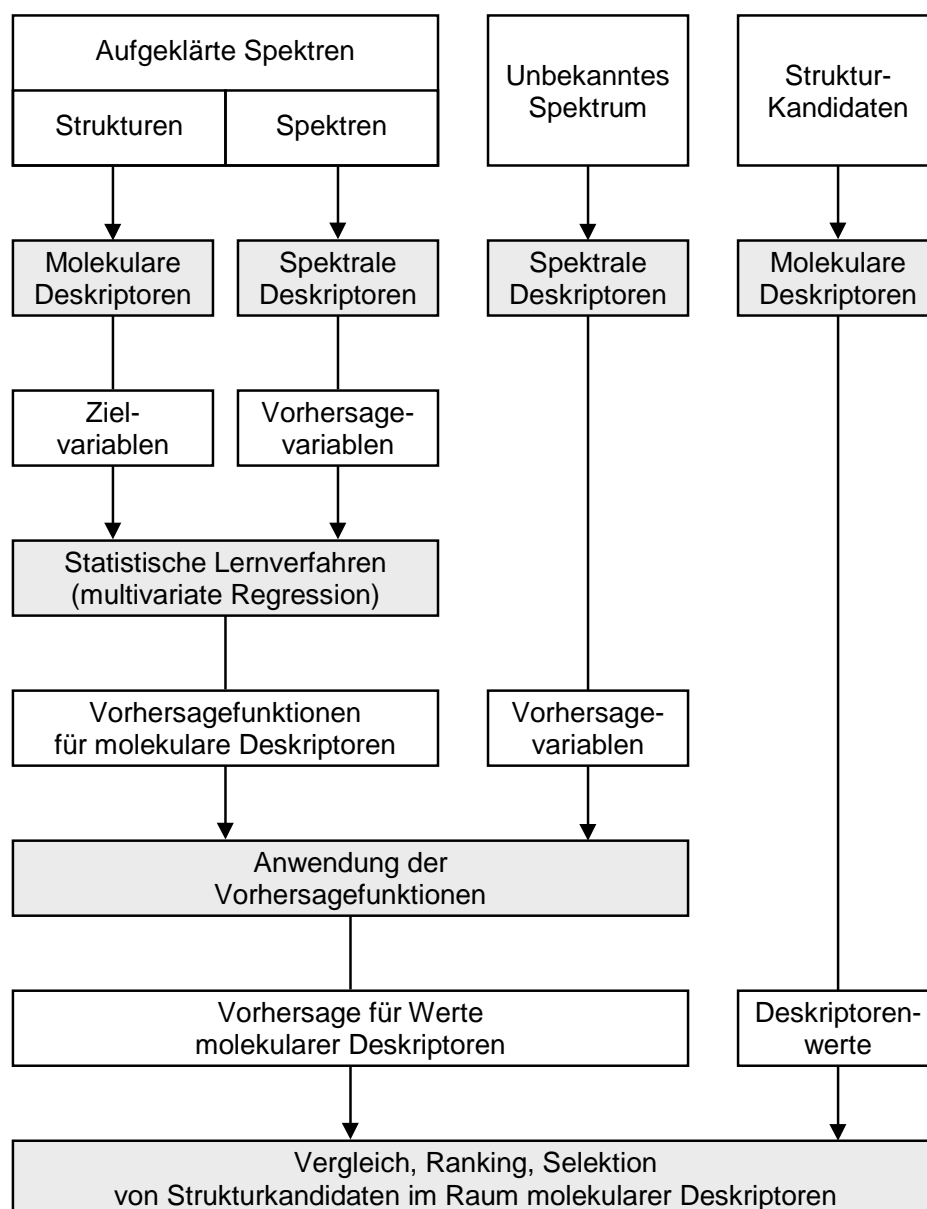


Abbildung 5.42: Vorgehensweise bei der Spektren-Verifikation durch quantitative Spektren-Struktur-Beziehungen

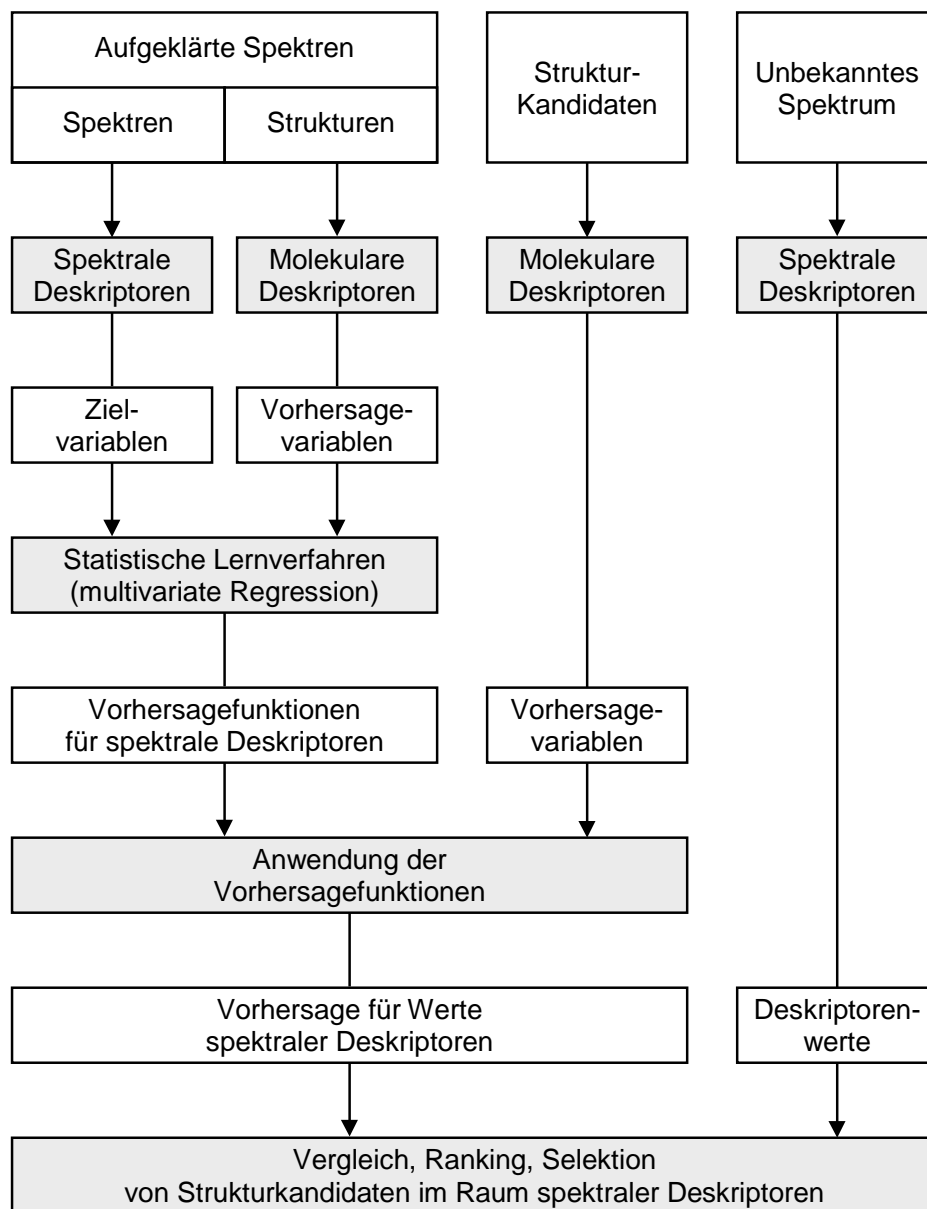


Abbildung 5.43: Vorgehensweise bei der Spektren-Verifikation durch quantitative Struktur-Spektren-Beziehungen

on. Sind solche Modelle gefunden, können für einen Strukturkandidaten Werte spektraler Deskriptoren vorhergesagt, und mit den tatsächlichen Deskriptorenwerten des gemessenen unbekanntes Spektrums verglichen werden. Bei  $n$  verwendeten spektralen Deskriptoren erhält derjenige Strukturkandidat die höchste Ranking-Position, der den kürzesten Abstand zu den vorhergesagten Werten spektraler Deskriptoren im  $R^n$  aufweist.

- Spektren-Verifikation durch *quantitative Struktur* × *Spektren-Kompatibilitäts-Beziehungen* (Abbildung 5.44): Bei dieser Methode wird zunächst das kartesische Produkt aus Spektren und Strukturen einer Datenbasis aufgeklärter Spektren gebildet. Paare aus zusammengehörigem Spektrum und Struktur erhalten den Wert 1, andere den Wert -1 zugewiesen. Diese Werte bilden die Zielvariable für ein statistisches Lernverfahren. Als Vorhersagevariable werden spektrale *und* molekulare Deskriptoren zu den Spektrum-Struktur-Paaren herangezogen. Wir erhalten eine Vorhersagefunktion, die für ein Paar aus einem unbekanntem Spektrum und einem Strukturkandidaten eine Prognose über die Kompatibilität dieses Paares erlaubt. Auf diese Weise können Vergleichswerte für Strukturkandidaten ermittelt werden.

Die vorgeschlagen Verfahren kombinieren drei mächtige Werkzeuge zur Spektren- und Struktur-Vorhersage,

- molekulare Deskriptoren,
- spektrale Deskriptoren und
- statistische Lernverfahren,

und wurden nach Wissen des Autors in dieser Weise noch nicht angewendet. Natürlich ist die Entwicklung dieser Verfahren mit einem beträchtlichen Forschungsaufwand verbunden, der den Rahmen der vorliegenden Arbeit endgültig sprengen würde. So muss untersucht werden, welche Deskriptoren als Ziel-, welche als Vorhersagevariable geeignet sind. Es wäre denkbar, neue molekulare Deskriptoren heranzuziehen, die etwa Informationen über die Fragmentierung einfließen lassen. Es muss geklärt werden, welche Norm auf  $R^n$  zum Ranking bestmögliche Ergebnisse liefert.

Insbesondere ist fraglich, ob eine große Datenbasis, wie etwa in Abschnitt 5.3.5 verwendet werden soll, oder ob es erfolgreicher ist, spezialisierte Modelle zu entwickeln, die auf das Ranking von Konstitutionsisomeren beschränkt sind. Die Bestimmung der Bruttoformel kann, wie wir im nächsten Abschnitt zeigen werden, schon jetzt durch Einsatz moderner massenspektrometrischer Hardware unter bestimmten Nebenbedingungen als gelöst betrachtet werden.

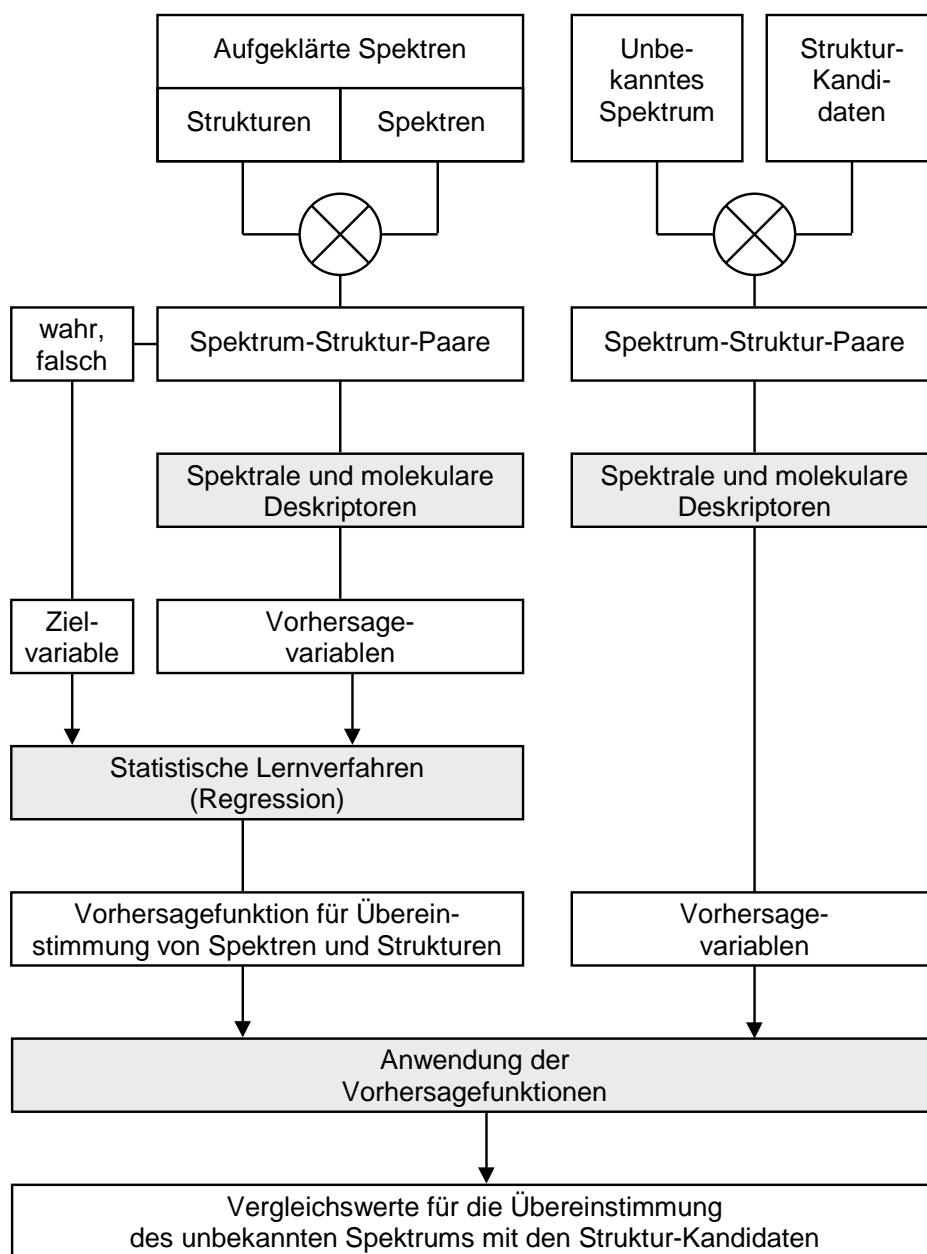


Abbildung 5.44: Vorgehensweise bei der Spektren-Verifikation durch quantitative Struktur×Spektren-Kompatibilitäts-Beziehungen



## 5.8 Hochauflösende Massenspektrometrie

Sowohl zur Interpretation als auch zur Verifikation von LR-MS wurden in den vorangegangenen Abschnitten Methoden getestet, die dem aktuellen Kenntnisstand der Wissenschaft entsprechen. Für eine automatische Strukturaufklärung sind die Ergebnisse jedoch noch nicht ausreichend. Dennoch können die vorgestellten Methoden zumindest eine Hilfestellung im Rahmen einer computerunterstützten Strukturaufklärung bieten, und als Ausgangsbasis für die weitere Forschung auf diesem Gebiet eingesetzt werden.

Die Lösung muss im Zusammenspiel mit anderen (massen-)spektroskopischen Methoden gesucht werden. Weiche Ionisierungsarten gewährleisten die Anwesenheit des Molekülions im Spektrum und erlauben somit die Bestimmung der Molekülmasse. Hochauflösende Massenspektrometrie liefert Massenzahlen mit einer Genauigkeit von mehreren Dezimalstellen. Damit kann die Summenformel oft auf wenige Kandidaten eingeschränkt, oder meist sogar eindeutig bestimmt werden. Auch hinsichtlich der Vergleichswert-Berechnung für Strukturformeln kann sich eine höhere Auflösung der Fragmentmassen positiv auswirken. Leider sind derzeit (noch) keine größeren Datenbestände an hochaufgelösten Massenspektren verfügbar, um dies statistisch zu belegen. Nicht zuletzt drängt sich Tandem-MS geradezu auf für eine rechnerunterstützte Auswertung und nährt die Hoffnung auf Fortschritte in der automatischen Strukturaufklärung via massenspektrometrischer Verfahren. Leider befinden sich Datenbanken zu den genannten verfeinerten massenspektrometrischen Methoden erst im Aufbau. Wir wollen dieses Kapitel mit einer theoretischen Betrachtung zu der Leistungsfähigkeit und den Grenzen hochauflösender Massenspektrometrie beschließen.

### Exakte Isotopenmassen

Tabelle 5.20 enthält *hochaufgelöste Isotopenmassen* und Isotopenverteilungen für die Elemente aus  $\mathcal{E}_{11}$  (vgl. Tabelle 5.2). Dabei bezeichnet  $m_{X,i}$  die exakte Masse des Isotops von  $X$ , das sich um  $i$  ganzzahlige Masseneinheiten von  $\tilde{m}_X$  unterscheidet, und  $I_{X,i}$  dessen relative Häufigkeit. Isotope mit relativer Häufigkeit unter 0,001 wurden in dieser Tabelle nicht berücksichtigt. Dies betrifft  ${}^2\text{H}$  mit  $m_{\text{H},1}=2,014102$  amu,  $I_{\text{H},1}=0,00015$  und  ${}^{36}\text{S}$  mit  $m_{\text{S},4}=35,967079$  amu,  $I_{\text{S},4}=0,00020$ . Aus den exakten Isotopenmassen berechnet man die gebräuchliche Größe der mittleren Atommasse eines chemischen Elements. Dabei werden die exakten Massen der Isotope gemäß ihrer Häufigkeit gewichtet:

$X$	$m_{X,0}$	$I_{X,0}$	$m_{X,1}$	$I_{X,1}$	$m_{X,2}$	$I_{X,2}$
H	1,007825	1,0000				
C	12,000000	0,9890	13,003355	0,0110		
N	14,003074	0,9963	15,000109	0,0037		
O	15,994915	0,9976	16,999131	0,0004	17,999159	0,0020
F	18,998403	1,0000				
Si	27,976928	0,9223	28,976496	0,0467	29,973772	0,0310
P	30,973763	1,0000				
S	31,972072	0,9504	32,971459	0,0075	33,967868	0,0421
Cl	34,968853	0,7577	36,965903	0,2423		
Br	78,918336	0,5069	80,916290	0,4931		
I	126,904477	1,0000				

Tabelle 5.20: Hochaufgelöste Isotopenmassen und Isotopenverteilungen für die Elemente aus  $\mathcal{E}_{11}$

### 5.8.1 Definition:

Für ein chemisches Element  $X \in \mathcal{E}$  ist seine *mittlere Atommasse* definiert als

$$\bar{m}_X := \sum_i m_{X,i} I_{X,i}.$$

Diese spielt jedoch im Bereich der Massenspektrometrie keine Rolle. Wir wollen im Folgenden untersuchen, inwiefern sich durch genauere Massenangaben die Kandidatenmengen für die Summenformel einschränken lassen. Des weiteren interessiert uns die Frage, unter welchen Bedingungen es möglich ist, die Bruttoformel eindeutig anhand ihrer exakten Masse zu bestimmen und mit welcher Genauigkeit dazu die Molekülmasse gemessen werden muss.

### Bruttoformeln mit identischen exakten Massen

Präzise hochauflösende Massenspektrometer können auf bis zu 6 Dezimalstellen genaue Massenangaben liefern. Zunächst wollen wir hier feststellen, dass damit die Summenformel generell nicht eindeutig bestimmt werden kann. So treten schon bei Masse 129,954034 die ersten Paare von Bruttoformeln in  $\mathcal{B}_{\mathcal{E}_{11}}^C$  mit gleichen Molekülmassen auf. Einige solcher Paare von Bruttoformeln gleicher Masse sind:

$$\begin{aligned}
 m_{\text{H}_4\text{F}_2\text{SSi}_2} &= m_{\text{HClO}_2\text{P}_2} &= 129,954034 \\
 m_{\text{CH}_4\text{F}_2\text{SSi}_2} &= m_{\text{CHClO}_2\text{P}_2} &= 141,954034 \\
 m_{\text{H}_3\text{F}_2\text{NSSi}_2} &= m_{\text{CINO}_2\text{P}_2} &= 142,949283 \\
 m_{\text{CH}_6\text{F}_2\text{SSi}_2} &= m_{\text{CH}_3\text{ClO}_2\text{P}_2} &= 143,969684 \\
 m_{\text{H}_5\text{F}_2\text{NSSi}_2} &= m_{\text{H}_2\text{CINO}_2\text{P}_2} &= 144,964933
 \end{aligned}$$

Natürlich könnte man einwenden, dass keine chemischen Verbindungen mit solchen Bruttoformeln existieren. Jedoch erhält man beispielsweise durch Hinzufügen gleichvieler C- oder H-Atome wiederum Paare von Bruttoformeln gleicher Masse. Die ersten beiden Bruttoformeln gleicher Masse mit mindestens 5 C- und 10 H-Atomen sind  $C_5H_{14}F_2SSi_2$  und  $C_5H_{11}ClO_2P_2$  bei 200,032284 amu.

### Massendifferenzen zwischen Bruttoformeln

Nun stellt man sich die Frage, wie dicht im Allgemeinen die exakten nominalen Massen von Bruttoformeln zusammenhängender molekularer Graphen liegen. Dazu generieren wir alle Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_{11}}^C$  mit exakter Masse bis  $m = 500$ : Es gibt  $n = 84204768$  solcher Bruttoformeln. Wir sortieren die exakten Massen dieser Bruttoformeln aufsteigend

$$m_i \leq m_{i+1}, \quad i \in n - 1$$

und ordnen sie ganzzahligen Massenbereichen zu:

$$\Omega_j := \{i \in n \mid m_i \in [j, j + 1[ \}, \quad j \in m.$$

Dann ermitteln wir die Differenzen aufeinander folgender Massen und bilden deren Minima, Maxima und arithmetische Mittel für die ganzzahligen Massenbereiche  $j \in m$  mit  $\Omega_j \neq \emptyset$ :

$$\begin{aligned} \text{Min}MD_j &:= \min\{m_i - m_{i-1} \mid i \in \Omega_j\}, \\ \text{Max}MD_j &:= \max\{m_i - m_{i-1} \mid i \in \Omega_j\}, \\ \text{Mean}MD_j &:= |\Omega_j|^{-1} \sum_{i \in \Omega_j} m_i - m_{i-1} \\ &= |\Omega_j|^{-1} (m_{\max \Omega_j} - m_{\min \Omega_j - 1}). \end{aligned}$$

Darüber hinaus interessiert uns, wie weit für jede Bruttoformel die nächstgelegene exakte Masse weiterer Bruttoformeln entfernt ist. Zu diesem Zweck definieren wir die *minimalen Massendifferenzen* (kurz *MMD*)

$$MMD_i := \min\{m_{i+1} - m_i, m_i - m_{i-1}\}, \quad 1 \leq i < n - 1$$

Auch für die *MMD* bilden wir wieder Minima, Maxima und arithmetische Mittel über ganzzahlige Massenbereiche:

$$\begin{aligned} \text{Min}MMD_j &:= \min\{MMD_i \mid i \in \Omega_j\}, \\ \text{Max}MMD_j &:= \max\{MMD_i \mid i \in \Omega_j\}, \\ \text{Mean}MMD_j &:= |\Omega_j|^{-1} \sum_{i \in \Omega_j} MMD_i. \end{aligned}$$

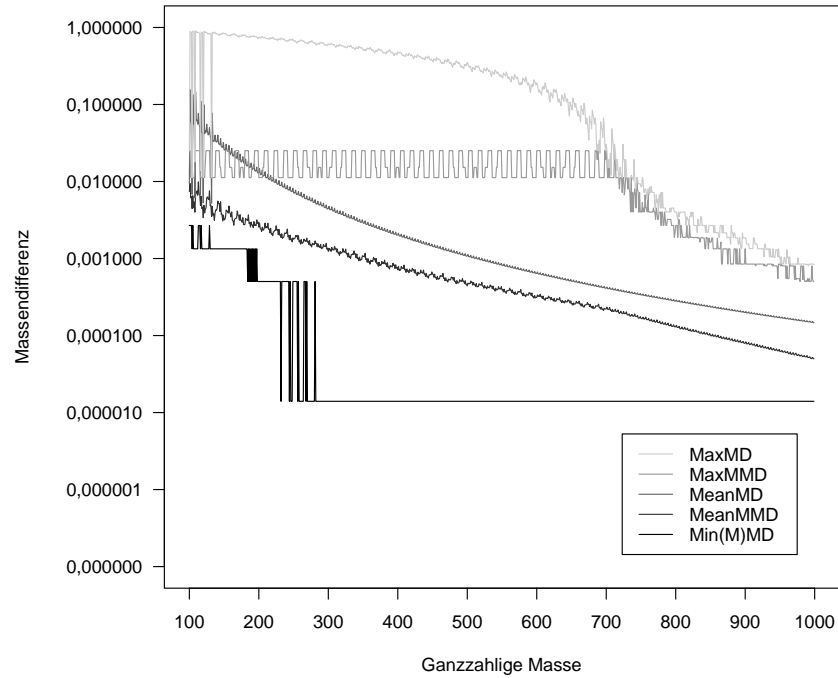


Abbildung 5.45: Minima, Maxima und arithmetische Mittel der Massendifferenzen für Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_4}^C$

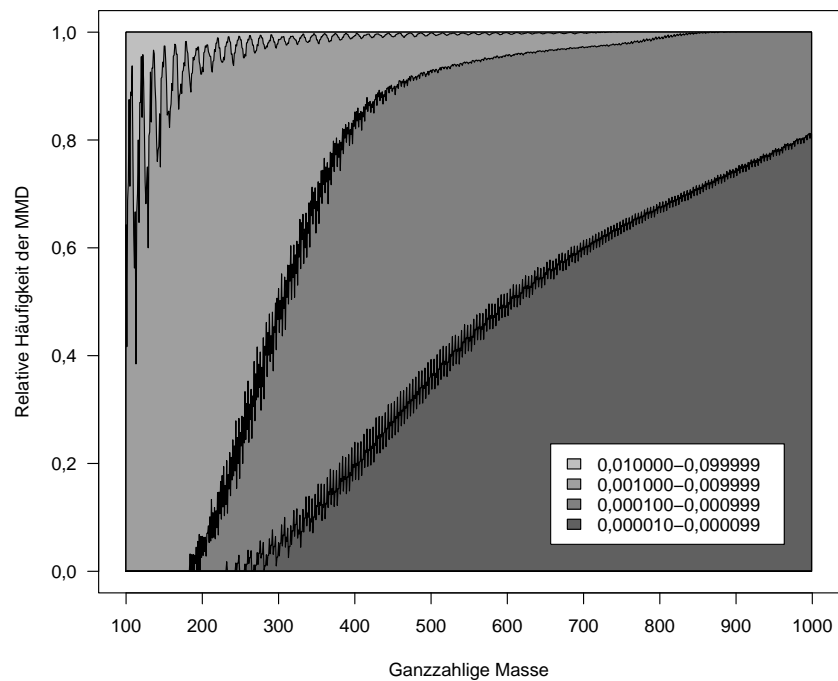


Abbildung 5.46: Relative Häufigkeiten der MMD für Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_4}^C$

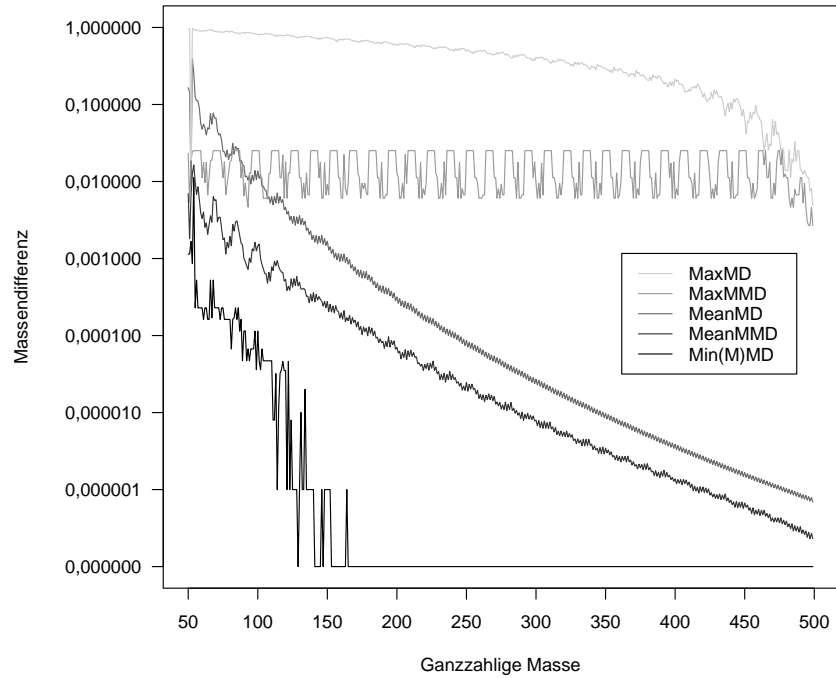


Abbildung 5.47: Minima, Maxima und arithmetische Mittel der Massendifferenzen für Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_{11}}^C$

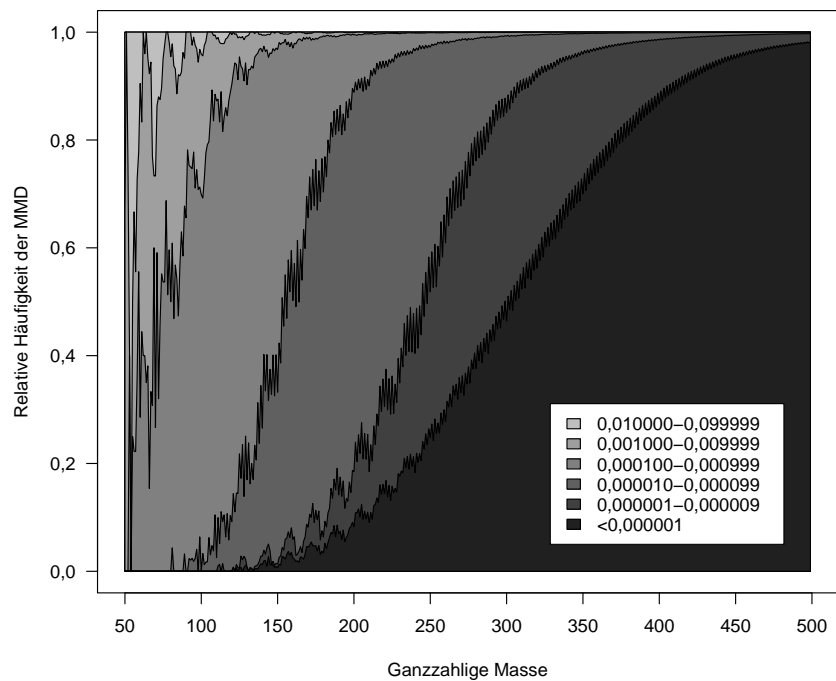


Abbildung 5.48: Relative Häufigkeiten der MMD für Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_{11}}^C$

So kann man die Dichte exakter Massen von Bruttoformeln in Abhängigkeit zu ganzzahligen Masseneinheiten darstellen. Entsprechendes führen wir auch für Bruttoformeln aus  $\mathcal{B}_{\mathcal{E}_4}^C$  mit einer exakten Masse bis  $m = 1000$  durch ( $n = 1737956$  Bruttoformeln).

Abbildungen 5.45 und 5.47 visualisieren  $MinMD_j$ ,  $MaxMD_j$ ,  $MeanMD_j$ ,  $MinMMD_j$ ,  $MaxMMD_j$  und  $MeanMMD_j$  in Abhängigkeit von ganzzahligen Massen  $j \geq 100$ . In diesem Bereich sind  $MinMD_j$  und  $MinMMD_j$  identisch, und werden somit durch die gleiche Kurve dargestellt.

$MinMD$  und  $MinMMD_j$  sind für uns von besonderem Interesse. Sie geben Auskunft wie nahe die exakten Massen von Bruttoformel-Kandidaten beieinander liegen können. Daraus kann man auf die Messgenauigkeit schließen, die notwendig ist, um die Bruttoformel anhand der experimentell ermittelten exakten Masse stets eindeutig identifizieren zu können.

Für Bruttoformeln zu Elementen aus  $\mathcal{E}_4$  stellen wir fest, dass  $MinMMD_j$  einen Wert von 0,000014 amu nicht unterschreiten. Basierend auf den Elementen aus  $\mathcal{E}_4$  genügt es also, die Masse des Moleküls mit einer Fehlertoleranz kleiner  $\pm 0,000007$  zu messen, um die Bruttoformel immer eindeutig bestimmen zu können. Dies ist erfüllt, wenn die Masse auf 5 Dezimalstellen genau ermittelt werden kann.

Anders stellt sich die Situation bei Berücksichtigung aller Elemente aus  $\mathcal{E}_{11}$  dar. Dann ist  $MinMMD_j$  bereits für  $j = 129$  erstmals Null, da in dem Massenintervall  $[129, 130[$  Bruttoformeln mit identischer exakter Masse existieren. Aber natürlich existieren auch in diesem Fall Bruttoformeln, für die es nicht leere Massenintervalle gibt, in denen sich keine weiteren Bruttoformeln befinden. Abbildung 5.48 zeigt die relativen Häufigkeiten der MMD aufgeteilt in Intervalle  $[10^{-d}, 10^{-d+1}]$ ,  $d = 0, \dots, 6$ . So haben bei einer ganzzahligen Masse von 300 amu noch über die Hälfte der Bruttoformeln eine  $MMD_i \geq 0,000001$ . Abbildung 5.46 zeigt die relativen Häufigkeiten der MMD für Bruttoformeln mit aus  $\mathcal{B}_{\mathcal{E}_4}^C$ .

### Bestimmung von Bruttoformel-Kandidaten anhand exakter Molekülmassen

Doch was bedeuten diese Erkenntnisse für die Anzahl von Kandidaten für die Bruttoformel bei verschiedenen Genauigkeiten von Massenspektrometern? Wenn die Massenauflösung — oder sagen wir besser der Messfehler für die exakte Molekülmasse einer Summenformel mit Masse  $m_i$  kleiner als die Hälfte des Abstands zu den jeweils benachbarten Massen weiterer Kandidaten mit benachbarten Massen  $m_{i-1}$  und  $m_{i+1}$  ist, kann die Summenformel eindeutig bestimmt werden. Wir illustrieren diese Tatsache anhand zweier Stichproben von Substanzen aus unserer MS-Datenbasis (Abschnitt 5.3.5). Dazu wurden

	$d$	Min.	1. Quart.	Median	Mittel	3. Quart.	Max.
$\mathcal{E}_4$	0	6	60,00	107,0	230,759	254,00	3355
	1	1	19,00	37,0	52,089	64,25	504
	2	1	2,00	4,0	5,398	6,00	49
	3	1	1,00	1,0	1,232	1,00	5
	4	1	1,00	1,0	1,012	1,00	2
	5	1	1,00	1,0	1,000	1,00	1
	6	1	1,00	1,0	1,000	1,00	1
$\mathcal{E}_{11}$	0	19	2072,00	8421,0	670560,053	75144,00	170389109
	1	1	158,75	492,0	35464,212	3181,00	16064782
	2	1	11,00	47,5	3579,303	281,50	1610209
	3	1	2,00	6,0	355,120	28,25	159045
	4	1	1,00	1,0	36,673	4,00	16140
	5	1	1,00	1,0	4,747	1,00	1738
	6	1	1,00	1,0	1,368	1,00	153

Tabelle 5.21: Anzahlen von Bruttoformel-Kandidaten für eine Stichprobe von Verbindungen aus  $\mathcal{E}_4$  bzw.  $\mathcal{E}_{11}$

1000 Verbindungen per Zufall ausgewählt, die sich nur aus Elementen von  $\mathcal{E}_4$  bzw.  $\mathcal{E}_{11}$  zusammensetzen.

Es wurden jeweils die exakten nominalen Molekülmassen  $m$  berechnet, auf  $d = 0, \dots, 6$  Dezimalstellen gerundet, und dann alle Bruttoformeln mit Elementen aus  $\mathcal{E}_4$  bzw.  $\mathcal{E}_{11}$  berechnet, die von der gerundeten Masse um maximal  $0,5 \cdot 10^{-d}$  abweichen. Tabelle 5.21 fasst die Ergebnisse für die Anzahlen von Kandidaten zusammen.

Insbesondere sind wir natürlich daran interessiert in wie vielen Fällen die Bruttoformel eindeutig bestimmt werden kann. Folgende Tabelle gibt die Anzahlen einelementiger Kandidatenmengen an:

$\mathcal{E}$	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$
$\mathcal{E}_4$	8	134	812	988	1000	1000
$\mathcal{E}_{11}$	10	78	213	562	835	938

In Übereinstimmung mit unseren vorangegangenen Betrachtungen zu den MMD bei  $\mathcal{E}_4$  ist ab einer Genauigkeit von  $d \geq 5$  Dezimalstellen eine eindeutige Bestimmung der Bruttoformel sichergestellt. Für  $\mathcal{E}_{11}$  gibt es selbst für  $d = 6$  Dezimalstellen Beispiele in unserer Stichprobe, für die die Bruttoformel nicht eindeutig ermittelt werden kann. Im extremen Fall liegen 153 Bruttoformeln innerhalb des relevanten Intervalls für die exakte Molekülmasse. Abbildungen 5.49 und 5.51 zeigen Boxplots (mit `range = 5`) der Anzahlen

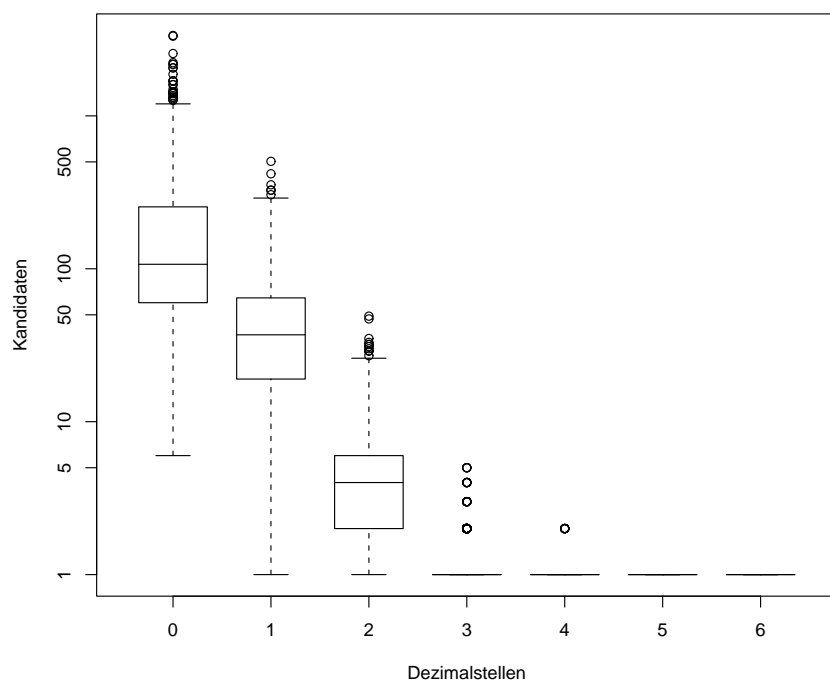


Abbildung 5.49: Boxplot der Anzahlen von Bruttoformel-Kandidaten für eine Stichprobe von Verbindungen aus  $\mathcal{E}_4$

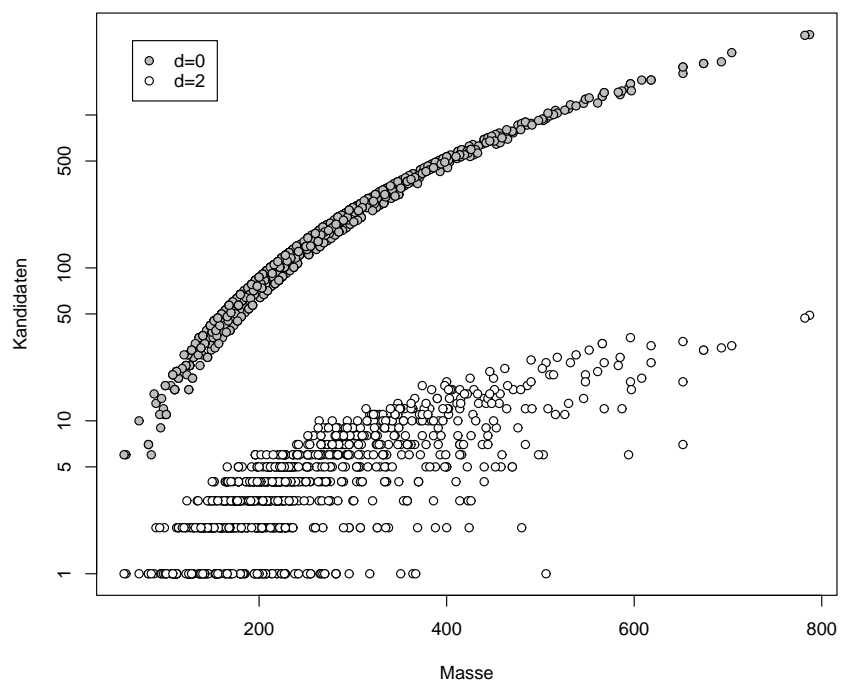


Abbildung 5.50: Plot der Anzahlen von Bruttoformel-Kandidaten und der Molekülmasse für eine Stichprobe von Verbindungen aus  $\mathcal{E}_4$



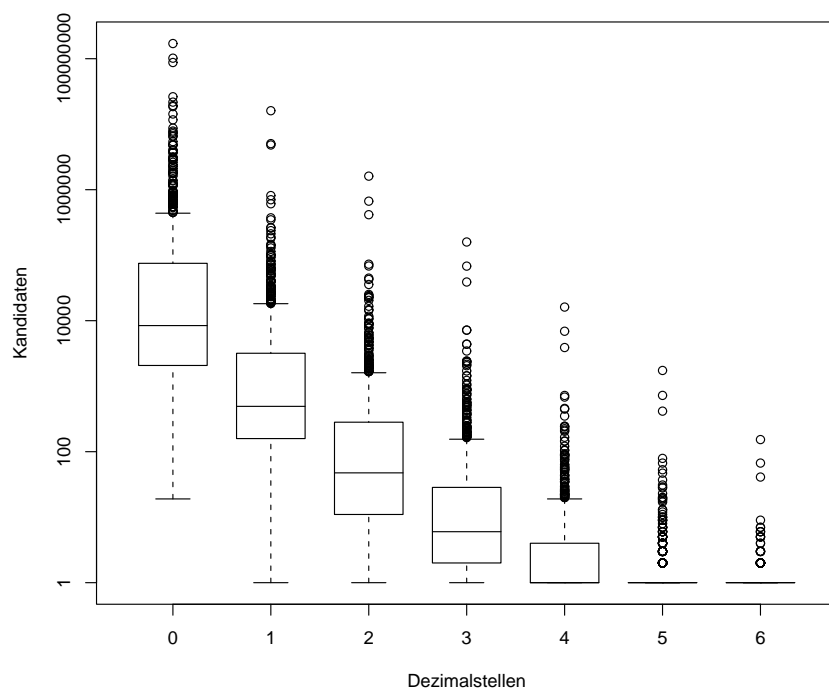


Abbildung 5.51: Boxplot der Anzahlen von Bruttoformel-Kandidaten für eine Stichprobe von Verbindungen aus  $\mathcal{E}_{11}$

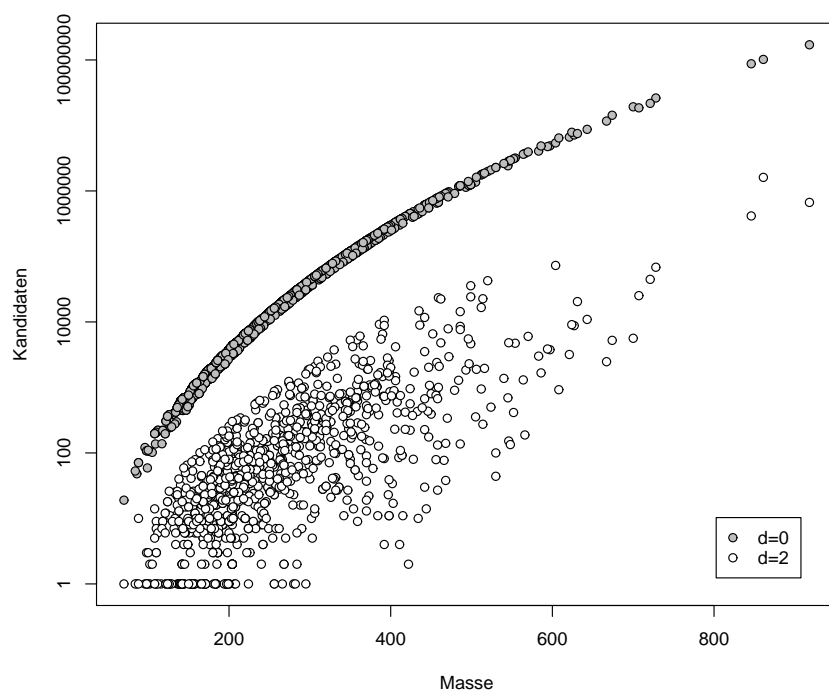


Abbildung 5.52: Plot der Anzahlen von Bruttoformel-Kandidaten und der Molekülmasse für eine Stichprobe von Verbindungen aus  $\mathcal{E}_{11}$

von Bruttoformel-Kandidaten für die beiden Stichproben.

Verständlicherweise ist die Anzahl der Bruttoformel-Kandidaten abhängig von den Molekülmassen in unseren Stichproben. Zunächst wollen wir uns eine Übersicht der Molekülmassen in den beiden Stichproben verschaffen:

$\mathcal{E}$	Min.	1. Quart.	Median	Mittel	3. Quart.	Max.
$\mathcal{E}_4$	56,03	180,15	226,63	258,42	310,24	787,30
$\mathcal{E}_{11}$	70,08	184,04	233,12	265,18	328,16	918,43

Wir sehen, dass in allen Kennwerten die Stichprobe mit Elementen aus  $\mathcal{E}_{11}$  größere Werte annimmt als die, welche auf Elemente aus  $\mathcal{E}_4$  beschränkt ist. Dies rührt daher, dass Elemente aus  $\mathcal{E}_{11} \setminus \mathcal{E}_4$  schwerer sind als die aus  $\mathcal{E}_4$  und spiegelt sich in den Molekülmassen der Stichproben wieder. Abbildungen 5.50 und 5.52 zeigen Scatterplots, in denen die Anzahl von Bruttoformel-Kandidaten in Abhängigkeit von der Molekülmasse dargestellt wird. Wir betrachten dabei exemplarisch die Genauigkeiten von  $d = 0$  und  $d = 2$  Dezimalstellen. Man sieht dabei sowohl das exponentielle Wachstum der Kandidaten-Anzahl mit zunehmender Masse als auch die Verringerung der Kandidaten-Anzahl bei Berücksichtigung von 2 im Vergleich zu 0 Dezimalstellen.

In der Praxis wird man sich oft nicht auf Elemente aus  $\mathcal{E}_4$  beschränken können, aber es wird auch nicht nötig sein, alle Bruttoformeln mit Elementen aus  $\mathcal{E}_{11}$  zu berücksichtigen. Vielmehr wird man die Anzahl von Atomen für die einzelnen Elemente einschränken können. Dann kann man anhand der oben beschriebenen Vorgehensweise entscheiden, welche Messgenauigkeit ein Massenspektrometer gewährleisten muss, um die Bruttoformel eines Analyten eindeutig anhand der Molekülmasse bestimmen zu können.

# Kapitel 6

## Patentwesen in der Chemie

In den vorangegangenen beiden Kapiteln wurden Methoden vorgestellt, die die chemische Synthese und Analyse unterstützen. In Kapitel 4 wurde demonstriert, wie mathematische Verfahren die Generierung virtueller kombinatorischer Bibliotheken ermöglichen, und die Suche nach chemischen Verbindungen mit bestimmten gewünschten physiko-chemischen oder biologisch-pharmazeutischen Eigenschaften erleichtern. In Kapitel 5 haben wir gesehen, wie die Struktur chemischer Verbindungen identifiziert werden kann. All diese Aufgaben sind klar dem *naturwissenschaftlich-technischen* Problemkreis zuzuordnen.

Die beschriebenen Methoden zur Darstellung und Generierung molekularer Strukturen können aber auch auf *ökonomisch-juristisch* ausgerichtete Fragestellungen der Chemie angewendet werden. Wurde eine neue chemische Verbindung gefunden, deren Eigenschaften sich als nützlich erweisen, und die gewinnbringend produziert und vertrieben werden kann, so ist der Erfinder oder Entwickler daran interessiert, dieses Know-How rechtlich schützen zu lassen. Diesen Schutz kann er sich über ein Patent garantieren lassen. Neben technischen Spezifikationen zur Synthese und Anwendung ist insbesondere die Strukturformel der neuen Verbindung wichtiger Bestandteil eines solchen Patents.

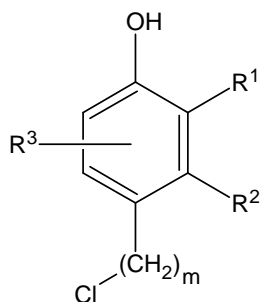
Oft besitzen chemische Verbindungen, deren Strukturen sich nur geringfügig unterscheiden auch annähernd identische Eigenschaften. Um einen wirksamen Patentschutz zu ermöglichen, wird dem Erfinder gestattet, nicht nur eine einzige Strukturformel anzugeben, sondern einen Strukturraum zu spezifizieren, der ähnliche Strukturen einschließt. Solche Strukturräume werden in Form von generischen Strukturformeln und *Markush-Formeln* (siehe [159], Kapitel 12–23) beschrieben.

Insbesondere liegt es natürlich im Interesse des Patentinhabers, den für ihn geschützten Strukturraum so groß wie möglich abzustecken. Patentämter

und –anwälte sind mit dem Problem konfrontiert, bei neuen Patentanträgen etwaige Überschneidungen mit bestehenden Patenten festzustellen. Wegen der zuweilen immensen Größe der Strukturräume und insbesondere wegen der ständig wachsenden Anzahl bestehender Patente drängt sich für diese Aufgabe eine Rechnerunterstützung auf. Zwar existieren bereits Datenbanken chemischer Patente, die Recherchemöglichkeiten sind aber in erster Linie textbasiert und auf struktureller Ebene von starker manueller Interaktion geprägt.

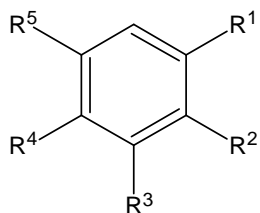
### Beispiele generischer Strukturformeln

Wir wollen zwei einfache Beispiele generischer Strukturformeln aufzeigen und ein nahe liegendes Verfahren zum Vergleich der beiden Strukturräume demonstrieren [74]. Bereits in Abschnitt 2.3 wurde der durch die generische Strukturformel



R<sup>1</sup> : CH<sub>3</sub> oder C<sub>2</sub>H<sub>5</sub>  
 R<sup>2</sup> : Alkyl (1–6 C–Atome)  
 R<sup>3</sup> : NH<sub>2</sub>  
 m : 1–3

beschriebene Strukturraum  $\mathcal{L}_1$  generiert. Wir wollen überprüfen, ob sich  $\mathcal{L}_1$  mit dem durch folgende generische Strukturformel definierten Strukturraum  $\mathcal{L}_2$  überschneidet:



R<sup>1</sup> : CH<sub>3</sub>, C<sub>2</sub>H<sub>5</sub>, OH  
 R<sup>2</sup> : Alkyl (1–6 C–Atome)  
 R<sup>3</sup> : OH, OCH<sub>3</sub>, OC<sub>2</sub>H<sub>5</sub>, CH<sub>3</sub>, C<sub>2</sub>H<sub>5</sub>  
 R<sup>4</sup> : OH, CH<sub>2</sub>Cl, NH<sub>2</sub>  
 R<sup>5</sup> : H, CH<sub>3</sub>, C<sub>2</sub>H<sub>5</sub>, NH<sub>2</sub>

Hierbei kommt lediglich Variation von Substituenten zum Einsatz, allerdings an fünf Positionen. Ebenso wie im zuvor gezeigten Beispiel handelt es sich bei dieser generischen Strukturformel (nach Stand unseres Wissens) nicht um ein real existierendes Patent. Sie wurde hier lediglich zu Demonstrationszwecken entworfen.

## Generierung der Strukturräume

Zur Generierung von  $\mathcal{L}_1$  mussten zunächst die 33 Alkylreste mit 1–6 C-Atomen bereitgestellt werden. Insgesamt erhielten wir

$$|\mathcal{L}_1| = 396$$

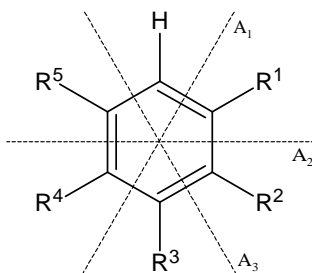
Strukturen. In diesem Fall entspricht die Mächtigkeit des Strukturraums  $|\mathcal{L}_1| = 2 \cdot 33 \cdot 2 \cdot 3$  dem Produkt der Anzahl von Variationsmöglichkeiten für die einzelnen Substituenten. Anders verhält es sich bei der zweiten generischen Strukturformel: *MOLGEN-COMB* generiert

$$|\mathcal{L}_2| = 5939$$

Strukturen, obwohl es  $3 \cdot 33 \cdot 5 \cdot 3 \cdot 4 = 5940$  Kombinationen von Substituenten gibt. Aufgrund der Symmetrie des Benzol-Grundgerüsts tritt jedoch eine Strukturformel zweimal auf.

## Identifikation der Dublette

Das Benzol-Grundgerüst der zweiten generischen Strukturformel hat unter Vernachlässigung der Substituenten und unter Berücksichtigung der Einfach- und Doppelbindungen  $S_3$  als Symmetriegruppe. Wir wollen überlegen, welche Belegungen der Substituenten zu der Dublette führen. Geometrisch veranschaulicht umfasst  $S_3$  neben der Identität zwei Drehungen und drei Achsenspiegelungen:



Wir untersuchen für welche Belegungen der Substituenten unsere generische Strukturformel einen nicht trivialen Automorphismus besitzt. Den entscheidenden Hinweis liefert der einzige Substituent, der nicht variiert werden kann, das H-Atom. Die Drehung um  $120^\circ$  scheidet aus, da sich unter den möglichen Substituenten für R<sup>4</sup> kein einzelnes H-Atom befindet. Mit analogen Argumenten kann man auch die Drehung um  $240^\circ$  und die Achsenspiegelungen an A<sub>1</sub> und A<sub>2</sub> ausschließen. Verbleibt die Spiegelung an A<sub>3</sub> als einziger möglicher nicht trivialer Automorphismus. Somit ist  $R^5 = H$  für unsere Dublette festgelegt. Wir bilden die Durchschnitte der möglichen Substituenten für R<sup>1</sup> und

$R^4$  sowie  $R^2$  und  $R^3$ :

$$\begin{aligned} \{\text{CH}_3, \text{C}_2\text{H}_5, \text{OH}\} \cap \{\text{OH}, \text{CH}_2\text{Cl}, \text{NH}_2\} &= \{\text{OH}\} \\ \{\text{Alkyle mit 1-6 C}\} \cap \{\text{OH}, \text{OCH}_3, \text{OC}_2\text{H}_5, \text{CH}_3, \text{C}_2\text{H}_5\} &= \{\text{CH}_3, \text{C}_2\text{H}_5\} \end{aligned}$$

Die beiden isomorphen Strukturen der Dublette erhält man also für  $R^5 = \text{H}$ ,  $R^1 = R^4 = \text{OH}$  und  $R^2 = \text{C}_2\text{H}_5$ ,  $R^3 = \text{CH}_3$  bzw.  $R^2 = \text{CH}_3$ ,  $R^3 = \text{C}_2\text{H}_5$ .

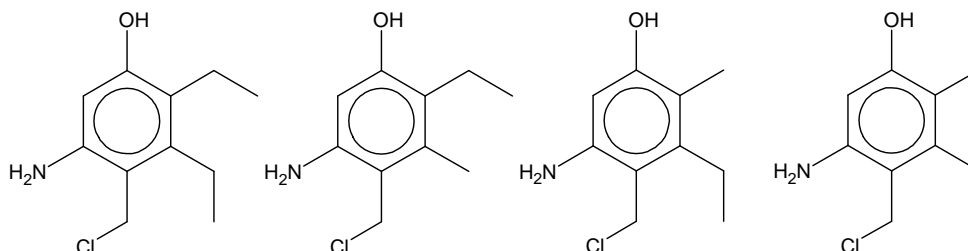
Für dieses einfache Beispiel konnte man die Suche nach Dubletten noch mit Papier und Bleistift erledigen. Es lässt sich aber bereits erahnen, dass bei komplexeren generischen Strukturformel zur redundanzfreien Generierung eine Rechnerunterstützung dringend erforderlich ist.

### Vergleich der Strukturräume

Zurück zu unserem eigentlichen Problem, der Suche nach Überschneidungen von  $\mathcal{L}_1$  und  $\mathcal{L}_2$ : Dazu identifizieren wir zunächst aromatische Bindungen und berechnen kanonische Nummerierungen der Strukturen (vgl. Abschnitt 4.5.2). Der nachfolgende Vergleich ergab

$$|\mathcal{L}_1 \cap \mathcal{L}_2| = 4$$

gemeinsame Strukturen. Die Patentverletzungen sind im Einzelnen:



Natürlich wird die vollständige Generierung aller Strukturen bei größeren Strukturräumen nicht das einzige Mittel zur Lösung des Problems bleiben. Einen Ausblick auf andere Verfahren zeigt beispielsweise [8]. Dort wird insbesondere die Berechnung von Deskriptoren für große, durch generische Strukturformeln definierte kombinatorische Bibliotheken beschrieben. In jedem Falle werden Kanonisierungsverfahren und Normalformen beim Vergleich generischer Strukturformeln und beim Aufbau von Datenbanken zu diesem Zweck eine wichtige Rolle spielen.

# Anhang





# Anhang A

## Datenstrukturen für molekulare Graphen

Je nachdem zu welchem Verwendungszweck molekulare Graphen im Speicher eines Rechners dargestellt werden, ist es sinnvoll, verschiedene Datenstrukturen bereit zu halten. In manchen Situationen, etwa bei bruttoformelbasierter Strukturgenerierung ist es wichtig, bestmöglichen *Zufallszugriff* auf die Kanten zu haben. Dann ist die Verwendung einer *Adjazenzmatrix* mit den Bindungen als Einträgen sinnvoll. Diese Art der Speicherung resultiert in einem großen Platzbedarf.

In anderen Situationen, beispielsweise für Substruktursuche, ist ein schneller *sequentieller Zugriff* vorteilhaft. Dann empfiehlt sich eine *vollständige Nachbarschaftsliste* als Datenstruktur. Vollständig heißt, dass zu jedem Knoten eine Liste aller benachbarten Atome geführt wird.

Oft ist man an einer möglichst *platzsparenden* Speicherung interessiert, wohingegen ein schneller Zugriff auf Atome und Bindungen völlig irrelevant ist. Dies ist immer dann der Fall, wenn Objekte auf Massenspeichermedien, d.h. außerhalb des Arbeitsspeichers abgelegt werden sollen, oder aber eine Aufbewahrung im Arbeitsspeicher nur dem Vergleich auf Identität dient. Dies trifft beispielsweise zu, wenn eine Strukturgenerierung mit Hilfe kanonischer Nummerierung durchgeführt wird. Eine solche kompakte Speicherung kann mit Hilfe einer *reduzierten Nachbarschaftsliste* realisiert werden. Dabei wird jede Kante nur *einmal* aufgeführt.

Im Folgenden wird das derzeit in *MOLGEN* verwendete Format zur kompakten Speicherung eines molekularen Graphen spezifiziert. Dabei wird auch berücksichtigt, dass H-Atome explizit oder implizit dargestellt werden können. Zudem ist optional die Speicherung von zwei- und dreidimensionalen Platzierungen sowie eines Molekülnamens möglich.

**1 Byte:** Angaben zu Umfang und Art der abgelegten Informationen

- Bit 0: *bName* (1, falls ein Name gespeichert wird)
- Bit 1: *bCoo2D* (1, falls 2D-Koordinaten gespeichert werden)
- Bit 2: *bCoo3D* (1, falls 3D-Koordinaten gespeichert werden)
- Bit 2: *bExplH* (1, falls H-Atome explizit gespeichert werden)
- Bit 4-7: derzeit ohne Verwendung (frei für Erweiterungen)

**if** *bName* = 1

- 4 Byte:** Anzahl *k* der Zeichen im Namen (**unsigned int**)
- k* **Byte:** Name des molekularen Graphen

**end**

**if** *bExplH* = 1

- 2 Byte:** Anzahl *n* der Atome (**unsigned short**)

**else**

- 2 Byte:** Anzahl *n* der Nicht-H-Atome (**unsigned short**)

**end**

**for each** (Nicht-H-)Atom *i*

- 1 Byte:** Ordnungszahl des Atoms -1 (**unsigned char**)
- 1 Byte:** Massendifferenz zum häufigsten Isotop (**signed char**)
- 1 Byte:** Ladung (**signed char**)
- 1 Byte:** Radikalstelle (**unsigned char**)
- if** *bCoo2D* = 1
  - 4 Byte:** 1. Koordinate der 2D-Platzierung (**float**)
  - 4 Byte:** 2. Koordinate der 2D-Platzierung (**float**)
- end**
- if** *bCoo3D* = 1
  - 4 Byte:** 1. Koordinate der 3D-Platzierung (**float**)
  - 4 Byte:** 2. Koordinate der 3D-Platzierung (**float**)
  - 4 Byte:** 3. Koordinate der 3D-Platzierung (**float**)
- end**
- if** *bExplH* = 0
  - 1 Byte:** Anzahl zu *i* benachbarter H-Atome (**unsigned char**)
- end**
- 1 Byte:** Anzahl zu *i* benachbarter (Nicht-H-)Atome *j* mit *j* > *i* (**unsigned char**)
- for each** zu *i* benachbartem (Nicht-H-)Atom *j* > *i*
  - s* **Byte:** *j*, wobei *s* = 1, falls *n* ≤ 256, *s* = 2 sonst (**unsigned char** bzw. **unsigned short**)
  - 1 Byte:** Bindungsvielfachheit zwischen *i* und *j*, 4 für aromatische Bindungen (**unsigned char**)

Für einen molekularen Graphen mit  $n$  Atomen und  $b$  Bindungen werden ohne die Speicherung von Koordinaten und Namen bei expliziten H-Atomen

$$1 + 2 + n \cdot (4 + 1) + b \cdot 2 = 3 + 5n + 2b$$

Byte benötigt. Sind unter den  $n$  Atomen  $h$  H-Atome, so beläuft sich bei impliziter Behandlung der H-Atome der Speicherplatzbedarf auf

$$1 + 2 + n \cdot (4 + 1 + 1) + (b - h) \cdot 2 = 3 + 6n - 8h + 2b$$

Byte. Bei chemischen Verbindungen in Datenbanken ist der durchschnittliche Anteil von H-Atomen etwa 0,5. So gesehen ist die implizite Speicherung von H-Atomen vorzuziehen. In einigen Spezialfällen ( $\text{H}_2$ ,  $\text{H}^+$ ,  $\text{H}^\bullet$ ) ist die explizite Speicherung unumgänglich. Ansonsten wird explizite Speicherung dann gewählt, wenn Koordinaten für alle Atome abgelegt werden sollen.

Diese Datenstruktur erhebt keinen Anspruch auf maximale Kompression. Insbesondere könnten bei der Kodierung der Atome weitere Einsparungen erzielt werden, etwa indem man zunächst eine Liste der auftretenden Atomzustände einführt, und bei den einzelnen Atomen dann nur noch mit einem Byte auf das Listenelement verweist.



# Anhang B

## Molekulare Deskriptoren

Die folgenden Seiten zeigen eine Aufstellung aller in *MOLGEN-QSPR* enthaltenen arithmetischen, topologischen und geometrischen Indizes. Eine genaue Spezifikation findet man in [119]. Für die Namen der Deskriptoren wurden die englischsprachigen Bezeichnungen beibehalten.

### B.1 Arithmetische Indizes

$A$	number of atoms
$A$ ( <i>incl. H</i> )	number of atoms (incl. H-atoms)
$N_H$	number of H-atoms
<i>rel.</i> $N_H$	relative number of H-atoms
$N_C$	number of C-atoms
<i>rel.</i> $N_C$	relative number of C-atoms
$N_O$	number of O-atoms
<i>rel.</i> $N_O$	relative number of O-atoms
$N_N$	number of N-atoms
<i>rel.</i> $N_N$	relative number of N-atoms
$N_S$	number of S-atoms
<i>rel.</i> $N_S$	relative number of S-atoms
$N_F$	number of F-atoms
<i>rel.</i> $N_F$	relative number of F-atoms
$N_{Cl}$	number of Cl-atoms
<i>rel.</i> $N_{Cl}$	relative number of Cl-atoms
$N_{Br}$	number of Br-atoms
<i>rel.</i> $N_{Br}$	relative number of Br-atoms
$N_I$	number of I-atoms
<i>rel.</i> $N_I$	relative number of I-atoms

$N_P$	number of P-atoms
<i>rel. N<sub>P</sub></i>	relative number of P-atoms
$B$	number of bonds
$B$ ( <i>incl. H</i> )	number of bonds (incl. H-atoms)
<i>loc. B</i>	number of localized bonding electron pairs
<i>loc. B</i> ( <i>incl. H</i> )	number of localized bonding electron pairs (incl. H)
$n-$	number of single bonds
<i>rel. n-</i>	relative number of single bonds
$n-$ ( <i>incl. H</i> )	number of single bonds (incl. H-atoms)
<i>rel. n-</i> ( <i>incl. H</i> )	relative number of single bonds (incl. H-atoms)
$n=$	number of double bonds
<i>rel. n=</i>	relative number of double bonds
$n=$ ( <i>incl. H</i> )	number of double bonds (incl. H-atoms)
$n\#$	number of triple bonds
<i>rel. n\#</i>	relative number of triple bonds
<i>rel. n\#</i> ( <i>incl. H</i> )	relative number of triple bonds (incl. H-atoms)
$n_{aroma}$	number of aromatic bonds
<i>rel. n<sub>aroma</sub></i>	relative number of aromatic bonds
<i>rel. n<sub>aroma</sub></i> ( <i>incl. H</i> )	relative number of aromatic bonds (incl. H-atoms)
$C$	cyclomatic number
$MW$	molecular weight
<i>mean AW</i>	mean atomic weight
$MW$ ( <i>incl. H</i> )	molecular weight (incl. H-atoms)
<i>mean AW</i> ( <i>incl. H</i> )	mean atomic weight (incl. H-atoms)
$cha$	total charge
$rad$	number of radicals
$HBD$	number of hydrogen bond donors
$HBA$	number of hydrogen bond acceptors

## B.2 Topologische Indizes

$W$	Wiener index
$M_1$	1st Zagreb index
$M_2$	2nd Zagreb index
${}^m M_1$	1st modified Zagreb index
${}^m M_2$	2nd modified Zagreb index
${}^0 \chi$	Randic index of order 0
${}^1 \chi$	Randic index of order 1
${}^2 \chi$	Randic index of order 2
${}^0 \chi^s$	solvation connectivity index of order 0

${}^1\chi^s$	solvation connectivity index of order 1
${}^2\chi^s$	solvation connectivity index of order 2
${}^3\chi^s$	solvation connectivity index of order 3
${}^3\chi_C^s$	solvation connectivity index for clusters
${}^0\chi^v$	Kier and Hall index of order 0
${}^1\chi^v$	Kier and Hall index of order 1
${}^2\chi^v$	Kier and Hall index of order 2
${}^3\chi^v$	Kier and Hall index of order 3
${}^1\kappa$	Kier shape index 1
${}^2\kappa$	Kier shape index 2
${}^3\kappa$	Kier shape index 3
$\Phi_{\bar{\alpha}}$	Kier molecular flexibility index non-alpha-modified
${}^1\kappa_{\alpha}$	Kier alpha-modified shape index 1
${}^2\kappa_{\alpha}$	Kier alpha-modified shape index 2
${}^3\kappa_{\alpha}$	Kier alpha-modified shape index 3
$\Phi$	Kier molecular flexibility index
$F$	Platt number
$N_{GS}$	Gordon-Scantlebury index
$J$	Balaban index
$J_{unsat}$	unsaturated Balaban index
$MTI$	Schultz molecular topological index
$MTI'$	MTI'-index
$H$	Harary number
$twc$	total walk count
$mwC^{(2)}$	molecular walk count of length 2
$mwC^{(3)}$	molecular walk count of length 3
$mwC^{(4)}$	molecular walk count of length 4
$mwC^{(5)}$	molecular walk count of length 5
$mwC^{(6)}$	molecular walk count of length 6
$mwC^{(7)}$	molecular walk count of length 7
$mwC^{(8)}$	molecular walk count of length 8
$twC_{unsat}$	unsaturated total walk count
$mwC_{unsat}^{(2)}$	unsaturated molecular walk count of length 2
$mwC_{unsat}^{(3)}$	unsaturated molecular walk count of length 3
$mwC_{unsat}^{(4)}$	unsaturated molecular walk count of length 4
$mwC_{unsat}^{(5)}$	unsaturated molecular walk count of length 5
$mwC_{unsat}^{(6)}$	unsaturated molecular walk count of length 6
$mwC_{unsat}^{(7)}$	unsaturated molecular walk count of length 7
$mwC_{unsat}^{(8)}$	unsaturated molecular walk count of length 8

$G_1$ ( <i>topol.</i> )	gravitational index (pairs, topol. dist.)
$G_1$ ( <i>topol., incl. H</i> )	gravitational index (pairs, topol. dist., incl. H-atoms)
$G_2$ ( <i>topol.</i> )	gravitational index (bonds, topol. dist.)
$G_2$ ( <i>topol., incl. H</i> )	gravitational index (bonds, topol. dist., incl. H-atoms)
$Z$	Hosoya $Z$ -index
$IC_0$	Basak information content of order 0
$TIC_0$	Basak total information content of order 0
$CIC_0$	Basak complementary information content of order 0
$N * CIC_0$	total complementary information content of order 0
$SIC_0$	Basak structural information content of order 0
$N * SIC_0$	total structural information content of order 0
$BIC_0$	bonding information content of order 0
$N * BIC_0$	total bonding information content of order 0
$IC_1$	Basak information content of order 1
$TIC_1$	Basak total information content of order 1
$CIC_1$	Basak complementary information content of order 1
$N * CIC_1$	total complementary information content of order 1
$SIC_1$	Basak structural information content of order 1
$N * SIC_1$	total structural information content of order 1
$BIC_1$	bonding information content of order 1
$N * BIC_1$	total bonding information content of order 1
$IC_2$	Basak information content of order 2
$TIC_2$	Basak total information content of order 2
$CIC_2$	Basak complementary information content of order 2
$N * CIC_2$	total complementary information content of order 2
$SIC_2$	Basak structural information content of order 2
$N * SIC_2$	total structural information content of order 2
$BIC_2$	bonding information content of order 2
$N * BIC_2$	total bonding information content of order 2
$MSD$	mean square distance index
$w$	detour index
$w_{diag}$	detour index (incl. half main diagonal)
$P_{acyc}$	total acyclic path count
${}^2P_{acyc}$	molecular acyclic path count of length 2
${}^3P_{acyc}$	molecular acyclic path count of length 3
${}^4P_{acyc}$	molecular acyclic path count of length 4
${}^5P_{acyc}$	molecular acyclic path count of length 5
${}^6P_{acyc}$	molecular acyclic path count of length 6
${}^7P_{acyc}$	molecular acyclic path count of length 7
${}^8P_{acyc}$	molecular acyclic path count of length 8
$\geq^9P_{acyc}$	molecular acyclic path count of length 9 and higher



$P$	total path count
${}^2P$	molecular path count of length 2
${}^3P$	molecular path count of length 3
${}^4P$	molecular path count of length 4
${}^5P$	molecular path count of length 5
${}^6P$	molecular path count of length 6
${}^7P$	molecular path count of length 7
${}^8P$	molecular path count of length 8
$\geq^9P$	molecular path count of length 9 and higher
<i>rings</i>	total ring count
${}^3rings$	molecular ring count of length 3
${}^4rings$	molecular ring count of length 4
${}^5rings$	molecular ring count of length 5
${}^6rings$	molecular ring count of length 6
${}^7rings$	molecular ring count of length 7
${}^8rings$	molecular ring count of length 8
$\geq^9rings$	molecular ring count of length 9 and higher
<i>ch. <math>G_1</math></i>	topological charge index of order 1
<i>ch. <math>G_2</math></i>	topological charge index of order 2
<i>ch. <math>G_3</math></i>	topological charge index of order 3
<i>ch. <math>G_4</math></i>	topological charge index of order 4
<i>ch. <math>G_5</math></i>	topological charge index of order 5
<i>ch. <math>G_6</math></i>	topological charge index of order 6
<i>ch. <math>G_7</math></i>	topological charge index of order 7
<i>ch. <math>G_8</math></i>	topological charge index of order 8
<i>ch. <math>J_1</math></i>	mean topological charge index of order 1
<i>ch. <math>J_2</math></i>	mean topological charge index of order 2
<i>ch. <math>J_3</math></i>	mean topological charge index of order 3
<i>ch. <math>J_4</math></i>	mean topological charge index of order 4
<i>ch. <math>J_5</math></i>	mean topological charge index of order 5
<i>ch. <math>J_6</math></i>	mean topological charge index of order 6
<i>ch. <math>J_7</math></i>	mean topological charge index of order 7
<i>ch. <math>J_8</math></i>	mean topological charge index of order 8
<i>ch. <math>J</math></i>	global topological charge index
$D$	topological diameter
$\xi^c$	eccentric connectivity index
$\lambda_1^A$	principal eigenvalue of $A$
$SCA1$	sum of coefficients of princ. eigenvec. of $A$
$SCA2$	mean coefficient of princ. eigenvec. of $A$
$SCA3$	log of sum of coeff. of princ. eigenvec. of $A$
$\lambda_1^D$	principal eigenvalue of $D$

$\chi_T$	total $\chi$ index
$T_m$	number of methyl groups
$T_3$	number of pairs of methyl groups at distance 3
$FRB$	freely rotatable bonds
$SZD$	Szeged index
$SZDp$	hyper-Szeged index
${}^3\chi_p$	connectivity index ${}^3\chi$ path
${}^4\chi_p$	connectivity index ${}^4\chi$ path
${}^5\chi_p$	connectivity index ${}^5\chi$ path
${}^6\chi_p$	connectivity index ${}^6\chi$ path
${}^3\chi_c$	connectivity index ${}^3\chi$ cluster
${}^4\chi_c$	connectivity index ${}^4\chi$ cluster
${}^5\chi_c$	connectivity index ${}^5\chi$ cluster
${}^6\chi_c$	connectivity index ${}^6\chi$ cluster
${}^4\chi_{pc}$	connectivity index ${}^4\chi$ path-cluster
${}^5\chi_{pc}$	connectivity index ${}^5\chi$ path-cluster
${}^6\chi_{pc}$	connectivity index ${}^6\chi$ path-cluster
${}^3\chi_{ch}$	connectivity index ${}^3\chi$ chain
${}^4\chi_{ch}$	connectivity index ${}^4\chi$ chain
${}^5\chi_{ch}$	connectivity index ${}^5\chi$ chain
${}^6\chi_{ch}$	connectivity index ${}^6\chi$ chain
${}^3\chi_p^v$	connectivity index ${}^3\chi^v$ path
${}^4\chi_p^v$	connectivity index ${}^4\chi^v$ path
${}^5\chi_p^v$	connectivity index ${}^5\chi^v$ path
${}^6\chi_p^v$	connectivity index ${}^6\chi^v$ path
${}^3\chi_c^v$	connectivity index ${}^3\chi^v$ cluster
${}^4\chi_c^v$	connectivity index ${}^4\chi^v$ cluster
${}^5\chi_c^v$	connectivity index ${}^5\chi^v$ cluster
${}^6\chi_c^v$	connectivity index ${}^6\chi^v$ cluster
${}^4\chi_{pc}^v$	connectivity index ${}^4\chi^v$ path-cluster
${}^5\chi_{pc}^v$	connectivity index ${}^5\chi^v$ path-cluster
${}^6\chi_{pc}^v$	connectivity index ${}^6\chi^v$ path-cluster
${}^3\chi_{ch}^v$	connectivity index ${}^3\chi^v$ chain
${}^4\chi_{ch}^v$	connectivity index ${}^4\chi^v$ chain
${}^5\chi_{ch}^v$	connectivity index ${}^5\chi^v$ chain
${}^6\chi_{ch}^v$	connectivity index ${}^6\chi^v$ chain
$sym$	size of topological symmetry group
$R$	topological radius
$con. comp.$	number of connectivity components
$gt planar$	graph-theoretical planarity

## B.3 Geometrische Indizes

$G_1$	gravitational index (pairs, 3D-dist.)
$G_1$ ( <i>incl. H</i> )	gravitational index (pairs, 3D-dist., incl. H-atoms)
$G_2$	gravitational index (bonds, 3D-dist.)
$G_2$ ( <i>incl. H</i> )	gravitational index (bonds, 3D-dist., incl. H-atoms)
$I_A$	moment of inertia A
$I_B$	moment of inertia B
$I_C$	moment of inertia C
<i>st. energy</i>	steric energy
<i>SHDW1</i>	XY shadow
<i>SHDW2</i>	XZ shadow
<i>SHDW3</i>	YZ shadow
<i>SHDW4</i>	standardized XY shadow
<i>SHDW5</i>	standardized XZ shadow
<i>SHDW6</i>	standardized YZ shadow
<i>SHDW1/SHDW2</i>	XY/XZ shadow
<i>SHDW1/SHDW3</i>	XY/YZ shadow
<i>SHDW2/SHDW3</i>	XZ/YZ shadow
<i>ssSHDW1</i>	size sorted shadow 1
<i>ssSHDW2</i>	size sorted shadow 2
<i>ssSHDW3</i>	size sorted shadow 3
<i>ssSHDW4</i>	size sorted standardized shadow 1
<i>ssSHDW5</i>	size sorted standardized shadow 2
<i>ssSHDW6</i>	size sorted standardized shadow 3
<i>ssSHDW1/SHDW2</i>	size sorted shadow 1/2
<i>ssSHDW1/SHDW3</i>	size sorted shadow 1/3
<i>ssSHDW2/SHDW3</i>	size sorted shadow 2/3
$V_{vdw}$	Van der Waals volume
$\rho_{vdw}$	density by Van der Waals volume
$V_{vdw}^s$	standardized Van der Waals volume
$V_{cub}$	enclosing cuboid
$S_{vdw}$	Van der Waals surface
$SAS_{H_2O}$	solvent-accessible surface (H <sub>2</sub> O)
$SAS_H$	solvent-accessible surface (H)
$D_{3D}$	geometrical diameter
$V_{sphere}$	enclosing sphere



# Anhang C

## Substrukturen für MS-Klassifikatoren

Im Folgenden werden die strukturellen Eigenschaften beschrieben, nach denen die Software *MSclass* [156] Massenspektren klassifizieren kann. *MSclass* umfasst Klassifikatoren zu insgesamt 85 verschiedenen SP, für jede dieser Eigenschaften werden bis zu 4 Klassifikatoren angeboten.

Die SP werden identifiziert durch einen bis zu 10 Zeichen langen Namen. Zudem ist für jede SP eine englischsprachige Beschreibung und eine Graphik angegeben. Die 85 SP sind in 5 Kategorien aufgeteilt: *Alkyle* (13 SP), *Aromaten* (40 SP), *Bindungen* (2 SP), *Elemente* (10 SP), *funktionelle Gruppen* (19 SP) und *Ringe* (1 SP). Die SP sind innerhalb dieser Kategorien gemäß ihrer Beschreibung alphabetisch aufsteigend angeordnet. Diese Informationen stimmen überein mit den Angaben im Klassifikator Handbuch [157] zu *MSclass* und wurden mit Genehmigung der Autoren zur Verfügung gestellt. Um die Klassifikationsergebnisse für einen Strukturgenerator wie *MOLGEN* nutzen zu können, müssen die SP durch Restriktionen beschrieben werden, die der Generator verarbeiten kann. Zu diesem Zweck wurden die strukturellen Informationen auf arithmetischer und topologischer Ebene in einem für *MOLGEN-MS* lesbaren Format kodiert.

Auf arithmetischer Ebene sind dies in erster Linie Intervalle, die die Anzahlen von Atomen für die chemischen Elemente einschränken, auf topologischer Ebene ein oder mehrere logische verknüpfte Substrukturen, die *MOLGEN* als strukturelle Restriktion verarbeitet. Im Folgenden sind für jede strukturelle Eigenschaft *S* die arithmetischen Restriktionen (AR) und strukturellen Restriktionen (SR) für das Vorhandensein (Klasse 1) und die Abwesenheit (Klasse 0) von *S* entsprechend dem derzeitigen Entwicklungsstand von *MOLGEN-MS* zusammengestellt.

Auf eine explizite graphische Darstellung der einzelnen Substrukturen wurde

aus Platzgründen verzichtet. In den meisten Fällen geht diese ohnehin aus der Beschreibung oder der Graphik der SP hervor. Stattdessen wurden in Fällen, wo die Kodierung durch strukturelle Restriktionen nicht offensichtlich ist, Bemerkungen über die verwendeten Methoden beigelegt.

Nicht immer können die in einer SP enthaltenen Informationen vollständig für die Strukturgenerierung durch *MOLGEN* genutzt werden. Manche SP sind zudem nicht präzise genug für eine automatische Verarbeitung formuliert. Betroffene SP sind mit einem Kreuz (†) gekennzeichnet.

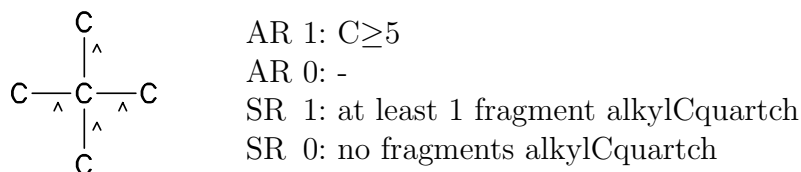
Schließlich wurden die hier aufgeführten SP teilweise zur Entwicklung neuer Klassifikatoren herangezogen (Abschnitt 5.5.2). SP, die dabei nicht berücksichtigt wurden, sind mit einem Doppelkreuz (‡) markiert. Gründe dafür waren entweder, dass diese SP für die Strukturgenerierung durch *MOLGEN* nicht nutzbar sind, oder dass keine entsprechend großen Datenbestände für Lern- und Testsatz in der Datenbasis (Abschnitt 5.3.5) vorlagen.

Dabei verteilen sich die SP, deren Informationen nicht vollständig für *MOLGEN* genutzt werden können und solche, die nicht für die Klassifikation in Abschnitt 5.5.2 herangezogen wurden, wie folgt auf die einzelnen Kategorien:

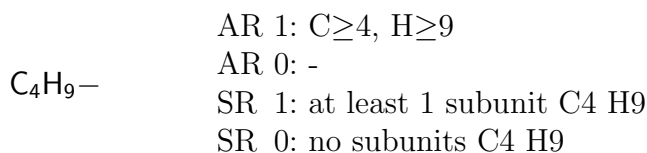
	SP	†	‡
Alkyle	13	2	0
Aromaten	40	9	7
Bindungen	2	0	0
Elemente	10	0	1
fkt. Gruppen	19	0	0
Ringe	1	0	0
$\Sigma$	85	11	8

## C.1 Alkyle

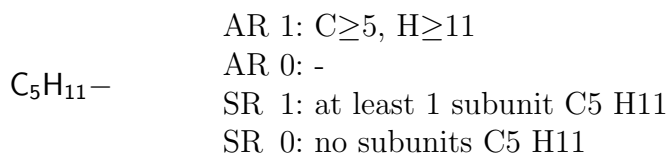
**C quart ch:** *C quarternary (4 chain-bonds to carbon atoms)*<sup>1</sup>



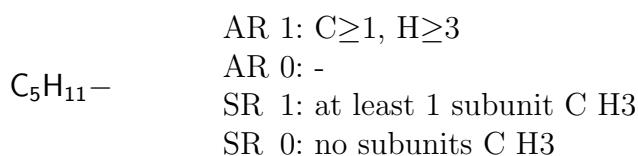
**C4 H9:** *C04 H09*<sup>2</sup>



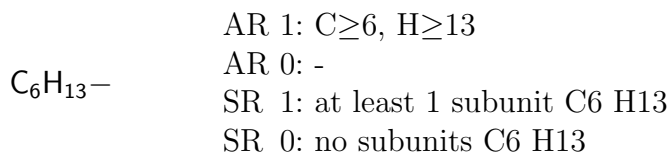
**C5 H11:** *C05 H11*



**C5 H11\*:** *C05 H11 or other alkyl*<sup>†,3</sup>



**C6 H13:** *C06 H13*



<sup>1</sup>Azyklische Bindungen ( $\wedge$ ) werden unter Verwendung von Substruktur-Restriktionen *Ring* beschrieben.

<sup>2</sup>Alkylreste mit gegebener Summenformel werden durch Summenformel-Substrukturen beschrieben.

<sup>3</sup>Hier ist die Definition der SP nicht präzise. Für einen beliebigen Alkylrest kann man lediglich eine CH<sub>3</sub>-Gruppe vorschreiben.

**C6 H13 n:** *C06 H13 (n-)*

$\text{CH}_3(\text{CH}_2)_5-$  AR 1:  $C \geq 6, H \geq 13$   
 AR 0: -  
 SR 1: at least 1 fragment alkylC6H13(n-)  
 SR 0: no fragments alkylC6H13(n-)

**C7 H15:** *C07 H15*

$\text{C}_7\text{H}_{15}-$  AR 1:  $C \geq 7, H \geq 15$   
 AR 0: -  
 SR 1: at least 1 subunit C7 H15  
 SR 0: no subunits C7 H15

**C8 H17:** *C08 H17*

$\text{C}_8\text{H}_{17}-$  AR 1:  $C \geq 8, H \geq 17$   
 AR 0: -  
 SR 1: at least 1 subunit C8 H17  
 SR 0: no subunits C8 H17

**C9 H19:** *C09 H19*

$\text{C}_9\text{H}_{19}-$  AR 1:  $C \geq 9, H \geq 19$   
 AR 0: -  
 SR 1: at least 1 subunit C9 H19  
 SR 0: no subunits C9 H19

**C10 H21:** *C10 H21*

$\text{C}_{10}\text{H}_{21}-$  AR 1:  $C \geq 10, H \geq 21$   
 AR 0: -  
 SR 1: at least 1 subunit C10 H21  
 SR 0: no subunits C10 H21

**C11 H23:** *C11 H23*<sup>†,4</sup>

$\text{C}_{\geq 11}\text{H}_{23}-$  AR 1:  $C \geq 11, H \geq 23$   
 AR 0: -  
 SR 1: -  
 SR 0: -

---

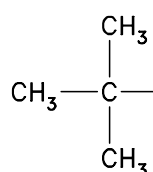
<sup>4</sup>Die Definition der SP besagt hier, dass  $\text{C}_{11}\text{H}_{23}$  oder ein größerer Alkylrest vorliegt. Diese Bedingung kann derzeit nicht in *MOLGEN*-Syntax beschrieben werden. Für die Bestimmung von Klassifikatoren in Abschnitt 5.5.2 wurde die Bedingung durch  $\text{C}_{11}\text{H}_{23}$  ersetzt.



**hydr carb:** *hydrocarbon*

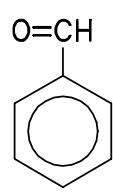
$C_xH_y$  AR 1:  $C \geq 1$   $H \geq 1$   
 no hetero atoms  
 AR 0: at least 1 hetero atom  
 SR 1: -  
 SR 0: -

**(CH<sub>3</sub>)<sub>3</sub>-C:** *tertiary butyl*

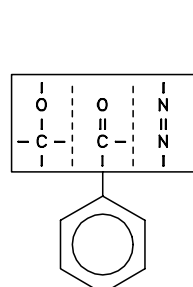
 AR 1:  $C \geq 4$ ,  $H \geq 9$   
 AR 0: -  
 SR 1: at least 1 fragment alkyl(CH<sub>3</sub>)<sub>3</sub>-C  
 SR 0: no fragments alkyl(CH<sub>3</sub>)<sub>3</sub>-C

## C.2 Aromaten

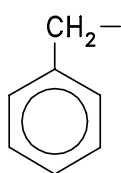
**ar-CHO:** *aldehyde aryl-CH=O*

 AR 1:  $C \geq 7$   $H \geq 1$   $O \geq 1$   
 $DBE \geq 5$   
 AR 0: -  
 SR 1: at least 1 fragment aromaar-CHO  
 SR 0: no fragments aromaar-CHO

**ar-CO,N2:** *aryl - -C-O or -C=O or -N=N*<sup>5</sup>

 AR 1:  $C \geq 6$   
 $DBE \geq 4$   
 AR 0: -  
 SR 1: 1-2 terms of  
 at least 1 fragment aromaar-COsd,  
 at least 1 fragment aromaar-N2  
 SR 0: exactly 2 terms of  
 no fragments aromaar-COsd,  
 no fragments aromaar-N2

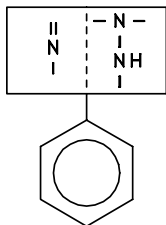
<sup>5</sup>Die Alternativen  $-C-O$  und  $-C=O$  werden als *ein* MMG durch Verwendung mehrerer möglicher Bindungen dargestellt. Die dritte Alternative wird durch Verwendung eines weiteren Substruktur-Eintrags in einer strukturellen Restriktion *Substruktur* realisiert.

**ar-CH2:** *aryl - -CH2 or -CH3*AR 1:  $C \geq 7$   $H \geq 2$   
DBE  $\geq 4$ 

AR 0: -

SR 1: at least 1 fragment aromaar-CH2

SR 0: no fragments aromaar-CH2

**ar-N,NHN:** *aryl - -N= or -NH-N* <sup>6</sup>AR 1:  $C \geq 6$   $N \geq 1$   
DBE  $\geq 4$ 

AR 0: -

SR 1: 1-2 terms of

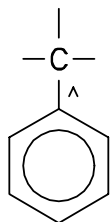
at least 1 fragment aromaar-Nsp2,

at least 1 fragment aromaar-NHN

SR 0: exactly 2 terms of

no fragments aromaar-Nsp2,

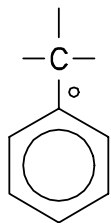
no fragments aromaar-NHN

**ar-C ch:** *aryl - C (chain-bond)*AR 1:  $C \geq 7$   
DBE  $\geq 4$ 

AR 0: -

SR 1: at least 1 fragment aromaar-Cch

SR 0: no fragments aromaar-Cch

**ar-C r:** *aryl - C (ring bond)* <sup>7</sup>AR 1:  $C \geq 7$   
DBE  $\geq 5$ 

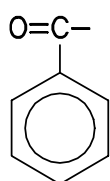
AR 0: -

SR 1: at least 1 fragment aromaar-Cr

SR 0: no fragments aromaar-Cr

<sup>6</sup>Die Substruktur  $-N=$  wird unter Verwendung einer Substruktur-Restriktion *Hybridisierung* beschrieben.

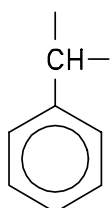
<sup>7</sup>Zyklische Bindungen (o) werden ebenso wie azyklische Bindungen durch Substruktur-Restriktionen *Ring* beschrieben.

**ar-CO:** *aryl - C=O*AR 1:  $C \geq 7$   $O \geq 1$   
DBE  $\geq 5$ 

AR 0: -

SR 1: at least 1 fragment aromaar-CO

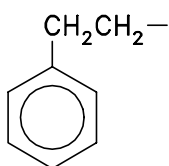
SR 0: no fragments aromaar-CO

**ar-CH:** *aryl - CH*AR 1:  $C \geq 7$   $H \geq 1$   
DBE  $\geq 4$ 

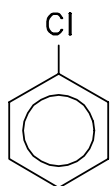
AR 0: -

SR 1: at least 1 fragment aromaar-CH

SR 0: no fragments aromaar-CH

**ar-CH<sub>2</sub>CH<sub>2</sub>:** *aryl - CH<sub>2</sub>-CH<sub>2</sub>*AR 1:  $C \geq 8$   $H \geq 4$   
DBE  $\geq 4$ 

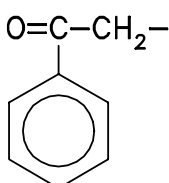
AR 0: -

SR 1: at least 1 fragment aromaar-CH<sub>2</sub>CH<sub>2</sub>SR 0: no fragments aromaar-CH<sub>2</sub>CH<sub>2</sub>**ar-Cl:** *aryl - Cl*AR 1:  $C \geq 6$   $Cl \geq 1$   
DBE  $\geq 4$ 

AR 0: -

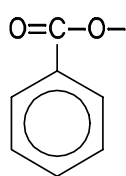
SR 1: at least 1 fragment aromaar-Cl

SR 0: no fragments aromaar-Cl

**ar-CO-CH<sub>2</sub>:** *aryl - CO-CH<sub>2</sub>*AR 1:  $C \geq 8$   $H \geq 2$   $O \geq 1$   
DBE  $\geq 5$ 

AR 0: -

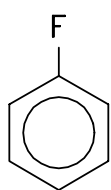
SR 1: at least 1 fragment aromaar-CO-CH<sub>2</sub>SR 0: no fragments aromaar-CO-CH<sub>2</sub>

**ar-COO:** *aryl - COO (benzoic acid/ester)*AR 1:  $C \geq 7$   $O \geq 2$   
DBE  $\geq 5$ 

AR 0: -

SR 1: at least 1 fragment aromaar-COO

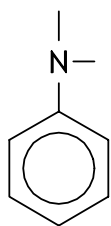
SR 0: no fragments aromaar-COO

**ar-F:** *aryl - F*AR 1:  $C \geq 6$   $F \geq 1$   
DBE  $\geq 4$ 

AR 0: -

SR 1: at least 1 fragment aromaar-F

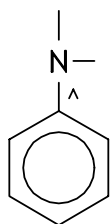
SR 0: no fragments aromaar-F

**ar-N:** *aryl - N*AR 1:  $C \geq 6$   $N \geq 1$   
DBE  $\geq 4$ 

AR 0: -

SR 1: at least 1 fragment aromaar-N

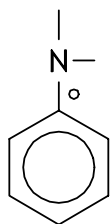
SR 0: no fragments aromaar-N

**ar-N ch:** *aryl - N (chain-bond)*AR 1:  $C \geq 6$   $N \geq 1$   
DBE  $\geq 4$ 

AR 0: -

SR 1: at least 1 fragment aromaar-Nch

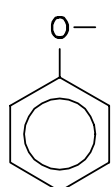
SR 0: no fragments aromaar-Nch

**ar-N r:** *aryl - N (ring-bond)*AR 1:  $C \geq 6$   $N \geq 1$   
DBE  $\geq 5$ 

AR 0: -

SR 1: at least 1 fragment aromaar-Nr

SR 0: no fragments aromaar-Nr

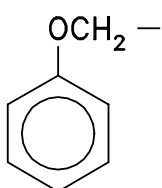
**ar-O:** *aryl - O*

AR 1:  $C \geq 6$   $O \geq 1$   
 DBE  $\geq 4$

AR 0: -

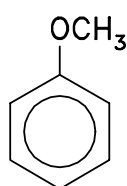
SR 1: at least 1 fragment aromaar-O

SR 0: no fragments aromaar-O

**ar-O-CH<sub>2</sub>:** *aryl - O-CH<sub>2</sub>*

AR 1:  $C \geq 7$   $H \geq 2$   $O \geq 1$   
 DBE  $\geq 4$

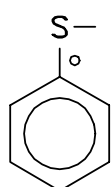
AR 0: -

SR 1: at least 1 fragment aromaar-O-CH<sub>2</sub>SR 0: no fragments aromaar-O-CH<sub>2</sub>**ar-O-CH<sub>3</sub>:** *aryl - O-CH<sub>3</sub> (methoxy)*

AR 1:  $C \geq 7$   $H \geq 3$   $O \geq 1$   
 DBE  $\geq 4$

AR 0: -

SR 1: at least 1 fragment aromaar-O-CH

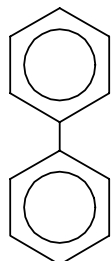
SR 0: no fragments aromaar-O-CH<sub>3</sub>**ar-S r:** *aryl - S (S in ring)*

AR 1:  $C \geq 6$   $S \geq 1$   
 DBE  $\geq 5$

AR 0: -

SR 1: at least 1 fragment aromaar-Sr

SR 0: no fragments aromaar-Sr

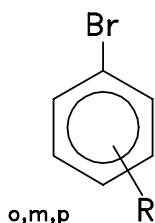
**biphenyl:** *biphenyl*

AR 1:  $C \geq 12$   
 DBE  $\geq 8$

AR 0: -

SR 1: at least 1 fragment aromabiphenyl

SR 0: no fragments aromabiphenyl

**C6H4-Br:**  $C_6H_4 - Br$  (*o,m,p substituted*)<sup>8</sup>

AR 1:  $C \geq 6$   $H \geq 4$   $Br \geq 1$   
 $DBE \geq 4$

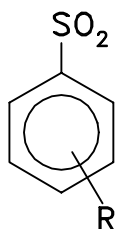
AR 0: -

SR 1: 1-5 terms of

at least 1 fragment aromaC6H4-Bra,  
 at least 1 fragment aromaC6H4-Brb,  
 at least 1 fragment aromaC6H4-Brc,  
 at least 1 fragment aromaC6H4-Brd,  
 at least 1 fragment aromaC6H4-Bre

SR 0: exactly 5 terms of

no fragments aromaC6H4-Bra,  
 no fragments aromaC6H4-Brb,  
 no fragments aromaC6H4-Brc,  
 no fragments aromaC6H4-Brd,  
 no fragments aromaC6H4-Bre

**C6H4-SO2:**  $C_6H_4 - SO_2$  (*o,m,p substituted*)<sup>‡,9</sup>

AR 1:  $C \geq 6$ ,  $H \geq 4$ ,  $O \geq 2$ ,  $S \geq 1$   
 $DBE \geq 4$

AR 0: -

SR 1: 1-5 terms of

at least 1 fragment aromaC6H4-SO2a,  
 at least 1 fragment aromaC6H4-SO2b,  
 at least 1 fragment aromaC6H4-SO2c,  
 at least 1 fragment aromaC6H4-SO2d,  
 at least 1 fragment aromaC6H4-SO2e

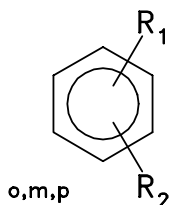
SR 0: exactly 5 terms of

no fragments aromaC6H4-SO2a,  
 no fragments aromaC6H4-SO2b,  
 no fragments aromaC6H4-SO2c,  
 no fragments aromaC6H4-SO2d,  
 no fragments aromaC6H4-SO2e

<sup>8</sup>Alternativen für die Position eines Substituenten werden durch Verwendung mehrerer Substruktur-Einträge in einer strukturellen Restriktion *Substruktur* realisiert.

<sup>9</sup>Für diese SP wurden zur Bestimmung von Klassifikatoren in Abschnitt 5.5.2 nicht genügend Spektren in der Datenbasis gefunden.

**C6H4 omp:**  $C_6H_4$  di-substituted (*o,m,p*) benzene ring



AR 1:  $C \geq 6$   $H \geq 4$

DBE  $\geq 4$

AR 0: -

SR 1: 1-4 terms of

at least 1 fragment `aromaC6H4ompa`,

at least 1 fragment `aromaC6H4ompb`,

at least 1 fragment `aromaC6H4ompc`,

at least 1 fragment `aromaC6H4ompd`,

SR 0: exactly 4 terms of

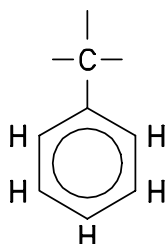
no fragments `aromaC6H4ompa`,

no fragments `aromaC6H4ompb`,

no fragments `aromaC6H4ompc`,

no fragments `aromaC6H4ompd`

**ph-C:**  $C_6H_5 - C$



AR 1:  $C \geq 7$   $H \geq 5$

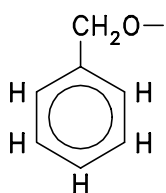
DBE  $\geq 4$

AR 0: -

SR 1: at least 1 fragment `aromaph-C`

SR 0: no fragments `aromaph-C`

**ph-CH<sub>2</sub>-O:**  $C_6H_5 - CH_2 - O$



AR 1:  $C \geq 7$   $H \geq 7$   $O \geq 1$

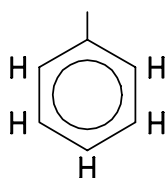
DBE  $\geq 4$

AR 0: -

SR 1: at least 1 fragment `aromaph-CH2-O`

SR 0: no fragments `aromaph-CH2-O`

**ph:**  $C_6H_5$ - (*phenyl*)



AR 1:  $C \geq 6$   $H \geq 5$

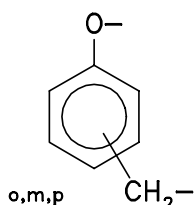
DBE  $\geq 4$

AR 0: -

SR 1: at least 1 fragment `aromaph`

SR 0: no fragments `aromaph`

**benz-O:**  $CH_2 - C_6H_4 - O - (o,m,p)$



AR 1:  $C \geq 7$   $H \geq 6$   $O \geq 1$   
DBE  $\geq 4$

AR 0: -

SR 1: 1-5 terms of

at least 1 fragment aromabenz-Oa,  
at least 1 fragment aromabenz-Ob,  
at least 1 fragment aromabenz-Oc,  
at least 1 fragment aromabenz-Od,  
at least 1 fragment aromabenz-Oe

SR 0: exactly 5 terms of

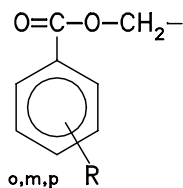
no fragments aromabenz-Oa,  
no fragments aromabenz-Ob,  
no fragments aromabenz-Oc,  
no fragments aromabenz-Od,  
no fragments aromabenz-Oe

**ar cond:** *condensed rings* †,‡

AR 1:  $C \geq 7$   
DBE  $\geq 6$

condensed aromatic rings  
AR 0: -  
SR 1: -  
SR 0: -

**ar-COOCH<sub>2</sub>\*:** *ester*  $C_6H_4-COO-CH_2-$  (and subst. at o,m, or p)



AR 1:  $C \geq 8$   $H \geq 6$   $O \geq 2$   
DBE  $\geq 5$

AR 0: -

SR 1: 1-5 terms of

at least 1 fragment aromaar-COOCH<sub>2</sub>a,  
at least 1 fragment aromaar-COOCH<sub>2</sub>b,  
at least 1 fragment aromaar-COOCH<sub>2</sub>c,  
at least 1 fragment aromaar-COOCH<sub>2</sub>d,  
at least 1 fragment aromaar-COOCH<sub>2</sub>e

SR 0: exactly 5 terms of

no fragments aromaar-COOCH<sub>2</sub>a,  
no fragments aromaar-COOCH<sub>2</sub>b,  
no fragments aromaar-COOCH<sub>2</sub>c,  
no fragments aromaar-COOCH<sub>2</sub>d,  
no fragments aromaar-COOCH<sub>2</sub>e



**ar het:** *hetero-aromatic* †,‡



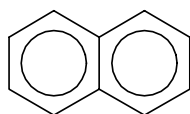
AR 1:  $C \geq 3$   
 DBE  $\geq 3$   
 AR 0: -  
 SR 1: -  
 SR 0: -

**ar poly:** *more than 1 aromatic ring (any type)* †,‡

more than one aromatic ring

AR 1:  $C \geq 7$   
 DBE  $\geq 6$   
 AR 0: -  
 SR 1: -  
 SR 0: -

**naph:** *naphthaline ring system*



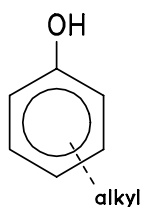
AR 1:  $C \geq 10$   
 DBE  $\geq 7$   
 AR 0: -  
 SR 1: 1-2 terms of  
 at least 1 fragment aromanapha,  
 at least 1 fragment aromanaphb,  
 SR 0: exactly 2 terms of  
 no fragments aromanapha,  
 no fragments aromanaphb

**non ar:** *non aromatic* †,‡

no aromatic ring

AR 1: -  
 AR 0:  $C \geq 3$   
 DBE  $\geq 3$   
 SR 1: -  
 SR 0: -

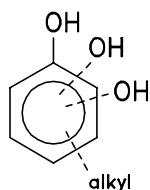
**phen-1-OH:** *phenol (1 OH), alkyl-subst.* †,10



AR 1:  $C \geq 6$   $H \geq 1$   $O \geq 1$   
 DBE  $\geq 4$   
 AR 0: -  
 SR 1: at least 1 fragment aromaar-OH  
 SR 0: no fragments aromaar-OH

<sup>10</sup>Die Definition der SP ist nicht präzise. Die optionale Substitution mit einem Alkylrest kann nicht berücksichtigt werden.

**phen:** *phenol (1-3 OH), alkyl-subst.* <sup>†,11</sup>



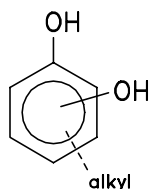
AR 1:  $C \geq 6$   $H \geq 1$   $O \geq 1$   
DBE  $\geq 4$

AR 0: -

SR 1: at least 1 fragment aromaar-OH

SR 0: no fragments aromaar-OH

**phen-2-OH:** *phenol (2 OH), alkyl-subst.* <sup>†</sup>



AR 1:  $C \geq 6$   $H \geq 2$   $O \geq 2$   
DBE  $\geq 4$

AR 0: -

SR 1: 1-4 terms of

at least 1 fragment aromaphen-2-OHa,

at least 1 fragment aromaphen-2-OHb,

at least 1 fragment aromaphen-2-OHc,

at least 1 fragment aromaphen-2-OHd,

SR 0: exactly 4 terms of

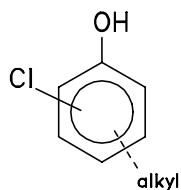
no fragments aromaphen-2-OHa,

no fragments aromaphen-2-OHb,

no fragments aromaphen-2-OHc,

no fragments aromaphen-2-OHd

**phen1-Cl1:** *phenol - Cl (1 OH, 1 Cl), alkyl-subst.* <sup>†,‡</sup>



AR 1:  $C \geq 6$   $H \geq 1$   $Cl \geq 1$   $O \geq 1$   
DBE  $\geq 4$

AR 0: -

SR 1: 1-5 terms of

at least 1 fragment aromaphen1-Cl1a,

at least 1 fragment aromaphen1-Cl1b,

at least 1 fragment aromaphen1-Cl1c,

at least 1 fragment aromaphen1-Cl1d,

at least 1 fragment aromaphen1-Cl1e

SR 0: exactly 5 terms of

no fragments aromaphen1-Cl1a,

no fragments aromaphen1-Cl1b,

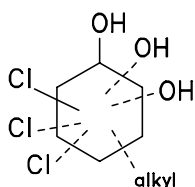
no fragments aromaphen1-Cl1c,

no fragments aromaphen1-Cl1d,

no fragments aromaphen1-Cl1e

<sup>11</sup>Die Definition der SP ist nicht präzise. Die optionale Substitution mit zwei weiteren OH-Gruppen und einem Alkylrest kann nicht berücksichtigt werden.

**phen-Cl:** *phenol - Cl (1-3 OH, 1-3 Cl), alkyl-subst.* †‡



AR 1:  $C \geq 6$   $H \geq 1$   $Cl \geq 1$   $O \geq 1$   
 DBE  $\geq 4$

AR 0: -

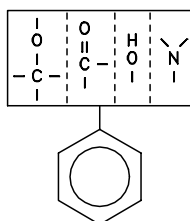
SR 1: 1-5 terms of

at least 1 fragment aromaphen1-Cl1a,  
 at least 1 fragment aromaphen1-Cl1b,  
 at least 1 fragment aromaphen1-Cl1c,  
 at least 1 fragment aromaphen1-Cl1d,  
 at least 1 fragment aromaphen1-Cl1e

SR 0: exactly 5 terms of

no fragments aromaphen1-Cl1a,  
 no fragments aromaphen1-Cl1b,  
 no fragments aromaphen1-Cl1c,  
 no fragments aromaphen1-Cl1d,  
 no fragments aromaphen1-Cl1e

**CO-C<sub>6</sub>H<sub>3</sub>-O:** *tri-subst. benzene ring: -C=O or -C-O or -OH or -N*



AR 1:  $C \geq 7$   $H \geq 3$   $O \geq 2$   
 DBE  $\geq 4$

AR 0: -

SR 1: 1-3 terms of

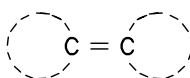
at least 1 fragment aromaar-COsd,  
 at least 1 fragment aromaar-OH,  
 at least 1 fragment aromaar-Nsp,

SR 0: exactly 5 terms of

no fragments aromaar-COsd,  
 no fragments aromaar-OH,  
 no fragments aromaar-Nsp

## C.3 Bindungen

**r>C=C<r:** *C=C between 2 rings*



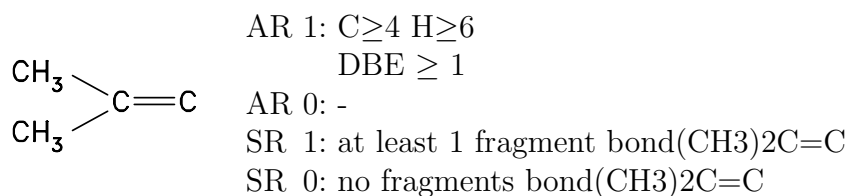
AR 1:  $C \geq 2$

DBE  $\geq 3$

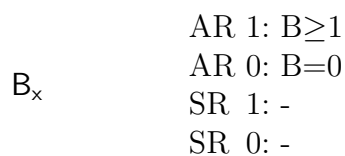
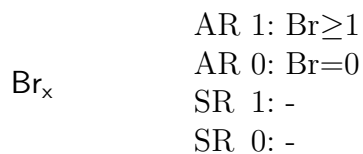
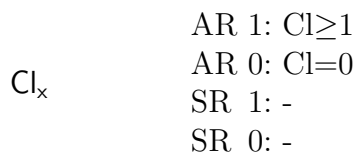
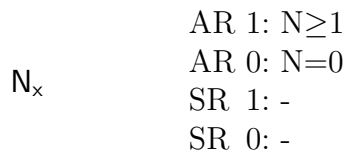
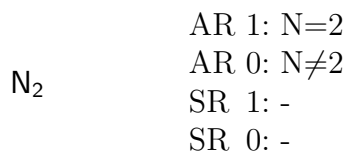
AR 0: -

SR 1: at least 1 fragment bondrC=Cr

SR 0: no fragments bondrC=Cr

**(CH<sub>3</sub>)<sub>2</sub>>C=C:** *isopropylidene*

## C.4 Elemente

**B:** *boron (any number)* ‡**Br:** *bromine (any number)***Cl:** *chlorine (any number)***N:** *nitrogen (any number)***N 2:** *nitrogen (any number)*

**P:** *phosphorous (any number)*

	AR 1: $P \geq 1$
$P_x$	AR 0: $P = 0$
	SR 1: -
	SR 0: -

**Si:** *silicon (any number)*

	AR 1: $Si \geq 1$
$Si_x$	AR 0: $Si = 0$
	SR 1: -
	SR 0: -

**Si 1:** *silicon: 1 atom*

	AR 1: $Si = 1$
$Si_1$	AR 0: $Si \neq 1$
	SR 1: -
	SR 0: -

**Si $\geq$ 2:** *silicon:  $\geq 2$  atoms*

	AR 1: $Si \geq 2$
$Si_{\geq 2}$	AR 0: $Si \leq 1$
	SR 1: -
	SR 0: -

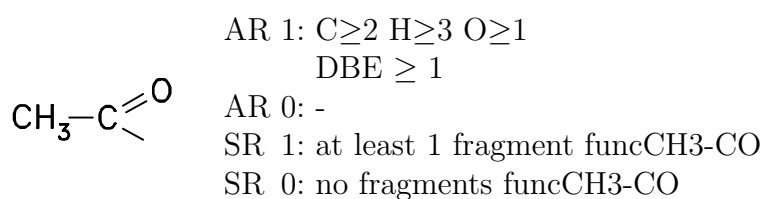
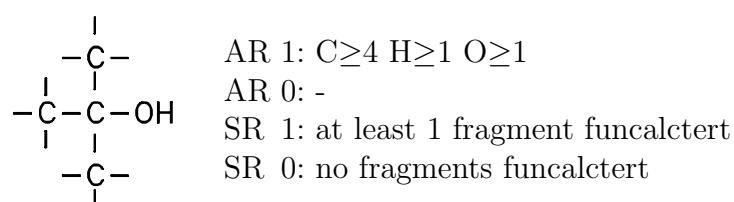
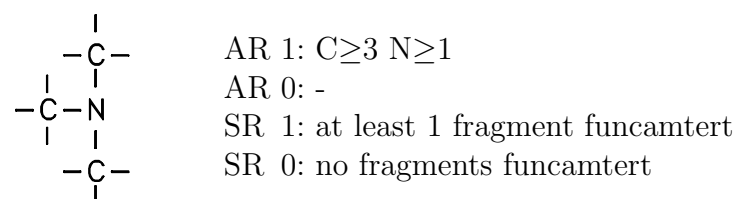
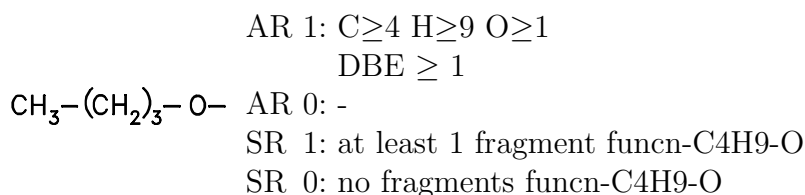
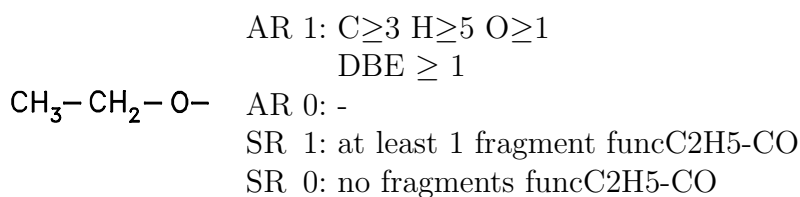
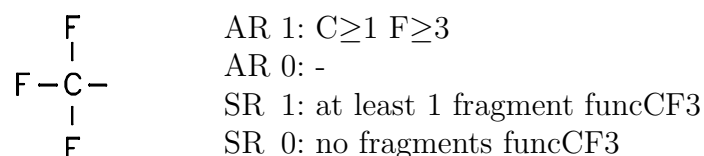
**S:** *sulphur (any number)*

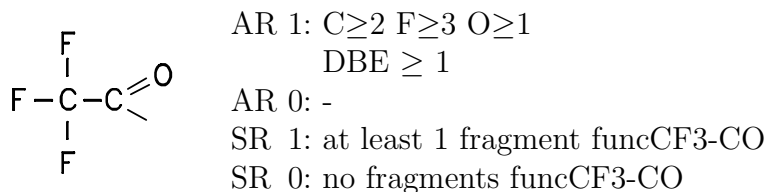
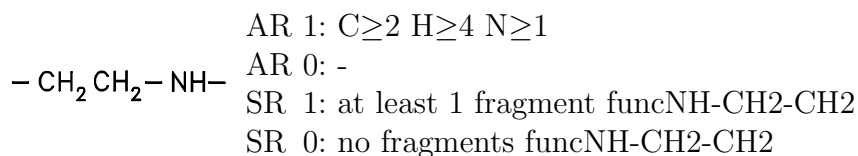
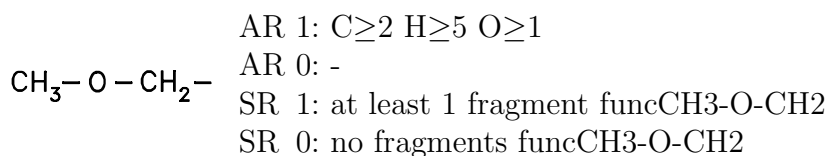
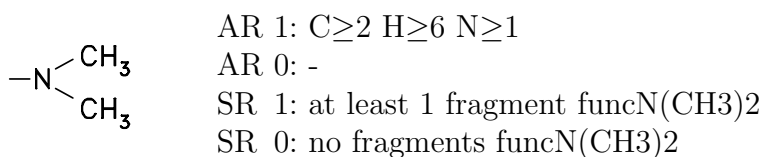
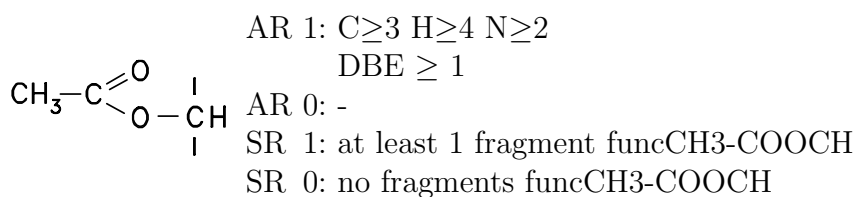
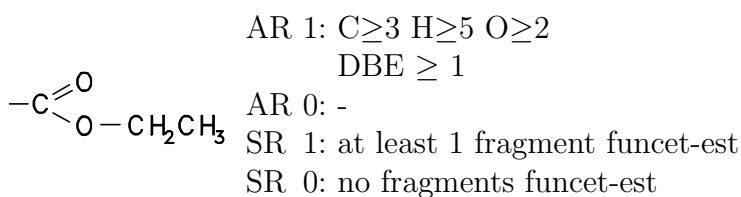
	AR 1: $S \geq 1$
$S_x$	AR 0: $S = 0$
	SR 1: -
	SR 0: -

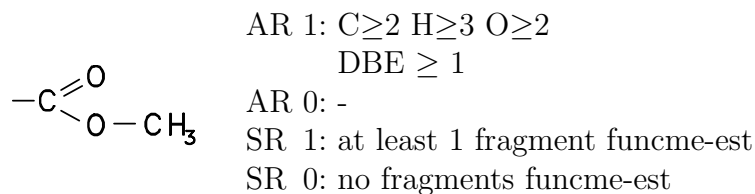
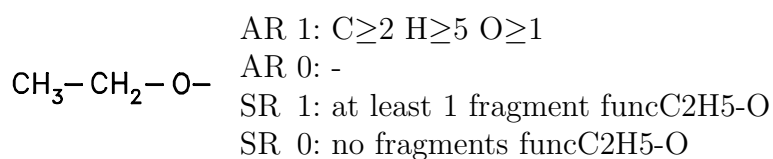
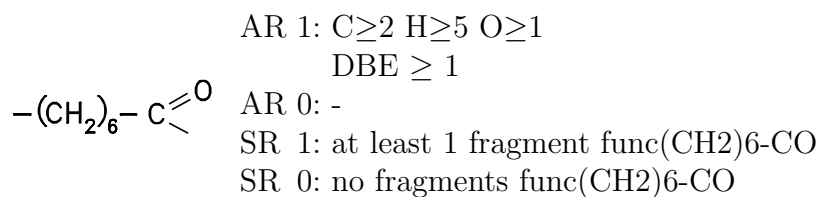
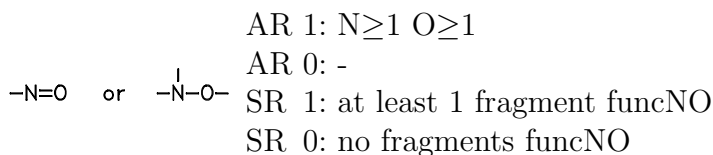
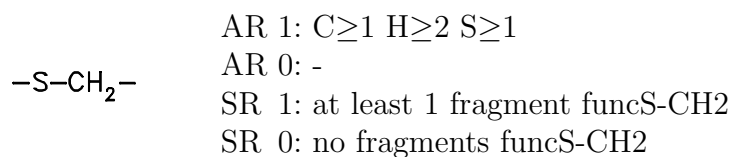
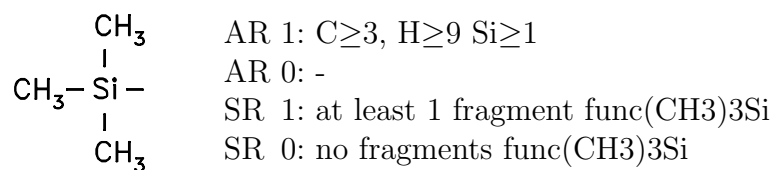
## C.5 Funktionelle Gruppen

**CH<sub>3</sub>-COO:** *acetoxy CH<sub>3</sub>-COO*

	AR 1: $C \geq 2$ $H \geq 3$ $O \geq 2$
	DBE $\geq 1$
$CH_3-C \begin{array}{l} \diagup O \\ \diagdown O- \end{array}$	AR 0: -
	SR 1: at least 1 fragment funcCH <sub>3</sub> -COO
	SR 0: no fragments funcCH <sub>3</sub> -COO

**CH<sub>3</sub>-CO:** *acetyl CH<sub>3</sub>-CO***alc tert:** *alcohol tertiary (no ester)***am tert:** *amine tertiary (no amide)***n-C<sub>4</sub>H<sub>9</sub>-O:** *butyl-oxy n-C<sub>4</sub>H<sub>9</sub> - O***C<sub>2</sub>H<sub>5</sub>-CO:** *C<sub>2</sub>H<sub>5</sub> - CO***CF<sub>3</sub>:** *CF<sub>3</sub> trifluoromethyl*

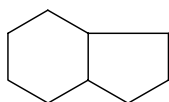
**CF<sub>3</sub>-CO:** *CF<sub>3</sub> - CO***NH-CH<sub>2</sub>-CH<sub>2</sub>:** *CH<sub>2</sub> - CH<sub>2</sub> - NH***CH<sub>3</sub>-O-CH<sub>2</sub>:** *CH<sub>3</sub> - O - CH<sub>2</sub>***N(CH<sub>3</sub>)<sub>2</sub>:** *dimethyl-amine -N(CH<sub>3</sub>)<sub>2</sub>***CH<sub>3</sub>-COOCH:** *ester of acetic acid CH<sub>3</sub>COO-CH<sub>2</sub>***et-est:** *ester: ethyl*

**me-est:** *ester: methyl***C2H5-O:** *ethoxy***(CH<sub>2</sub>)<sub>6</sub>-CO:** *ketone (CH<sub>2</sub>)<sub>6</sub> - CO***NO:** *nitrogen-oxygen bond***S-CH<sub>2</sub>:** *S - CH<sub>2</sub>***(CH<sub>3</sub>)<sub>3</sub> Si:** *trimethylsilyl*



## C.6 Ringe

**r 5+6:** 5-ring and 6-ring condensed <sup>12</sup>



AR 1: DBE  $\geq 1$

AR 0: -

SR 1: at least 1 fragment ringr5+6

SR 0: no fragments ringr5+6

---

<sup>12</sup>Die kondensierten Ringe werden als MMG unter Verwendung des Atomtyps *Beliebig* und mit Alternativen für die Bindungen dargestellt.



## Anhang D

# Bruttoformeln nach Masse und Ionentyp

Für ganzzahlige nominale Massen werden in den Tabellen D.1 – D.4 zunächst Gesamtzahlen von Bruttoformeln ohne weitere Einschränkungen angegeben. Die Spalte *Ionen* enthält die Anzahl von Bruttoformeln, die den Bedingungen (Gr2) und (Con) aus Satz 1.3.19 genügen. Die dritte Spalte enthält die Anzahl von Bruttoformeln, die zusätzlich Bedingung (Gr1) entsprechen. Dies sind gerade die Bruttoformeln, die für die Molekülon im MS in Frage kommen. Die nächste Spalte gibt schließlich die Anzahl derjenigen Bruttoformeln an, die (Gr2) und (Con) genügen, (Gr1) jedoch nicht erfüllen. Die CPU-Zeiten wurden auf einem PC mit PIII/850MHz unter Windows NT4 gemessen.

312 ANHANG D. BRUTTOFORMELN NACH MASSE UND IONENTYP

Masse	Gesamt	Ionen	OEI	E EI
1	1	0	0	0
2	1	1	1	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0
6	1	0	0	0
7	1	0	0	0
8	1	0	0	0
9	1	0	0	0
10	1	0	0	0
11	1	0	0	0
12	2	0	0	0
13	2	0	0	0
14	3	0	0	0
15	3	0	0	0
16	4	1	1	0
17	4	1	1	0
18	4	1	1	0
19	4	0	0	0
20	4	0	0	0
21	4	0	0	0
22	4	0	0	0
23	4	0	0	0
24	5	1	1	0
25	5	1	0	1
26	6	1	1	0
27	6	2	1	1
28	8	3	2	1
29	8	3	1	2
30	9	4	3	1
31	9	4	2	2
32	10	4	3	1
33	10	2	1	1
34	10	1	1	0
35	10	0	0	0
36	11	1	1	0
37	11	1	0	1
38	12	2	1	1
39	12	2	1	1
40	14	4	3	1
41	14	4	1	3
42	16	6	3	3
43	16	6	3	3
44	18	8	5	3
45	18	7	3	4
46	19	7	4	3
47	19	5	3	2
48	21	5	4	1
49	21	3	1	2
50	22	3	2	1

Masse	Gesamt	Ionen	OEI	E EI
51	22	2	1	1
52	24	4	3	1
53	24	4	1	3
54	26	6	3	3
55	26	6	3	3
56	29	9	6	3
57	29	9	3	6
58	31	11	6	5
59	31	10	5	5
60	34	12	8	4
61	34	10	4	6
62	36	10	6	4
63	36	7	4	3
64	39	8	6	2
65	39	6	2	4
66	41	7	4	3
67	41	6	3	3
68	44	9	6	3
69	44	9	3	6
70	47	12	6	6
71	47	12	6	6
72	51	16	10	6
73	51	15	6	9
74	54	17	9	8
75	54	15	8	7
76	58	17	11	6
77	58	14	6	8
78	61	14	8	6
79	61	11	6	5
80	65	13	9	4
81	65	11	4	7
82	68	13	7	6
83	68	12	6	6
84	73	17	11	6
85	73	17	6	11
86	77	21	11	10
87	77	20	10	10
88	82	24	15	9
89	82	22	9	13
90	86	24	13	11
91	86	21	11	10
92	91	23	15	8
93	91	19	8	11
94	95	20	11	9
95	95	17	9	8
96	101	21	14	7
97	101	19	7	12
98	106	23	12	11
99	106	22	11	11
100	112	28	17	11

Tabelle D.1: Anzahlen von Bruttoformeln zu nominalen Massen von 1 bis 100 mit Elementen aus  $\mathcal{E}_4$

Masse	Gesamt	Ionen	OEI	EEl
1	1	0	0	0
2	1	1	1	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0
6	1	0	0	0
7	1	0	0	0
8	1	0	0	0
9	1	0	0	0
10	1	0	0	0
11	1	0	0	0
12	2	0	0	0
13	2	0	0	0
14	3	0	0	0
15	3	0	0	0
16	4	1	1	0
17	4	1	1	0
18	4	1	1	0
19	5	0	0	0
20	5	1	1	0
21	5	0	0	0
22	5	0	0	0
23	5	0	0	0
24	6	1	1	0
25	6	1	0	1
26	7	1	1	0
27	7	2	1	1
28	10	3	2	1
29	10	3	1	2
30	11	4	3	1
31	13	4	2	2
32	15	5	4	1
33	16	2	1	1
34	16	4	4	0
35	18	1	1	0
36	19	3	3	0
37	19	1	0	1
38	21	3	2	1
39	21	2	1	1
40	24	5	4	1
41	24	5	1	4
42	27	7	4	3
43	29	9	4	5
44	33	12	8	4
45	35	13	6	7
46	37	15	9	6
47	42	14	7	7
48	45	16	12	4
49	47	11	5	6
50	50	10	8	2
51	53	5	3	2
52	57	10	9	1
53	57	6	2	4
54	62	10	6	4
55	64	10	4	6
56	71	17	13	4
57	74	19	6	13
58	78	23	13	10
59	85	28	10	18
60	92	34	25	9
61	97	36	14	22
62	103	40	26	14
63	112	35	14	21
64	120	41	31	10
65	123	31	13	18
66	132	31	22	9
67	138	21	9	12
68	147	28	23	5
69	151	21	8	13
70	162	32	19	13
71	170	34	14	20
72	180	43	30	13
73	189	49	20	29
74	199	58	32	26
75	213	69	26	43
76	227	79	53	26
77	236	79	35	44
78	251	85	53	32
79	266	80	32	48
80	282	87	62	25
81	291	71	31	40
82	309	72	49	23
83	323	62	25	37
84	341	71	53	18
85	354	63	26	37
86	372	78	48	30
87	391	85	32	53
88	412	103	71	32
89	430	112	46	66
90	452	130	77	53
91	477	145	53	92
92	504	161	108	53
93	523	159	67	92
94	553	174	113	61
95	581	166	65	101
96	610	172	121	51
97	632	151	63	88
98	667	161	108	53
99	696	145	59	86
100	727	158	110	48

Tabelle D.2: Anzahlen von Bruttoformeln zu nominalen Massen von 1 bis 100 mit Elementen aus  $\mathcal{E}_{11}$

314 ANHANG D. BRUTTOFORMELN NACH MASSE UND IONENTYP

Masse	Gesamt	Ionen	OEI	E EI	CPU
150	313	73	38	35	0,00 s
200	677	151	87	64	0,00 s
250	1244	270	138	132	0,00 s
300	2068	448	248	200	0,00 s
350	3188	676	344	332	0,01 s
400	4657	985	533	452	0,02 s
450	6515	1371	695	676	0,03 s
500	8815	1843	983	860	0,04 s
600	14916	3102	1639	1463	0,06 s
700	23332	4824	2530	2294	0,09 s
800	34433	7089	3697	3392	0,16 s
900	48591	9977	5180	4797	0,25 s
1000	66180	13552	7011	6541	0,37 s

Tabelle D.3: Anzahlen von Bruttoformeln zu nominalen Massen über 100 mit Elementen aus  $\mathcal{E}_4$

Masse	Gesamt	Ionen	OEI	E EI	CPU
150	5299	1259	764	495	0,00 s
200	26263	6383	3797	2586	0,03 s
250	101339	24162	14140	10022	0,12 s
300	327411	76144	43861	32283	0,43 s
350	925843	211769	120387	91382	1,33 s
400	2357940	533418	299361	234057	3,69 s
450	5518977	1238647	688903	549744	9,33 s
500	12045750	2685131	1481067	1204064	21,95 s
600	48507196	10684233	5821545	4862688	100,69 s
700	163873929	35750618	19293812	16456806	360,39 s
800	483527540	104698838	56072098	48626740	1165,95 s
900	1280954355	275682153	146725060	128957093	3373,54 s
1000	3107850498	665461540	352344362	313117178	8689,85 s

Tabelle D.4: Anzahlen von Bruttoformeln zu nominalen Massen über 100 mit Elementen aus  $\mathcal{E}_{11}$

# Anhang E

## Isomere nach Bruttoformel und Masse

In den folgenden Tabellen werden sortiert nach aufsteigender Masse  $m$  alle Bruttoformeln  $\beta$  zu Elementen aus  $\mathcal{E}_4$  mit mindestens einem C-Atom und die Anzahl von Konstitutionsisomeren  $|\bar{\mathcal{M}}_\beta^C|$  angegeben. Spalte  $BS$  enthält die Anzahl der in der *Beilstein*-Datenbank (Abschnitt 1.7) aufgelisteten Konstitutionsisomere, Spalte  $MS$  entsprechende Angaben bezogen auf die in Abschnitt 5.3.5 vorgestellte MS-Struktur-Datenbasis. Man beachte, dass es sich bei den letzten beiden Spalten lediglich um Momentaufnahmen handelt. Sie sind aber in jedem Fall nützlich, um einen Eindruck von den Anzahlen der mathematisch möglichen, als existent nachgewiesenen und massenspektrometrisch vermessenen Konstitutionsisomere zu gewinnen.

$m$	$\beta$	$ \bar{\mathcal{M}}_\beta^C $	$BS$	$MS$
16	CH <sub>4</sub>	1	1	1
24	C <sub>2</sub>	0	0	0
26	C <sub>2</sub> H <sub>2</sub>	1	1	1
27	CHN	1	1	1
28	C <sub>2</sub> H <sub>4</sub>	1	1	1
29	CH <sub>3</sub> N	1	1	0
30	CH <sub>2</sub> O	1	1	1
	C <sub>2</sub> H <sub>6</sub>	1	1	1
31	CH <sub>5</sub> N	1	1	1
32	CH <sub>4</sub> O	1	1	1
36	C <sub>3</sub>	1	0	0
38	C <sub>3</sub> H <sub>2</sub>	2	1	0
39	C <sub>2</sub> HN	2	0	0
40	CN <sub>2</sub>	1	0	0
	C <sub>2</sub> O	1	0	0
	C <sub>3</sub> H <sub>4</sub>	3	3	3
41	C <sub>2</sub> H <sub>3</sub> N	5	5	1
42	CH <sub>2</sub> N <sub>2</sub>	4	4	0
	C <sub>2</sub> H <sub>2</sub> O	3	3	1
	C <sub>3</sub> H <sub>6</sub>	2	2	2
43	CHNO	3	3	0

$m$	$\beta$	$ \bar{\mathcal{M}}_\beta^C $	$BS$	$MS$
	C <sub>2</sub> H <sub>5</sub> N	4	4	1
44	CO <sub>2</sub>	1	1	1
	CH <sub>4</sub> N <sub>2</sub>	4	4	0
	C <sub>2</sub> H <sub>4</sub> O	3	3	2
	C <sub>3</sub> H <sub>8</sub>	1	1	1
45	CH <sub>3</sub> NO	5	5	2
	C <sub>2</sub> H <sub>7</sub> N	2	2	2
46	CH <sub>2</sub> O <sub>2</sub>	2	2	1
	CH <sub>6</sub> N <sub>2</sub>	2	2	1
	C <sub>2</sub> H <sub>6</sub> O	2	2	2
47	CH <sub>5</sub> NO	3	3	1
48	CH <sub>4</sub> O <sub>2</sub>	2	2	0
	C <sub>4</sub>	3	0	0
50	C <sub>4</sub> H <sub>2</sub>	7	1	1
51	C <sub>3</sub> HN	7	1	1
52	C <sub>2</sub> N <sub>2</sub>	5	1	1
	C <sub>3</sub> O	2	0	0
	C <sub>4</sub> H <sub>4</sub>	11	7	1
53	C <sub>3</sub> H <sub>3</sub> N	19	6	1
54	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub>	19	4	0
	C <sub>3</sub> H <sub>2</sub> O	9	3	0

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	$C_4H_6$	9	9	7
55	$CHN_3$	6	1	0
	$C_2HNO$	11	1	0
	$C_3H_5N$	21	13	2
56	$CN_2O$	4	1	0
	$C_2O_2$	3	1	0
	$C_2H_4N_2$	27	9	0
	$C_3H_4O$	13	13	2
	$C_4H_8$	5	5	4
57	$CH_3N_3$	13	0	0
	$C_2H_3NO$	26	6	2
	$C_3H_7N$	12	12	6
58	$CH_2N_2O$	18	4	0
	$C_2H_2O_2$	9	3	1
	$C_2H_6N_2$	18	10	1
	$C_3H_6O$	9	9	6
	$C_4H_{10}$	2	2	2
59	$CHNO_2$	8	2	0
	$CH_5N_3$	11	4	1
	$C_2H_5NO$	22	10	3
	$C_3H_9N$	4	4	4
60	$CO_3$	1	1	0
	$CH_4N_2O$	21	7	2
	$C_2H_4O_2$	10	9	3
	$C_2H_8N_2$	6	5	4
	$C_3H_8O$	3	3	3
	$C_5$	6	0	0
61	$CH_3NO_2$	15	5	1
	$CH_7N_3$	4	1	0
	$C_2H_7NO$	8	7	3
62	$CH_2O_3$	4	2	0
	$CH_6N_2O$	8	2	0
	$C_2H_6O_2$	5	5	2
	$C_5H_2$	21	0	0
63	$CH_5NO_2$	8	1	0
	$C_4HN$	27	0	0
64	$CH_4O_3$	3	3	0
	$C_3N_2$	14	0	0
	$C_4O$	7	0	0
	$C_5H_4$	40	8	0
65	$C_4H_3N$	87	7	0
66	$C_3H_2N_2$	86	8	1
	$C_4H_2O$	36	2	0
	$C_5H_6$	40	20	4
67	$C_2HN_3$	34	3	0
	$C_3HNO$	46	1	0
	$C_4H_5N$	116	12	5
68	$CN_4$	6	0	0
	$C_2N_2O$	20	2	0
	$C_3O_2$	7	1	1
	$C_3H_4N_2$	155	19	5
	$C_4H_4O$	62	19	2
	$C_5H_8$	26	25	16
69	$C_2H_3N_3$	99	10	2
	$C_3H_3NO$	136	13	4
	$C_4H_7N$	85	30	6
70	$CH_2N_4$	31	4	1
	$C_2H_2N_2O$	114	8	2
	$C_3H_2O_2$	34	5	1
	$C_3H_6N_2$	136	23	2
	$C_4H_6O$	55	34	15
	$C_5H_{10}$	10	10	10

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
71	$CHN_3O$	34	2	0
	$C_2HNO_2$	40	4	0
	$C_2H_5N_3$	110	4	0
	$C_3H_5NO$	154	24	6
	$C_4H_9N$	35	32	10
72	$CN_2O_2$	12	0	0
	$CH_4N_4$	47	1	0
	$C_2O_3$	5	1	0
	$C_2H_4N_2O$	177	7	0
	$C_3H_4O_2$	52	15	4
	$C_3H_8N_2$	62	22	3
	$C_4H_8O$	26	26	17
	$C_5H_{12}$	3	3	3
	$C_6$	19	0	0
73	$CH_3N_3O$	86	2	0
	$C_2H_3NO_2$	99	10	0
	$C_2H_7N_3$	58	7	1
	$C_3H_7NO$	84	34	8
	$C_4H_{11}N$	8	8	7
74	$CH_2N_2O_2$	65	1	0
	$CH_6N_4$	29	2	1
	$C_2H_2O_3$	20	2	1
	$C_2H_6N_2O$	115	20	5
	$C_3H_6O_2$	34	21	7
	$C_3H_{10}N_2$	14	11	6
	$C_4H_{10}O$	7	7	7
	$C_6H_2$	85	1	0
75	$CHNO_3$	18	1	0
	$CH_5N_3O$	71	4	1
	$C_2H_5NO_2$	84	18	5
	$C_2H_9N_3$	14	2	0
	$C_3H_9NO$	21	18	6
	$C_5HN$	112	1	0
76	$CO_4$	2	0	0
	$CH_4N_2O_2$	75	5	1
	$CH_8N_4$	8	1	0
	$C_2H_4O_3$	22	9	1
	$C_2H_8N_2O$	31	9	1
	$C_3H_8O_2$	11	11	5
	$C_4N_2$	64	1	1
	$C_5O$	21	0	0
	$C_6H_4$	185	6	0
77	$CH_3NO_3$	34	1	0
	$CH_7N_3O$	21	0	0
	$C_2H_7NO_2$	28	7	1
	$C_5H_3N$	437	3	0
78	$CH_2O_4$	6	0	0
	$CH_6N_2O_2$	28	1	0
	$C_2H_6O_3$	10	8	0
	$C_4H_2N_2$	465	2	2
	$C_5H_2O$	151	2	0
	$C_6H_6$	217	29	5
79	$CH_5NO_3$	17	2	0
	$C_3HN_3$	194	1	0
	$C_4HNO$	216	1	0
	$C_5H_5N$	685	17	2
80	$CH_4O_4$	5	1	0
	$C_2N_4$	42	1	0
	$C_3N_2O$	88	1	0
	$C_4O_2$	28	0	0
	$C_4H_4N_2$	1005	12	5
	$C_5H_4O$	318	14	0



$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	$C_6H_8$	159	71	18
81	$C_3H_3N_3$	706	6	2
	$C_4H_3NO$	775	12	0
	$C_5H_7N$	593	33	9
82	$C_2H_2N_4$	272	3	1
	$C_3H_2N_2O$	703	5	0
	$C_4H_2O_2$	163	4	0
	$C_4H_6N_2$	1058	62	9
	$C_5H_6O$	337	63	9
	$C_6H_{10}$	77	69	33
83	$CHN_5$	42	1	0
	$C_2HN_3O$	256	1	0
	$C_3HNO_2$	202	1	0
	$C_3H_5N_3$	969	24	5
	$C_4H_5NO$	1069	47	2
	$C_5H_9N$	313	73	12
84	$CN_4O$	32	0	0
	$C_2N_2O_2$	76	1	0
	$C_2H_4N_4$	512	23	6
	$C_3O_3$	16	0	0
	$C_3H_4N_2O$	1371	29	4
	$C_4H_4O_2$	301	29	5
	$C_4H_8N_2$	633	59	10
	$C_5H_8O$	205	110	31
	$C_6H_{12}$	25	25	22
	$C_7$	50	0	0
85	$CH_3N_5$	131	7	2
	$C_2H_3N_3O$	826	19	0
	$C_3H_3NO_2$	641	16	1
	$C_3H_7N_3$	681	13	0
	$C_4H_7NO$	764	65	14
	$C_5H_{11}N$	100	69	15
86	$CH_2N_4O$	227	5	0
	$C_2H_2N_2O_2$	506	10	0
	$C_2H_6N_4$	439	4	1
	$C_3H_2O_3$	98	4	1
	$C_3H_6N_2O$	1194	32	3
	$C_4H_6O_2$	263	61	15
	$C_4H_{10}N_2$	218	61	8
	$C_5H_{10}O$	74	74	44
	$C_6H_{14}$	5	5	5
	$C_7H_2$	356	0	0
87	$CHN_3O_2$	137	1	0
	$CH_5N_5$	145	0	0
	$C_2HNO_3$	110	0	0
	$C_2H_5N_3O$	935	5	0
	$C_3H_5NO_2$	732	42	1
	$C_3H_9N_3$	259	16	1
	$C_4H_9NO$	299	85	15
	$C_5H_{13}N$	17	17	16
	$C_6HN$	540	0	0
88	$CN_2O_3$	29	0	0
	$CH_4N_4O$	361	2	0
	$C_2O_4$	10	2	0
	$C_2H_4N_2O_2$	807	15	3
	$C_2H_8N_4$	189	6	0
	$C_3H_4O_3$	152	15	4
	$C_3H_8N_2O$	527	40	4
	$C_4H_8O_2$	122	59	23
	$C_4H_{12}N_2$	38	28	17
	$C_5N_2$	271	0	0
	$C_5H_{12}O$	14	14	14

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$	
	$C_6O$	85	0	0	
	$C_7H_4$	920	3	0	
89	$CH_3N_3O_2$	369	1	0	
	$CH_7N_5$	73	1	0	
	$C_2H_3NO_3$	288	6	0	
	$C_2H_7N_3O$	481	13	0	
	$C_3H_7NO_2$	391	45	9	
	$C_3H_{11}N_3$	45	5	0	
	$C_4H_{11}NO$	56	42	11	
	$C_6H_3N$	2447	1	0	
90	$CH_2N_2O_3$	173	0	0	
	$CH_6N_4O$	225	1	1	
	$C_2H_2O_4$	41	2	1	
	$C_2H_6N_2O_2$	521	21	2	
	$C_2H_{10}N_4$	37	2	0	
	$C_3H_6O_3$	102	20	8	
	$C_3H_{10}N_2O$	102	17	3	
	$C_4H_{10}O_2$	28	26	12	
	$C_5H_2N_2$	2652	1	0	
	$C_6H_2O$	738	0	0	
	$C_7H_6$	1230	17	1	
	91	$CHNO_4$	34	0	0
$CH_5N_3O_2$		306	2	0	
$CH_9N_5$		15	0	0	
$C_2H_5NO_3$		246	9	1	
$C_2H_9N_3O$		101	0	0	
$C_3H_9NO_2$		90	20	3	
$C_4HN_3$		1224	2	0	
$C_5HNO$		1111	0	0	
$C_6H_5N$	4394	22	0		
92	$CO_5$	2	0	0	
	$CH_4N_2O_3$	207	2	0	
	$CH_8N_4O$	52	1	0	
	$C_2H_4O_4$	48	5	0	
	$C_2H_8N_2O_2$	132	1	0	
	$C_3N_4$	235	1	0	
	$C_3H_8O_3$	28	15	1	
	$C_4N_2O$	475	1	0	
	$C_5O_2$	98	1	0	
	$C_5H_4N_2$	6763	14	4	
	$C_6H_4O$	1823	15	1	
$C_7H_8$	1031	81	13		
93	$CH_3NO_4$	68	0	0	
	$CH_7N_3O_2$	86	0	0	
	$C_2H_7NO_3$	76	1	0	
	$C_4H_3N_3$	5245	15	4	
	$C_5H_3NO$	4738	8	1	
	$C_6H_7N$	4378	69	7	
94	$CH_2O_5$	9	1	0	
	$CH_6N_2O_3$	73	0	0	
	$C_2H_6O_4$	20	5	0	
	$C_3H_2N_4$	2165	9	0	
	$C_4H_2N_2O$	4628	11	0	
	$C_5H_2O_2$	812	1	0	
	$C_5H_6N_2$	8341	39	15	
	$C_6H_6O$	2237	65	3	
$C_7H_{10}$	575	183	27		
95	$CH_5NO_4$	33	0	0	
	$C_2HN_5$	425	2	0	
	$C_3HN_3O$	1861	5	0	
	$C_4HNO_2$	1127	0	0	
	$C_4H_5N_3$	8528	31	8	

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	C <sub>5</sub> H <sub>5</sub> NO	7687	43	5
	C <sub>6</sub> H <sub>9</sub> N	2732	72	9
96	CN <sub>6</sub>	35	1	0
	CH <sub>4</sub> O <sub>5</sub>	6	1	0
	C <sub>2</sub> N <sub>4</sub> O	280	0	0
	C <sub>3</sub> N <sub>2</sub> O <sub>2</sub>	412	0	0
	C <sub>3</sub> H <sub>4</sub> N <sub>4</sub>	5016	10	0
	C <sub>4</sub> O <sub>3</sub>	72	0	0
	C <sub>4</sub> H <sub>4</sub> N <sub>2</sub> O	10770	40	8
	C <sub>5</sub> H <sub>4</sub> O <sub>2</sub>	1821	22	6
	C <sub>5</sub> H <sub>8</sub> N <sub>2</sub>	5984	106	17
	C <sub>6</sub> H <sub>8</sub> O	1623	217	22
	C <sub>7</sub> H <sub>12</sub>	222	153	59
	C <sub>8</sub>	204	0	0
97	C <sub>2</sub> H <sub>3</sub> N <sub>5</sub>	1630	1	0
	C <sub>3</sub> H <sub>3</sub> N <sub>3</sub> O	7341	10	1
	C <sub>4</sub> H <sub>3</sub> NO <sub>2</sub>	4332	18	1
	C <sub>4</sub> H <sub>7</sub> N <sub>3</sub>	7301	65	5
	C <sub>5</sub> H <sub>7</sub> NO	6637	141	12
	C <sub>6</sub> H <sub>11</sub> N	1111	154	16
98	CH <sub>2</sub> N <sub>6</sub>	270	1	0
	C <sub>2</sub> H <sub>2</sub> N <sub>4</sub> O	2489	1	0
	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub> O <sub>2</sub>	3734	4	0
	C <sub>3</sub> H <sub>6</sub> N <sub>4</sub>	5328	38	7
	C <sub>4</sub> H <sub>2</sub> O <sub>3</sub>	551	5	1
	C <sub>4</sub> H <sub>6</sub> N <sub>2</sub> O	11514	93	13
	C <sub>5</sub> H <sub>6</sub> O <sub>2</sub>	1938	112	21
	C <sub>5</sub> H <sub>10</sub> N <sub>2</sub>	2668	128	14
	C <sub>6</sub> H <sub>10</sub> O	747	349	83
	C <sub>7</sub> H <sub>14</sub>	56	55	38
	C <sub>8</sub> H <sub>2</sub>	1804	1	0
99	CHN <sub>5</sub> O	352	0	0
	C <sub>2</sub> HN <sub>3</sub> O <sub>2</sub>	1258	1	0
	C <sub>2</sub> H <sub>5</sub> N <sub>5</sub>	2274	16	3
	C <sub>3</sub> HNO <sub>3</sub>	677	1	0
	C <sub>3</sub> H <sub>5</sub> N <sub>3</sub> O	10363	58	1
	C <sub>4</sub> H <sub>5</sub> NO <sub>2</sub>	6102	59	6
	C <sub>4</sub> H <sub>9</sub> N <sub>3</sub>	3654	39	0
	C <sub>5</sub> H <sub>9</sub> NO	3390	184	27
	C <sub>6</sub> H <sub>13</sub> N	284	143	24
	C <sub>7</sub> HN	2879	1	0
100	CN <sub>4</sub> O <sub>2</sub>	139	0	0
	CH <sub>4</sub> N <sub>6</sub>	528	3	0
	C <sub>2</sub> N <sub>2</sub> O <sub>3</sub>	215	0	0
	C <sub>2</sub> H <sub>4</sub> N <sub>4</sub> O	5039	18	1
	C <sub>3</sub> O <sub>4</sub>	36	0	0
	C <sub>3</sub> H <sub>4</sub> N <sub>2</sub> O <sub>2</sub>	7547	29	2
	C <sub>3</sub> H <sub>8</sub> N <sub>4</sub>	3080	8	0
	C <sub>4</sub> H <sub>4</sub> O <sub>3</sub>	1073	29	5
	C <sub>4</sub> H <sub>8</sub> N <sub>2</sub> O	6754	103	4
	C <sub>5</sub> H <sub>8</sub> O <sub>2</sub>	1168	206	42
	C <sub>5</sub> H <sub>12</sub> N <sub>2</sub>	716	116	12
	C <sub>6</sub> N <sub>2</sub>	1448	1	0
	C <sub>6</sub> H <sub>12</sub> O	211	188	79
	C <sub>7</sub> O	356	0	0
	C <sub>7</sub> H <sub>16</sub>	9	9	9
	C <sub>8</sub> H <sub>4</sub>	5308	0	0
101	CH <sub>3</sub> N <sub>5</sub> O	1206	1	0
	C <sub>2</sub> H <sub>3</sub> N <sub>3</sub> O <sub>2</sub>	4315	9	0
	C <sub>2</sub> H <sub>7</sub> N <sub>5</sub>	1567	2	0
	C <sub>3</sub> H <sub>3</sub> NO <sub>3</sub>	2279	10	0
	C <sub>3</sub> H <sub>7</sub> N <sub>3</sub> O	7227	16	1

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	C <sub>4</sub> H <sub>7</sub> NO <sub>2</sub>	4331	127	10
	C <sub>4</sub> H <sub>11</sub> N <sub>3</sub>	1055	26	0
	C <sub>5</sub> H <sub>11</sub> NO	1015	182	33
	C <sub>6</sub> H <sub>15</sub> N	39	39	18
	C <sub>7</sub> H <sub>3</sub> N	15052	0	0
102	CH <sub>2</sub> N <sub>4</sub> O <sub>2</sub>	1097	0	0
	CH <sub>6</sub> N <sub>6</sub>	447	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub> O <sub>3</sub>	1661	3	0
	C <sub>2</sub> H <sub>6</sub> N <sub>4</sub> O	4358	7	1
	C <sub>3</sub> H <sub>2</sub> O <sub>4</sub>	246	3	1
	C <sub>3</sub> H <sub>6</sub> N <sub>2</sub> O <sub>2</sub>	6618	52	5
	C <sub>3</sub> H <sub>10</sub> N <sub>4</sub>	978	11	0
	C <sub>4</sub> H <sub>6</sub> O <sub>3</sub>	949	68	7
	C <sub>4</sub> H <sub>10</sub> N <sub>2</sub> O	2201	90	10
	C <sub>5</sub> H <sub>10</sub> O <sub>2</sub>	400	181	41
	C <sub>5</sub> H <sub>14</sub> N <sub>2</sub>	97	58	15
	C <sub>6</sub> H <sub>2</sub> N <sub>2</sub>	16977	2	0
	C <sub>6</sub> H <sub>14</sub> O	32	32	32
	C <sub>7</sub> H <sub>2</sub> O	3971	0	0
	C <sub>8</sub> H <sub>6</sub>	7982	20	1
103	CHN <sub>3</sub> O <sub>3</sub>	421	0	0
	CH <sub>5</sub> N <sub>5</sub> O	1370	0	0
	C <sub>2</sub> HNO <sub>4</sub>	260	0	0
	C <sub>2</sub> H <sub>5</sub> N <sub>3</sub> O <sub>2</sub>	4978	10	3
	C <sub>2</sub> H <sub>9</sub> N <sub>5</sub>	560	3	0
	C <sub>3</sub> H <sub>5</sub> NO <sub>3</sub>	2644	19	0
	C <sub>3</sub> H <sub>9</sub> N <sub>3</sub> O	2630	28	0
	C <sub>4</sub> H <sub>9</sub> NO <sub>2</sub>	1640	102	21
	C <sub>4</sub> H <sub>13</sub> N <sub>3</sub>	146	13	1
	C <sub>5</sub> HN <sub>3</sub>	8172	1	0
	C <sub>5</sub> H <sub>13</sub> NO	149	93	13
	C <sub>6</sub> HNO	6340	0	0
	C <sub>7</sub> H <sub>5</sub> N	30478	11	3
104	CN <sub>2</sub> O <sub>4</sub>	61	0	0
	CH <sub>4</sub> N <sub>4</sub> O <sub>2</sub>	1819	1	0
	CH <sub>8</sub> N <sub>6</sub>	183	1	0
	C <sub>2</sub> O <sub>5</sub>	14	0	0
	C <sub>2</sub> H <sub>4</sub> N <sub>2</sub> O <sub>3</sub>	2763	8	0
	C <sub>2</sub> H <sub>8</sub> N <sub>4</sub> O	1818	1	0
	C <sub>3</sub> H <sub>4</sub> O <sub>4</sub>	401	12	1
	C <sub>3</sub> H <sub>8</sub> N <sub>2</sub> O <sub>2</sub>	2852	58	5
	C <sub>3</sub> H <sub>12</sub> N <sub>4</sub>	143	1	0
	C <sub>4</sub> N <sub>4</sub>	1616	1	0
	C <sub>4</sub> H <sub>8</sub> O <sub>3</sub>	425	74	17
	C <sub>4</sub> H <sub>12</sub> N <sub>2</sub> O	333	31	1
	C <sub>5</sub> N <sub>2</sub> O	2693	0	0
	C <sub>5</sub> H <sub>12</sub> O <sub>2</sub>	69	58	25
	C <sub>6</sub> O <sub>2</sub>	459	0	0
	C <sub>6</sub> H <sub>4</sub> N <sub>2</sub>	49516	19	3
	C <sub>7</sub> H <sub>4</sub> O	11332	6	0
	C <sub>8</sub> H <sub>8</sub>	7437	102	4
105	CH <sub>3</sub> N <sub>3</sub> O <sub>3</sub>	1198	0	0
	CH <sub>7</sub> N <sub>5</sub> O	674	0	0
	C <sub>2</sub> H <sub>3</sub> NO <sub>4</sub>	720	3	0
	C <sub>2</sub> H <sub>7</sub> N <sub>3</sub> O <sub>2</sub>	2529	4	0
	C <sub>2</sub> H <sub>11</sub> N <sub>5</sub>	88	0	0
	C <sub>3</sub> H <sub>7</sub> NO <sub>3</sub>	1391	30	4
	C <sub>3</sub> H <sub>11</sub> N <sub>3</sub> O	419	0	0
	C <sub>4</sub> H <sub>11</sub> NO <sub>2</sub>	284	36	4
	C <sub>5</sub> H <sub>3</sub> N <sub>3</sub>	40910	9	1
	C <sub>6</sub> H <sub>3</sub> NO	31325	0	0
	C <sub>7</sub> H <sub>7</sub> N	34152	42	4

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
106	CH <sub>2</sub> N <sub>2</sub> O <sub>4</sub>	404	0	0
	CH <sub>6</sub> N <sub>4</sub> O <sub>2</sub>	1131	0	0
	CH <sub>10</sub> N <sub>6</sub>	32	0	0
	C <sub>2</sub> H <sub>2</sub> O <sub>5</sub>	74	2	0
	C <sub>2</sub> H <sub>6</sub> N <sub>2</sub> O <sub>3</sub>	1778	5	0
	C <sub>2</sub> H <sub>10</sub> N <sub>4</sub> O	315	0	0
	C <sub>3</sub> H <sub>6</sub> O <sub>4</sub>	263	14	1
	C <sub>3</sub> H <sub>10</sub> N <sub>2</sub> O <sub>2</sub>	525	6	0
	C <sub>4</sub> H <sub>2</sub> N <sub>4</sub>	18307	3	0
	C <sub>4</sub> H <sub>10</sub> O <sub>3</sub>	88	31	5
	C <sub>5</sub> H <sub>2</sub> N <sub>2</sub> O	32187	0	0
	C <sub>6</sub> H <sub>2</sub> O <sub>2</sub>	4636	2	0
	C <sub>6</sub> H <sub>6</sub> N <sub>2</sub>	69352	67	11
	C <sub>7</sub> H <sub>6</sub> O	15804	34	6
	C <sub>8</sub> H <sub>10</sub>	4679	249	20
107	CHNO <sub>5</sub>	58	0	0
	CH <sub>5</sub> N <sub>3</sub> O <sub>3</sub>	1003	0	0
	CH <sub>9</sub> N <sub>5</sub> O	128	0	0
	C <sub>2</sub> H <sub>5</sub> NO <sub>4</sub>	620	2	0
	C <sub>2</sub> H <sub>9</sub> N <sub>3</sub> O <sub>2</sub>	508	0	0
	C <sub>3</sub> HN <sub>5</sub>	3969	0	0
	C <sub>3</sub> H <sub>9</sub> NO <sub>3</sub>	302	4	0
	C <sub>4</sub> HN <sub>3</sub> O	14015	0	0
	C <sub>5</sub> HNO <sub>2</sub>	6776	0	0
	C <sub>5</sub> H <sub>5</sub> N <sub>3</sub>	76376	38	2
	C <sub>6</sub> H <sub>5</sub> NO	58218	32	4
	C <sub>7</sub> H <sub>9</sub> N	24314	134	18
	108	CO <sub>6</sub>	3	0
CH <sub>4</sub> N <sub>2</sub> O <sub>4</sub>		492	0	0
CH <sub>8</sub> N <sub>4</sub> O <sub>2</sub>		254	0	0
C <sub>2</sub> N <sub>6</sub>		389	0	0
C <sub>2</sub> H <sub>4</sub> O <sub>5</sub>		88	0	0
C <sub>2</sub> H <sub>8</sub> N <sub>2</sub> O <sub>3</sub>		428	0	0
C <sub>3</sub> N <sub>4</sub> O		2218	0	0
C <sub>3</sub> H <sub>8</sub> O <sub>4</sub>		71	10	0
C <sub>4</sub> N <sub>2</sub> O <sub>2</sub>		2628	1	0
C <sub>4</sub> H <sub>4</sub> N <sub>4</sub>		49423	37	4
C <sub>5</sub> O <sub>3</sub>		327	0	0
C <sub>5</sub> H <sub>4</sub> N <sub>2</sub> O		87055	33	0
C <sub>6</sub> H <sub>4</sub> O <sub>2</sub>		12098	16	1
C <sub>6</sub> H <sub>8</sub> N <sub>2</sub>		57411	137	30
C <sub>7</sub> H <sub>8</sub> O		13177	194	15
C <sub>8</sub> H <sub>12</sub>		2082	426	76
C <sub>9</sub>		832	0	0
109	CH <sub>3</sub> NO <sub>5</sub>	121	0	0
	CH <sub>7</sub> N <sub>3</sub> O <sub>3</sub>	275	0	0
	C <sub>2</sub> H <sub>7</sub> NO <sub>4</sub>	186	0	0
	C <sub>3</sub> H <sub>3</sub> N <sub>5</sub>	18307	14	0
	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> O	65056	15	0
	C <sub>5</sub> H <sub>3</sub> NO <sub>2</sub>	30807	8	0
	C <sub>5</sub> H <sub>7</sub> N <sub>3</sub>	76138	78	8
	C <sub>6</sub> H <sub>7</sub> NO	58265	151	21
C <sub>7</sub> H <sub>11</sub> N	11673	184	12	
110	CH <sub>2</sub> O <sub>6</sub>	12	0	0
	CH <sub>6</sub> N <sub>2</sub> O <sub>4</sub>	174	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>6</sub>	3761	2	0
	C <sub>2</sub> H <sub>6</sub> O <sub>5</sub>	35	1	0
	C <sub>3</sub> H <sub>2</sub> N <sub>4</sub> O	24928	4	0
	C <sub>4</sub> H <sub>2</sub> N <sub>2</sub> O <sub>2</sub>	28722	3	0
	C <sub>4</sub> H <sub>6</sub> N <sub>4</sub>	61793	49	5
	C <sub>5</sub> H <sub>2</sub> O <sub>3</sub>	3292	1	0
	C <sub>5</sub> H <sub>6</sub> N <sub>2</sub> O	109134	145	27

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	C <sub>6</sub> H <sub>6</sub> O <sub>2</sub>	15066	104	9
	C <sub>6</sub> H <sub>10</sub> N <sub>2</sub>	30600	218	17
	C <sub>7</sub> H <sub>10</sub> O	7166	590	50
	C <sub>8</sub> H <sub>14</sub>	654	303	84
	C <sub>9</sub> H <sub>2</sub>	10064	0	0
111	CHN <sub>7</sub>	343	0	0
	CH <sub>5</sub> NO <sub>5</sub>	58	0	0
	C <sub>2</sub> HN <sub>5</sub> O	4418	0	0
	C <sub>3</sub> HN <sub>3</sub> O <sub>2</sub>	10828	1	0
	C <sub>3</sub> H <sub>5</sub> N <sub>5</sub>	30527	11	0
	C <sub>4</sub> HNO <sub>3</sub>	4429	0	0
	C <sub>4</sub> H <sub>5</sub> N <sub>3</sub> O	108769	94	6
	C <sub>5</sub> H <sub>5</sub> NO <sub>2</sub>	51235	73	8
	C <sub>5</sub> H <sub>9</sub> N <sub>3</sub>	46125	100	5
	C <sub>6</sub> H <sub>9</sub> NO	35759	356	14
C <sub>7</sub> H <sub>13</sub> N	3809	292	30	
C <sub>8</sub> HN	17198	0	0	
112	CN <sub>6</sub> O	280	0	0
	CH <sub>4</sub> O <sub>6</sub>	9	0	0
	C <sub>2</sub> N <sub>4</sub> O <sub>2</sub>	1451	0	0
	C <sub>2</sub> H <sub>4</sub> N <sub>6</sub>	8982	3	0
	C <sub>3</sub> N <sub>2</sub> O <sub>3</sub>	1463	1	0
	C <sub>3</sub> H <sub>4</sub> N <sub>4</sub> O	60869	34	1
	C <sub>4</sub> O <sub>4</sub>	194	1	0
	C <sub>4</sub> H <sub>4</sub> N <sub>2</sub> O <sub>2</sub>	69482	74	7
	C <sub>4</sub> H <sub>8</sub> N <sub>4</sub>	43697	68	5
	C <sub>5</sub> H <sub>4</sub> O <sub>3</sub>	7744	36	7
	C <sub>5</sub> H <sub>8</sub> N <sub>2</sub> O	77843	204	14
	C <sub>6</sub> H <sub>8</sub> O <sub>2</sub>	10893	389	43
	C <sub>6</sub> H <sub>12</sub> N <sub>2</sub>	10706	254	26
	C <sub>7</sub> N <sub>2</sub>	8260	0	0
	C <sub>7</sub> H <sub>12</sub> O	2589	782	96
	C <sub>8</sub> O	1804	0	0
	C <sub>8</sub> H <sub>16</sub>	139	115	79
C <sub>9</sub> H <sub>4</sub>	33860	2	0	
113	CH <sub>3</sub> N <sub>7</sub>	1372	0	0
	C <sub>2</sub> H <sub>3</sub> N <sub>5</sub> O	18494	5	0
	C <sub>3</sub> H <sub>3</sub> N <sub>3</sub> O <sub>2</sub>	45304	27	1
	C <sub>3</sub> H <sub>7</sub> N <sub>5</sub>	26040	31	2
	C <sub>4</sub> H <sub>3</sub> NO <sub>3</sub>	18082	16	0
	C <sub>4</sub> H <sub>7</sub> N <sub>3</sub> O	93323	118	2
	C <sub>5</sub> H <sub>7</sub> NO <sub>2</sub>	44336	178	13
	C <sub>5</sub> H <sub>11</sub> N <sub>3</sub>	17608	42	0
C <sub>6</sub> H <sub>11</sub> NO	13982	398	26	
C <sub>7</sub> H <sub>15</sub> N	801	236	31	
C <sub>8</sub> H <sub>3</sub> N	102012	1	0	
114	CH <sub>2</sub> N <sub>6</sub> O	2711	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>4</sub> O <sub>2</sub>	14394	6	0
	C <sub>2</sub> H <sub>6</sub> N <sub>6</sub>	9588	4	0
	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub> O <sub>3</sub>	14739	3	1
	C <sub>3</sub> H <sub>6</sub> N <sub>4</sub> O	65627	42	2
	C <sub>4</sub> H <sub>2</sub> O <sub>4</sub>	1635	8	2
	C <sub>4</sub> H <sub>6</sub> N <sub>2</sub> O <sub>2</sub>	75211	101	9
	C <sub>4</sub> H <sub>10</sub> N <sub>4</sub>	18611	21	0
	C <sub>5</sub> H <sub>6</sub> O <sub>3</sub>	8397	129	10
	C <sub>5</sub> H <sub>10</sub> N <sub>2</sub> O	33689	200	6
	C <sub>6</sub> H <sub>10</sub> O <sub>2</sub>	4869	597	91
	C <sub>6</sub> H <sub>14</sub> N <sub>2</sub>	2338	197	23
	C <sub>7</sub> H <sub>2</sub> N <sub>2</sub> O	117942	0	0
	C <sub>7</sub> H <sub>14</sub> O	596	397	88
	C <sub>8</sub> H <sub>2</sub> O	24021	0	0
C <sub>8</sub> H <sub>18</sub>	18	18	18	

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	$C_9H_6$	56437	1	0
115	$CHN_5O_2$	1868	0	0
	$CH_5N_7$	1941	0	0
	$C_2HN_3O_3$	4628	0	0
	$C_2H_5N_5O$	26619	4	0
	$C_3HNO_4$	1909	0	0
	$C_3H_5N_3O_2$	65434	34	1
	$C_3H_9N_5$	12629	5	0
	$C_4H_5NO_3$	26063	42	2
	$C_4H_9N_3O$	45798	40	2
	$C_5H_9NO_2$	22259	308	12
	$C_5H_{13}N_3$	4054	44	2
	$C_6HN_3$	59406	0	0
	$C_6H_{13}NO$	3345	356	44
	$C_7HNO$	39727	0	0
	$C_7H_{17}N$	89	73	18
	$C_8H_5N$	229260	2	0
116	$CN_4O_3$	448	0	0
	$CH_4N_6O$	5678	0	0
	$C_2N_2O_4$	549	0	0
	$C_2H_4N_4O_2$	30346	11	0
	$C_2H_8N_6$	5431	4	0
	$C_3O_5$	68	0	0
	$C_3H_4N_2O_3$	31006	13	1
	$C_3H_8N_4O$	37506	15	0
	$C_4H_4O_4$	3328	19	3
	$C_4H_8N_2O_2$	43731	134	9
	$C_4H_{12}N_4$	4618	17	1
	$C_5N_4$	11556	1	0
	$C_5H_8O_3$	4986	232	16
	$C_5H_{12}N_2O$	8585	170	14
	$C_6N_2O$	17171	0	0
	$C_6H_{12}O_2$	1313	472	61
	$C_6H_{16}N_2$	260	105	22
	$C_7O_2$	2254	1	0
	$C_7H_4N_2$	388019	9	0
	$C_7H_{16}O$	72	67	41
	$C_8H_4O$	77431	6	0
	$C_9H_8$	57771	39	7
117	$CH_3N_5O_2$	6821	0	0
	$CH_7N_7$	1317	0	0
	$C_2H_3N_3O_3$	16849	2	0
	$C_2H_7N_5O$	18299	3	0
	$C_3H_3NO_4$	6789	3	0
	$C_3H_7N_3O_2$	45626	32	2
	$C_3H_{11}N_5$	3429	2	0
	$C_4H_7NO_3$	18469	67	6
	$C_4H_{11}N_3O$	12676	41	0
	$C_5H_{11}NO_2$	6418	233	16
	$C_5H_{15}N_3$	453	16	1
	$C_6H_3N_3$	338610	7	3
	$C_6H_{15}NO$	398	156	14
	$C_7H_3NO$	223890	2	0
	$C_8H_7N$	284065	54	10
118	$CH_2N_4O_3$	3978	0	0
	$CH_6N_6O$	4880	0	0
	$C_2H_2N_2O_4$	4667	1	0
	$C_2H_6N_4O_2$	26505	9	2
	$C_2H_{10}N_6$	1646	0	0
	$C_3H_2O_5$	547	2	0
	$C_3H_6N_2O_3$	27415	27	1
	$C_3H_{10}N_4O$	11507	5	0

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	$C_4H_6O_4$	2958	36	6
	$C_4H_{10}N_2O_2$	13864	117	3
	$C_4H_{14}N_4$	542	5	0
	$C_5H_2N_4$	159874	5	1
	$C_5H_{10}O_3$	1656	197	23
	$C_5H_{14}N_2O$	1041	37	1
	$C_6H_2N_2O$	240339	4	0
	$C_6H_{14}O_2$	179	129	37
	$C_7H_2O_2$	28770	0	0
	$C_7H_6N_2$	606589	74	13
	$C_8H_6O$	120427	24	3
	$C_9H_{10}$	40139	157	17
119	$CHN_3O_4$	1092	0	0
	$CH_5N_5O_2$	7934	0	0
	$CH_9N_7$	452	0	0
	$C_2HNO_5$	543	0	0
	$C_2H_5N_3O_3$	19834	6	0
	$C_2H_9N_5O$	6353	0	0
	$C_3H_5NO_4$	8026	9	0
	$C_3H_9N_3O_2$	16276	8	0
	$C_3H_{13}N_5$	422	0	0
	$C_4HN_5$	37895	1	0
	$C_4H_9NO_3$	6836	82	8
	$C_4H_{13}N_3O$	1605	1	0
	$C_5HN_3O$	110452	2	0
	$C_5H_{13}NO_2$	874	68	7
	$C_6HNO_2$	44446	0	0
	$C_6H_5N_3$	710961	74	17
	$C_7H_5NO$	467617	40	8
	$C_8H_9N$	225296	103	11
120	$CN_2O_5$	115	0	0
	$CH_4N_4O_3$	6850	0	0
	$CH_8N_6O$	1959	0	0
	$C_2O_6$	23	0	0
	$C_2H_4N_2O_4$	8018	7	0
	$C_2H_8N_4O_2$	10925	0	0
	$C_2H_{12}N_6$	223	0	0
	$C_3N_6$	3697	0	0
	$C_3H_4O_5$	915	3	1
	$C_3H_8N_2O_3$	11666	12	1
	$C_3H_{12}N_4O$	1549	0	0
	$C_4N_4O$	18638	2	0
	$C_4H_8O_4$	1310	48	2
	$C_4H_{12}N_2O_2$	1981	10	0
	$C_5N_2O_2$	17438	0	0
	$C_5H_4N_4$	493258	52	5
	$C_5H_{12}O_3$	258	73	10
	$C_6O_3$	1796	0	0
	$C_6H_4N_2O$	738283	43	1
	$C_7H_4O_2$	86246	8	0
	$C_7H_8N_2$	563966	131	14
	$C_8H_8O$	112484	144	14
	$C_9H_{12}$	19983	415	40
	$C_{10}$	4330	0	0
121	$CH_3N_3O_4$	3274	0	0
	$CH_7N_5O_2$	3892	0	0
	$CH_{11}N_7$	66	0	0
	$C_2H_3NO_5$	1585	0	0
	$C_2H_7N_3O_3$	10035	0	1
	$C_2H_{11}N_5O$	926	0	0
	$C_3H_7NO_4$	4197	2	0
	$C_3H_{11}N_3O_2$	2479	0	0

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	C <sub>4</sub> H <sub>3</sub> N <sub>5</sub>	202151	13	3
	C <sub>4</sub> H <sub>11</sub> NO <sub>3</sub>	1116	10	1
	C <sub>5</sub> H <sub>3</sub> N <sub>3</sub> O	590494	16	0
	C <sub>6</sub> H <sub>3</sub> NO <sub>2</sub>	233143	6	1
	C <sub>6</sub> H <sub>7</sub> N <sub>3</sub>	801769	76	11
	C <sub>7</sub> H <sub>7</sub> NO	528227	124	12
	C <sub>8</sub> H <sub>11</sub> N	122819	218	36
122	CH <sub>2</sub> N <sub>2</sub> O <sub>5</sub>	823	0	0
	CH <sub>6</sub> N <sub>4</sub> O <sub>3</sub>	4295	0	0
	CH <sub>10</sub> N <sub>6</sub> O	313	0	0
	C <sub>2</sub> H <sub>2</sub> O <sub>6</sub>	127	2	0
	C <sub>2</sub> H <sub>6</sub> N <sub>2</sub> O <sub>4</sub>	5187	0	0
	C <sub>2</sub> H <sub>10</sub> N <sub>4</sub> O <sub>2</sub>	1824	0	0
	C <sub>3</sub> H <sub>2</sub> N <sub>6</sub>	46786	2	0
	C <sub>3</sub> H <sub>6</sub> O <sub>5</sub>	608	1	0
	C <sub>3</sub> H <sub>10</sub> N <sub>2</sub> O <sub>3</sub>	2051	1	0
	C <sub>4</sub> H <sub>2</sub> N <sub>4</sub> O	247932	0	0
	C <sub>4</sub> H <sub>10</sub> O <sub>4</sub>	255	21	1
	C <sub>5</sub> H <sub>2</sub> N <sub>2</sub> O <sub>2</sub>	229717	4	0
	C <sub>5</sub> H <sub>6</sub> N <sub>4</sub>	704153	52	1
	C <sub>6</sub> H <sub>2</sub> O <sub>3</sub>	21641	1	0
	C <sub>6</sub> H <sub>6</sub> N <sub>2</sub> O	1053290	100	10
	C <sub>7</sub> H <sub>6</sub> O <sub>2</sub>	122391	69	7
	C <sub>7</sub> H <sub>10</sub> N <sub>2</sub>	344434	273	31
	C <sub>8</sub> H <sub>10</sub> O	69669	553	40
	C <sub>9</sub> H <sub>14</sub>	7244	579	47
	C <sub>10</sub> H <sub>2</sub>	64352	1	0
123	CHNO <sub>6</sub>	92	0	0
	CH <sub>5</sub> N <sub>3</sub> O <sub>4</sub>	2791	0	0
	CH <sub>9</sub> N <sub>5</sub> O <sub>2</sub>	718	0	0
	C <sub>2</sub> HN <sub>7</sub>	5370	0	0
	C <sub>2</sub> H <sub>5</sub> NO <sub>5</sub>	1385	0	0
	C <sub>2</sub> H <sub>9</sub> N <sub>3</sub> O <sub>3</sub>	1951	0	0
	C <sub>3</sub> HN <sub>5</sub> O	50129	0	0
	C <sub>3</sub> H <sub>9</sub> NO <sub>4</sub>	877	0	0
	C <sub>4</sub> HN <sub>3</sub> O <sub>2</sub>	94422	0	0
	C <sub>4</sub> H <sub>5</sub> N <sub>5</sub>	388316	26	0
	C <sub>5</sub> HNO <sub>3</sub>	30775	0	0
	C <sub>5</sub> H <sub>5</sub> N <sub>3</sub> O	1133182	52	2
	C <sub>6</sub> H <sub>5</sub> NO <sub>2</sub>	444584	67	5
	C <sub>6</sub> H <sub>9</sub> N <sub>3</sub>	561140	182	8
	C <sub>7</sub> H <sub>9</sub> NO	372937	397	35
	C <sub>8</sub> H <sub>13</sub> N	47323	335	16
	C <sub>9</sub> HN	113702	1	0
124	CN <sub>8</sub>	251	0	0
	CO <sub>7</sub>	3	0	0
	CH <sub>4</sub> N <sub>2</sub> O <sub>5</sub>	1040	0	0
	CH <sub>8</sub> N <sub>4</sub> O <sub>3</sub>	942	0	0
	C <sub>2</sub> N <sub>6</sub> O	3866	0	0
	C <sub>2</sub> H <sub>4</sub> O <sub>6</sub>	157	0	0
	C <sub>2</sub> H <sub>8</sub> N <sub>2</sub> O <sub>4</sub>	1222	0	0
	C <sub>3</sub> N <sub>4</sub> O <sub>2</sub>	13654	0	0
	C <sub>3</sub> H <sub>4</sub> N <sub>6</sub>	131957	12	0
	C <sub>3</sub> H <sub>8</sub> O <sub>5</sub>	154	2	0
	C <sub>4</sub> N <sub>2</sub> O <sub>3</sub>	10835	0	0
	C <sub>4</sub> H <sub>4</sub> N <sub>4</sub> O	702522	13	1
	C <sub>5</sub> O <sub>4</sub>	1015	0	0
	C <sub>5</sub> H <sub>4</sub> N <sub>2</sub> O <sub>2</sub>	645384	26	1
	C <sub>5</sub> H <sub>8</sub> N <sub>4</sub>	579834	92	4
	C <sub>6</sub> H <sub>4</sub> O <sub>3</sub>	59327	27	1
	C <sub>6</sub> H <sub>8</sub> N <sub>2</sub> O	871629	334	18
	C <sub>7</sub> H <sub>8</sub> O <sub>2</sub>	102139	318	23

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	C <sub>7</sub> H <sub>12</sub> N <sub>2</sub>	143857	295	14
	C <sub>8</sub> N <sub>2</sub>	54288	1	0
	C <sub>8</sub> H <sub>12</sub> O	29797	1257	78
	C <sub>9</sub> O	10064	0	0
	C <sub>9</sub> H <sub>16</sub>	1902	431	90
	C <sub>10</sub> H <sub>4</sub>	241297	2	0
125	CH <sub>3</sub> NO <sub>6</sub>	201	0	0
	CH <sub>7</sub> N <sub>3</sub> O <sub>4</sub>	759	0	0
	C <sub>2</sub> H <sub>3</sub> N <sub>7</sub>	25873	0	0
	C <sub>2</sub> H <sub>7</sub> NO <sub>5</sub>	405	0	0
	C <sub>3</sub> H <sub>3</sub> N <sub>5</sub> O	247655	0	0
	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> O <sub>2</sub>	463377	6	0
	C <sub>4</sub> H <sub>7</sub> N <sub>5</sub>	388792	35	3
	C <sub>5</sub> H <sub>3</sub> NO <sub>3</sub>	148046	6	0
	C <sub>5</sub> H <sub>7</sub> N <sub>3</sub> O	1137301	208	6
	C <sub>6</sub> H <sub>7</sub> NO <sub>2</sub>	448029	280	10
	C <sub>6</sub> H <sub>11</sub> N <sub>3</sub>	258612	126	10
	C <sub>7</sub> H <sub>11</sub> NO	174763	573	18
	C <sub>8</sub> H <sub>15</sub> N	12770	499	26
	C <sub>9</sub> H <sub>3</sub> N	753600	0	0
126	CH <sub>2</sub> N <sub>8</sub>	2596	0	0
	CH <sub>2</sub> O <sub>7</sub>	16	0	0
	CH <sub>6</sub> N <sub>2</sub> O <sub>5</sub>	364	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>6</sub> O	46231	0	0
	C <sub>2</sub> H <sub>6</sub> O <sub>6</sub>	61	1	0
	C <sub>3</sub> H <sub>2</sub> N <sub>4</sub> O <sub>2</sub>	168157	0	0
	C <sub>3</sub> H <sub>6</sub> N <sub>6</sub>	167486	15	1
	C <sub>4</sub> H <sub>2</sub> N <sub>2</sub> O <sub>3</sub>	131318	3	0
	C <sub>4</sub> H <sub>6</sub> N <sub>4</sub> O	893672	121	8
	C <sub>5</sub> H <sub>2</sub> O <sub>4</sub>	11291	0	0
	C <sub>5</sub> H <sub>6</sub> N <sub>2</sub> O <sub>2</sub>	821421	199	16
	C <sub>5</sub> H <sub>10</sub> N <sub>4</sub>	300547	83	3
	C <sub>6</sub> H <sub>6</sub> O <sub>3</sub>	75331	140	18
	C <sub>6</sub> H <sub>10</sub> N <sub>2</sub> O	456982	331	10
	C <sub>7</sub> H <sub>10</sub> O <sub>2</sub>	54641	909	51
	C <sub>7</sub> H <sub>14</sub> N <sub>2</sub>	40953	296	20
	C <sub>8</sub> H <sub>2</sub> N <sub>2</sub>	896748	1	0
	C <sub>8</sub> H <sub>14</sub> O	8796	1347	113
	C <sub>9</sub> H <sub>2</sub> O	160114	0	0
	C <sub>9</sub> H <sub>18</sub>	338	165	62
	C <sub>10</sub> H <sub>6</sub>	439373	16	1
127	CHN <sub>7</sub> O	3781	0	0
	CH <sub>5</sub> NO <sub>6</sub>	97	0	0
	C <sub>2</sub> HN <sub>5</sub> O <sub>2</sub>	27707	0	0
	C <sub>2</sub> H <sub>5</sub> N <sub>7</sub>	44090	3	0
	C <sub>3</sub> HN <sub>3</sub> O <sub>3</sub>	46748	0	0
	C <sub>3</sub> H <sub>5</sub> N <sub>5</sub> O	424976	54	1
	C <sub>4</sub> HNO <sub>4</sub>	14493	0	0
	C <sub>4</sub> H <sub>5</sub> N <sub>3</sub> O <sub>2</sub>	793933	124	8
	C <sub>4</sub> H <sub>9</sub> N <sub>5</sub>	231598	34	2
	C <sub>5</sub> H <sub>5</sub> NO <sub>3</sub>	252373	82	2
	C <sub>5</sub> H <sub>9</sub> N <sub>3</sub> O	682547	158	3
	C <sub>6</sub> H <sub>9</sub> NO <sub>2</sub>	272736	448	16
	C <sub>6</sub> H <sub>13</sub> N <sub>3</sub>	78864	57	0
	C <sub>7</sub> HN <sub>3</sub>	465296	0	0
	C <sub>7</sub> H <sub>13</sub> NO	54700	706	45
	C <sub>8</sub> HNO	273106	0	0
	C <sub>8</sub> H <sub>17</sub> N	2258	343	39
	C <sub>9</sub> H <sub>5</sub> N	1863935	12	0
128	CN <sub>6</sub> O <sub>2</sub>	1554	0	0
	CH <sub>4</sub> N <sub>8</sub>	6392	0	0
	CH <sub>4</sub> O <sub>7</sub>	11	0	0

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	$C_2N_4O_3$	5575	0	0
	$C_2H_4N_6O$	117168	0	0
	$C_3N_2O_4$	4415	0	0
	$C_3H_4N_4O_2$	426683	35	1
	$C_3H_8N_6$	117563	3	0
	$C_4O_5$	429	0	0
	$C_4H_4N_2O_3$	330347	45	4
	$C_4H_8N_4O$	630529	39	1
	$C_5H_4O_4$	27721	25	0
	$C_5H_8N_2O_2$	585130	207	14
	$C_5H_{12}N_4$	99803	21	0
	$C_6N_4$	92041	1	1
	$C_6H_8O_3$	54343	379	23
	$C_6H_{12}N_2O$	154666	289	16
	$C_7N_2O$	118895	0	0
	$C_7H_{12}O_2$	19154	1138	95
	$C_7H_{16}N_2$	7436	256	21
	$C_8O_2$	13163	0	0
	$C_8H_4N_2$	3272676	6	3
	$C_8H_{16}O$	1684	637	125
	$C_9H_4O$	575884	1	0
	$C_9H_{20}$	35	35	35
	$C_{10}H_8$	488125	32	5
129	$CH_3N_7O$	16587	0	0
	$C_2H_3N_5O_2$	123010	1	0
	$C_2H_7N_7$	37610	2	0
	$C_3H_3N_3O_3$	206392	26	1
	$C_3H_7N_5O$	364469	5	0
	$C_4H_3NO_4$	62473	12	0
	$C_4H_7N_3O_2$	685212	72	1
	$C_4H_{11}N_5$	85111	6	2
	$C_5H_7NO_3$	219604	158	6
	$C_5H_{11}N_3O$	254221	58	2
	$C_6H_{11}NO_2$	104235	600	28
	$C_6H_{15}N_3$	14947	69	4
	$C_7H_3N_3$	2978179	7	3
	$C_7H_{15}NO$	10777	556	45
	$C_8H_3NO$	1729030	0	0
	$C_8H_{19}N$	211	117	26
	$C_9H_7N$	2521767	29	5
130	$CH_2N_6O_2$	16785	0	0
	$CH_6N_8$	6863	0	0
	$C_2H_2N_4O_3$	61544	1	0
	$C_2H_6N_6O$	127460	3	0
	$C_3H_2N_2O_4$	48402	5	0
	$C_3H_6N_4O_2$	466623	19	1
	$C_3H_{10}N_6$	48674	2	0
	$C_4H_2O_5$	4228	2	0
	$C_4H_6N_2O_3$	362688	49	2
	$C_4H_{10}N_4O$	263477	22	0
	$C_5H_6O_4$	30434	67	5
	$C_5H_{10}N_2O_2$	249379	198	9
	$C_5H_{14}N_4$	20046	17	1
	$C_6H_2N_4$	1475564	8	1
	$C_6H_{10}O_3$	23838	580	45
	$C_6H_{14}N_2O$	31984	239	9
	$C_7H_2N_2O$	1921208	0	0
	$C_7H_{14}O_2$	4177	726	65
	$C_7H_{18}N_2$	688	111	14
	$C_8H_2O_2$	197786	0	0
	$C_8H_6N_2$	5625815	51	8
	$C_8H_{18}O$	171	130	58

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	$C_9H_6O$	985744	17	2
	$C_{10}H_{10}$	369067	176	21
131	$CHN_5O_3$	7424	0	0
	$CH_5N_7O$	24294	0	0
	$C_2HN_3O_4$	14217	0	0
	$C_2H_5N_5O_2$	181597	0	0
	$C_2H_9N_7$	17891	0	0
	$C_3HNO_5$	4718	0	0
	$C_3H_5N_3O_3$	305195	10	1
	$C_3H_9N_5O$	174553	2	0
	$C_4H_5NO_4$	92108	20	0
	$C_4H_9N_3O_2$	333234	70	3
	$C_4H_{13}N_5$	18381	1	0
	$C_5HN_5$	371403	0	0
	$C_5H_9NO_3$	109126	188	14
	$C_5H_{13}N_3O$	56067	72	0
	$C_6HN_3O$	922258	0	0
	$C_6H_{13}NO_2$	23946	417	22
	$C_6H_{17}N_3$	1395	29	1
	$C_7HNO_2$	316272	0	0
	$C_7H_5N_3$	6927201	28	3
	$C_7H_{17}NO$	1068	216	10
	$C_8H_5NO$	3999703	11	4
	$C_9H_9N$	2190926	142	36
132	$CN_4O_4$	1257	0	0
	$CH_4N_6O_2$	36687	0	0
	$CH_8N_8$	3831	0	0
	$C_2N_2O_5$	1221	0	0
	$C_2H_4N_4O_3$	134993	1	0
	$C_2H_8N_6O$	71735	1	0
	$C_3O_6$	130	0	0
	$C_3H_4N_2O_4$	105410	3	1
	$C_3H_8N_4O_2$	266035	20	0
	$C_3H_{12}N_6$	11469	1	0
	$C_4N_6$	38349	1	0
	$C_4H_4O_5$	8952	7	0
	$C_4H_8N_2O_3$	210267	60	4
	$C_4H_{12}N_4O$	62968	4	0
	$C_5N_4O$	161998	0	0
	$C_5H_8O_4$	18023	122	9
	$C_5H_{12}N_2O_2$	61555	175	5
	$C_5H_{16}N_4$	1933	3	0
	$C_6N_2O_2$	127485	0	0
	$C_6H_4N_4$	5094755	25	2
	$C_6H_{12}O_3$	6171	424	39
	$C_6H_{16}N_2O$	3218	51	1
	$C_7O_3$	10604	0	0
	$C_7H_4N_2O$	6599812	13	1
	$C_7H_{16}O_2$	463	241	29
	$C_8H_4O_2$	666395	6	1
	$C_8H_8N_2$	5767073	229	13
	$C_9H_8O$	1013745	117	17
	$C_{10}H_{12}$	201578	379	44
	$C_{11}$	25227	0	0
133	$CH_3N_5O_3$	28677	0	0
	$CH_7N_7O$	16557	0	0
	$C_2H_3N_3O_4$	54494	0	0
	$C_2H_7N_5O_2$	125466	4	0
	$C_2H_{11}N_7$	4665	0	0
	$C_3H_3NO_5$	17677	1	0
	$C_3H_7N_3O_3$	213775	16	0
	$C_3H_{11}N_5O$	45932	2	0

$m$	$\beta$	$ M_\beta^C $	$BS$	$MS$
	C <sub>4</sub> H <sub>7</sub> NO <sub>4</sub>	65500	33	2
	C <sub>4</sub> H <sub>11</sub> N <sub>3</sub> O <sub>2</sub>	89999	10	0
	C <sub>4</sub> H <sub>15</sub> N <sub>5</sub>	1854	0	0
	C <sub>5</sub> H <sub>3</sub> N <sub>5</sub>	2249646	3	0
	C <sub>5</sub> H <sub>11</sub> NO <sub>3</sub>	30610	188	13
	C <sub>5</sub> H <sub>15</sub> N <sub>3</sub> O	5825	4	0
	C <sub>6</sub> H <sub>3</sub> N <sub>3</sub> O	5572831	6	0
	C <sub>6</sub> H <sub>15</sub> NO <sub>2</sub>	2659	89	6
	C <sub>7</sub> H <sub>3</sub> NO <sub>2</sub>	1882049	4	0
	C <sub>7</sub> H <sub>7</sub> N <sub>3</sub>	8666101	188	16
	C <sub>8</sub> H <sub>7</sub> NO	5005355	150	21
	C <sub>9</sub> H <sub>11</sub> N	1323028	252	19
134	CH <sub>2</sub> N <sub>4</sub> O <sub>4</sub>	12052	0	0
	CH <sub>6</sub> N <sub>6</sub> O <sub>2</sub>	32046	0	0
	CH <sub>10</sub> N <sub>8</sub>	1121	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub> O <sub>5</sub>	11406	0	0
	C <sub>2</sub> H <sub>6</sub> N <sub>4</sub> O <sub>3</sub>	119536	0	0
	C <sub>2</sub> H <sub>10</sub> N <sub>6</sub> O	21136	0	0
	C <sub>3</sub> H <sub>2</sub> O <sub>6</sub>	1115	0	0
	C <sub>3</sub> H <sub>6</sub> N <sub>2</sub> O <sub>4</sub>	94380	11	0
	C <sub>3</sub> H <sub>10</sub> N <sub>4</sub> O <sub>2</sub>	80120	0	0
	C <sub>3</sub> H <sub>14</sub> N <sub>6</sub>	1230	0	0
	C <sub>4</sub> H <sub>2</sub> N <sub>6</sub>	574379	2	0
	C <sub>4</sub> H <sub>6</sub> O <sub>5</sub>	8070	17	3
	C <sub>4</sub> H <sub>10</sub> N <sub>2</sub> O <sub>3</sub>	65347	34	1
	C <sub>4</sub> H <sub>14</sub> N <sub>4</sub> O	6877	0	0
	C <sub>5</sub> H <sub>2</sub> N <sub>4</sub> O	2504259	2	0
	C <sub>5</sub> H <sub>10</sub> O <sub>4</sub>	5841	120	6
	C <sub>5</sub> H <sub>14</sub> N <sub>2</sub> O <sub>2</sub>	7079	13	0
	C <sub>6</sub> H <sub>2</sub> N <sub>2</sub> O <sub>2</sub>	1942993	1	0
	C <sub>6</sub> H <sub>6</sub> N <sub>4</sub>	8123295	154	16
	C <sub>6</sub> H <sub>14</sub> O <sub>3</sub>	772	145	11
	C <sub>7</sub> H <sub>2</sub> O <sub>3</sub>	153809	0	0
	C <sub>7</sub> H <sub>6</sub> N <sub>2</sub> O	10504307	184	11
	C <sub>8</sub> H <sub>6</sub> O <sub>2</sub>	1055605	63	10
	C <sub>8</sub> H <sub>10</sub> N <sub>2</sub>	3928605	286	28
	C <sub>9</sub> H <sub>10</sub> O	697708	368	47
	C <sub>10</sub> H <sub>14</sub>	81909	639	68
	C <sub>11</sub> H <sub>2</sub>	455822	0	0
135	CHN <sub>3</sub> O <sub>5</sub>	2487	0	0
	CH <sub>5</sub> N <sub>5</sub> O <sub>3</sub>	34223	0	0
	CH <sub>9</sub> N <sub>7</sub> O	5557	0	0
	C <sub>2</sub> HNO <sub>6</sub>	1039	0	0
	C <sub>2</sub> H <sub>5</sub> N <sub>3</sub> O <sub>4</sub>	65614	1	0
	C <sub>2</sub> H <sub>9</sub> N <sub>5</sub> O <sub>2</sub>	43012	0	0
	C <sub>2</sub> H <sub>13</sub> N <sub>7</sub>	540	0	0
	C <sub>3</sub> HN <sub>7</sub>	75194	0	0
	C <sub>3</sub> H <sub>5</sub> NO <sub>5</sub>	21341	1	0
	C <sub>3</sub> H <sub>9</sub> N <sub>3</sub> O <sub>3</sub>	75329	2	0
	C <sub>3</sub> H <sub>13</sub> N <sub>5</sub> O	5343	0	0
	C <sub>4</sub> HN <sub>5</sub> O	556979	1	0
	C <sub>4</sub> H <sub>9</sub> NO <sub>4</sub>	23900	10	0
	C <sub>4</sub> H <sub>13</sub> N <sub>3</sub> O <sub>2</sub>	10903	0	0
	C <sub>5</sub> HN <sub>3</sub> O <sub>2</sub>	848498	0	0
	C <sub>5</sub> H <sub>5</sub> N <sub>5</sub>	4864651	93	10
	C <sub>5</sub> H <sub>13</sub> NO <sub>3</sub>	3949	13	0
	C <sub>6</sub> HNO <sub>3</sub>	229417	0	0
	C <sub>6</sub> H <sub>5</sub> N <sub>3</sub> O	12015117	105	6
	C <sub>7</sub> H <sub>5</sub> NO <sub>2</sub>	4032639	68	6
	C <sub>7</sub> H <sub>9</sub> N <sub>3</sub>	6807596	126	1
	C <sub>8</sub> H <sub>9</sub> NO	3955938	405	36
	C <sub>9</sub> H <sub>13</sub> N	577485	374	48

$m$	$\beta$	$ M_\beta^C $	$BS$	$MS$
	C <sub>10</sub> HN	828373	0	0
136	CN <sub>2</sub> O <sub>6</sub>	199	0	0
	CH <sub>4</sub> N <sub>4</sub> O <sub>4</sub>	21546	1	0
	CH <sub>8</sub> N <sub>6</sub> O <sub>2</sub>	12792	0	0
	CH <sub>12</sub> N <sub>8</sub>	142	0	0
	C <sub>2</sub> N <sub>8</sub>	4124	0	0
	C <sub>2</sub> O <sub>7</sub>	31	0	0
	C <sub>2</sub> H <sub>4</sub> N <sub>2</sub> O <sub>5</sub>	20331	0	0
	C <sub>2</sub> H <sub>8</sub> N <sub>4</sub> O <sub>3</sub>	49011	0	0
	C <sub>2</sub> H <sub>12</sub> N <sub>6</sub> O	2659	0	0
	C <sub>3</sub> N <sub>6</sub> O	47721	0	0
	C <sub>3</sub> H <sub>4</sub> O <sub>6</sub>	1945	2	0
	C <sub>3</sub> H <sub>8</sub> N <sub>2</sub> O <sub>4</sub>	39896	1	1
	C <sub>3</sub> H <sub>12</sub> N <sub>4</sub> O <sub>2</sub>	10407	0	0
	C <sub>4</sub> N <sub>4</sub> O <sub>2</sub>	132427	1	0
	C <sub>4</sub> H <sub>4</sub> N <sub>6</sub>	1845453	34	0
	C <sub>4</sub> H <sub>8</sub> O <sub>5</sub>	3528	8	0
	C <sub>4</sub> H <sub>12</sub> N <sub>2</sub> O <sub>3</sub>	8937	2	0
	C <sub>5</sub> N <sub>2</sub> O <sub>3</sub>	83751	0	0
	C <sub>5</sub> H <sub>4</sub> N <sub>4</sub> O	8047925	103	6
	C <sub>5</sub> H <sub>12</sub> O <sub>4</sub>	869	35	3
	C <sub>6</sub> O <sub>4</sub>	6361	1	0
	C <sub>6</sub> H <sub>4</sub> N <sub>2</sub> O <sub>2</sub>	6190115	53	2
	C <sub>6</sub> H <sub>8</sub> N <sub>4</sub>	7553343	101	3
	C <sub>7</sub> H <sub>4</sub> O <sub>3</sub>	481262	8	1
	C <sub>7</sub> H <sub>8</sub> N <sub>2</sub> O	9795506	288	26
	C <sub>8</sub> H <sub>8</sub> O <sub>2</sub>	989647	329	37
	C <sub>8</sub> H <sub>12</sub> N <sub>2</sub>	1874516	455	41
	C <sub>9</sub> N <sub>2</sub>	391470	0	0
	C <sub>9</sub> H <sub>12</sub> O	338761	1020	97
	C <sub>10</sub> O	64352	0	0
	C <sub>10</sub> H <sub>16</sub>	24938	932	135
	C <sub>11</sub> H <sub>4</sub>	1885531	1	0
137	CH <sub>3</sub> N <sub>3</sub> O <sub>5</sub>	7849	0	0
	CH <sub>7</sub> N <sub>5</sub> O <sub>3</sub>	16846	0	0
	CH <sub>11</sub> N <sub>7</sub> O	763	0	0
	C <sub>2</sub> H <sub>3</sub> NO <sub>6</sub>	3198	0	0
	C <sub>2</sub> H <sub>7</sub> N <sub>3</sub> O <sub>4</sub>	33266	0	0
	C <sub>2</sub> H <sub>11</sub> N <sub>5</sub> O <sub>2</sub>	6080	0	0
	C <sub>3</sub> H <sub>3</sub> N <sub>7</sub>	422513	2	0
	C <sub>3</sub> H <sub>7</sub> NO <sub>5</sub>	11146	0	0
	C <sub>3</sub> H <sub>11</sub> N <sub>3</sub> O <sub>3</sub>	11117	0	0
	C <sub>4</sub> H <sub>3</sub> N <sub>5</sub> O	3155630	37	3
	C <sub>4</sub> H <sub>11</sub> NO <sub>4</sub>	3760	2	0
	C <sub>5</sub> H <sub>3</sub> N <sub>3</sub> O <sub>2</sub>	4773841	16	0
	C <sub>5</sub> H <sub>7</sub> N <sub>5</sub>	5534563	39	1
	C <sub>6</sub> H <sub>3</sub> NO <sub>3</sub>	1268434	6	0
	C <sub>6</sub> H <sub>7</sub> N <sub>3</sub> O	13675413	144	5
	C <sub>7</sub> H <sub>7</sub> NO <sub>2</sub>	4598367	250	23
	C <sub>7</sub> H <sub>11</sub> N <sub>3</sub>	3609741	222	1
	C <sub>8</sub> H <sub>11</sub> NO	2123287	734	44
	C <sub>9</sub> H <sub>15</sub> N	184124	435	19
	C <sub>10</sub> H <sub>3</sub> N	6090422	1	0
138	CH <sub>2</sub> N <sub>2</sub> O <sub>6</sub>	1553	0	0
	CH <sub>6</sub> N <sub>4</sub> O <sub>4</sub>	13680	0	0
	CH <sub>10</sub> N <sub>6</sub> O <sub>2</sub>	1995	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>8</sub>	53906	5	0
	C <sub>2</sub> H <sub>2</sub> O <sub>7</sub>	202	0	0
	C <sub>2</sub> H <sub>6</sub> N <sub>2</sub> O <sub>5</sub>	13267	0	0
	C <sub>2</sub> H <sub>10</sub> N <sub>4</sub> O <sub>3</sub>	7961	0	0
	C <sub>3</sub> H <sub>2</sub> N <sub>6</sub> O	685244	2	0
	C <sub>3</sub> H <sub>6</sub> O <sub>6</sub>	1294	0	0

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	C <sub>3</sub> H <sub>10</sub> N <sub>2</sub> O <sub>4</sub>	6847	0	0
	C <sub>4</sub> H <sub>2</sub> N <sub>4</sub> O <sub>2</sub>	1910769	16	1
	C <sub>4</sub> H <sub>6</sub> N <sub>6</sub>	2681554	10	0
	C <sub>4</sub> H <sub>10</sub> O <sub>5</sub>	655	10	0
	C <sub>5</sub> H <sub>2</sub> N <sub>2</sub> O <sub>3</sub>	1197634	5	0
	C <sub>5</sub> H <sub>6</sub> N <sub>4</sub> O	11689522	85	0
	C <sub>6</sub> H <sub>2</sub> O <sub>4</sub>	84414	2	0
	C <sub>6</sub> H <sub>6</sub> N <sub>2</sub> O <sub>2</sub>	8978366	143	14
	C <sub>6</sub> H <sub>10</sub> N <sub>4</sub>	4533685	129	4
	C <sub>7</sub> H <sub>6</sub> O <sub>3</sub>	696019	103	13
	C <sub>7</sub> H <sub>10</sub> N <sub>2</sub> O	5926666	487	25
	C <sub>8</sub> H <sub>10</sub> O <sub>2</sub>	607376	876	81
	C <sub>8</sub> H <sub>14</sub> N <sub>2</sub>	637380	399	13
	C <sub>9</sub> H <sub>2</sub> N <sub>2</sub>	7380691	0	0
	C <sub>9</sub> H <sub>14</sub> O	118215	2032	96
	C <sub>10</sub> H <sub>2</sub> O	1175685	0	0
	C <sub>10</sub> H <sub>18</sub>	5568	633	113
	C <sub>11</sub> H <sub>6</sub>	3717018	1	0
139	CHNO <sub>7</sub>	137	0	0
	CHN <sub>9</sub>	3140	0	0
	CH <sub>5</sub> N <sub>3</sub> O <sub>5</sub>	6836	0	0
	CH <sub>9</sub> N <sub>5</sub> O <sub>3</sub>	3058	0	0
	C <sub>2</sub> HN <sub>7</sub> O	71821	0	0
	C <sub>2</sub> H <sub>5</sub> NO <sub>6</sub>	2839	0	0
	C <sub>2</sub> H <sub>9</sub> N <sub>3</sub> O <sub>4</sub>	6367	0	0
	C <sub>3</sub> HN <sub>5</sub> O <sub>2</sub>	363438	0	0
	C <sub>3</sub> H <sub>5</sub> N <sub>7</sub>	833476	1	0
	C <sub>3</sub> H <sub>9</sub> NO <sub>5</sub>	2270	0	0
	C <sub>4</sub> HN <sub>3</sub> O <sub>3</sub>	466716	0	0
	C <sub>4</sub> H <sub>5</sub> N <sub>5</sub> O	6228822	14	0
	C <sub>5</sub> HNO <sub>4</sub>	114952	0	0
	C <sub>5</sub> H <sub>5</sub> N <sub>3</sub> O <sub>2</sub>	9390618	58	0
	C <sub>5</sub> H <sub>9</sub> N <sub>5</sub>	3841244	51	0
	C <sub>6</sub> H <sub>5</sub> NO <sub>3</sub>	2480437	50	5
	C <sub>6</sub> H <sub>9</sub> N <sub>3</sub> O	9536191	306	6
	C <sub>7</sub> H <sub>9</sub> NO <sub>2</sub>	3237132	551	17
	C <sub>7</sub> H <sub>13</sub> N <sub>3</sub>	1327095	176	9
	C <sub>8</sub> HN <sub>3</sub>	3928846	0	0
	C <sub>8</sub> H <sub>13</sub> NO	795607	922	25
	C <sub>9</sub> HNO	2047874	0	0
	C <sub>9</sub> H <sub>17</sub> N	41989	623	34
	C <sub>10</sub> H <sub>5</sub> N	16335064	0	0
140	CN <sub>8</sub> O	2681	0	0
	CO <sub>8</sub>	4	0	0
	CH <sub>4</sub> N <sub>2</sub> O <sub>6</sub>	2021	0	0
	CH <sub>8</sub> N <sub>4</sub> O <sub>4</sub>	2990	0	0
	C <sub>2</sub> N <sub>6</sub> O <sub>2</sub>	25376	0	0
	C <sub>2</sub> H <sub>4</sub> N <sub>8</sub>	156863	0	0
	C <sub>2</sub> H <sub>4</sub> O <sub>7</sub>	256	0	0
	C <sub>2</sub> H <sub>8</sub> N <sub>2</sub> O <sub>5</sub>	3077	0	0
	C <sub>3</sub> N <sub>4</sub> O <sub>3</sub>	62223	0	0
	C <sub>3</sub> H <sub>4</sub> N <sub>6</sub> O	2021309	0	0
	C <sub>3</sub> H <sub>8</sub> O <sub>6</sub>	324	0	0
	C <sub>4</sub> N <sub>2</sub> O <sub>4</sub>	37712	0	0
	C <sub>4</sub> H <sub>4</sub> N <sub>4</sub> O <sub>2</sub>	5615022	7	0
	C <sub>4</sub> H <sub>8</sub> N <sub>6</sub>	2207858	27	1
	C <sub>5</sub> O <sub>5</sub>	2711	1	0
	C <sub>5</sub> H <sub>4</sub> N <sub>2</sub> O <sub>3</sub>	3489419	44	2
	C <sub>5</sub> H <sub>8</sub> N <sub>4</sub> O	9647405	192	3
	C <sub>6</sub> H <sub>4</sub> O <sub>4</sub>	240785	24	3
	C <sub>6</sub> H <sub>8</sub> N <sub>2</sub> O <sub>2</sub>	7451672	405	14
	C <sub>6</sub> H <sub>12</sub> N <sub>4</sub>	1828942	78	2
	C <sub>7</sub> N <sub>4</sub>	785412	0	0
	C <sub>7</sub> H <sub>8</sub> O <sub>3</sub>	582423	336	31
	C <sub>7</sub> H <sub>12</sub> N <sub>2</sub> O	2424077	469	7
	C <sub>8</sub> N <sub>2</sub> O	901869	0	0
	C <sub>8</sub> H <sub>12</sub> O <sub>2</sub>	254468	1881	106
	C <sub>8</sub> H <sub>16</sub> N <sub>2</sub>	151696	398	22
	C <sub>9</sub> O <sub>2</sub>	84548	0	0
	C <sub>9</sub> H <sub>4</sub> N <sub>2</sub>	29566078	0	0
	C <sub>9</sub> H <sub>16</sub> O	29172	1885	107
	C <sub>10</sub> H <sub>4</sub> O	4654419	0	0
	C <sub>10</sub> H <sub>20</sub>	852	252	89
	C <sub>11</sub> H <sub>8</sub>	4442438	29	2
141	CH <sub>3</sub> NO <sub>7</sub>	314	0	0
	CH <sub>3</sub> N <sub>9</sub>	15658	0	0
	CH <sub>7</sub> N <sub>3</sub> O <sub>5</sub>	1859	0	0
	C <sub>2</sub> H <sub>3</sub> N <sub>7</sub> O	372568	0	0
	C <sub>2</sub> H <sub>7</sub> NO <sub>6</sub>	824	0	0
	C <sub>3</sub> H <sub>3</sub> N <sub>5</sub> O <sub>2</sub>	1892347	0	0
	C <sub>3</sub> H <sub>7</sub> N <sub>7</sub>	840842	9	0
	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> O <sub>3</sub>	2408635	0	0
	C <sub>4</sub> H <sub>7</sub> N <sub>5</sub> O	6291833	99	1
	C <sub>5</sub> H <sub>3</sub> NO <sub>4</sub>	581752	0	0
	C <sub>5</sub> H <sub>7</sub> N <sub>3</sub> O <sub>2</sub>	9510665	240	11
	C <sub>5</sub> H <sub>11</sub> N <sub>5</sub>	1727027	48	0
	C <sub>6</sub> H <sub>7</sub> NO <sub>3</sub>	2522498	211	6
	C <sub>6</sub> H <sub>11</sub> N <sub>3</sub> O	4328819	191	4
	C <sub>7</sub> H <sub>11</sub> NO <sub>2</sub>	1495599	784	26
	C <sub>7</sub> H <sub>15</sub> N <sub>3</sub>	333757	63	1
	C <sub>8</sub> H <sub>3</sub> N <sub>3</sub>	27869664	3	0
	C <sub>8</sub> H <sub>15</sub> NO	205672	1039	39
	C <sub>9</sub> H <sub>3</sub> NO	14390891	0	0
	C <sub>9</sub> H <sub>19</sub> N	6355	440	29
	C <sub>10</sub> H <sub>7</sub> N	23895548	25	0
142	CH <sub>2</sub> N <sub>8</sub> O	33796	0	0
	CH <sub>2</sub> O <sub>8</sub>	20	0	0
	CH <sub>6</sub> N <sub>2</sub> O <sub>6</sub>	717	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>6</sub> O <sub>2</sub>	332923	1	0
	C <sub>2</sub> H <sub>6</sub> N <sub>8</sub>	201872	1	0
	C <sub>2</sub> H <sub>6</sub> O <sub>7</sub>	98	0	0
	C <sub>3</sub> H <sub>2</sub> N <sub>4</sub> O <sub>3</sub>	830461	0	0
	C <sub>3</sub> H <sub>6</sub> N <sub>6</sub> O	2611977	19	0
	C <sub>4</sub> H <sub>2</sub> N <sub>2</sub> O <sub>4</sub>	493119	6	1
	C <sub>4</sub> H <sub>6</sub> N <sub>4</sub> O <sub>2</sub>	7260203	99	3
	C <sub>4</sub> H <sub>10</sub> N <sub>6</sub>	1126112	6	0
	C <sub>5</sub> H <sub>2</sub> O <sub>5</sub>	33662	2	1
	C <sub>5</sub> H <sub>6</sub> N <sub>2</sub> O <sub>3</sub>	4513867	133	4
	C <sub>5</sub> H <sub>10</sub> N <sub>4</sub> O	4951073	57	0
	C <sub>6</sub> H <sub>6</sub> O <sub>4</sub>	310776	134	13
	C <sub>6</sub> H <sub>10</sub> N <sub>2</sub> O <sub>2</sub>	3874178	378	14
	C <sub>6</sub> H <sub>14</sub> N <sub>4</sub>	492658	31	2
	C <sub>7</sub> H <sub>2</sub> N <sub>4</sub>	14352119	2	1
	C <sub>7</sub> H <sub>10</sub> O <sub>3</sub>	308660	854	27
	C <sub>7</sub> H <sub>14</sub> N <sub>2</sub> O	666580	400	15
	C <sub>8</sub> H <sub>2</sub> N <sub>2</sub> O	16462667	0	0
	C <sub>8</sub> H <sub>14</sub> O <sub>2</sub>	72534	1915	128
	C <sub>8</sub> H <sub>18</sub> N <sub>2</sub>	23437	308	14
	C <sub>9</sub> H <sub>2</sub> O <sub>2</sub>	1481754	0	0
	C <sub>9</sub> H <sub>6</sub> N <sub>2</sub>	55296968	50	6
	C <sub>9</sub> H <sub>18</sub> O	4745	800	78
	C <sub>10</sub> H <sub>6</sub> O	8671508	12	0
	C <sub>10</sub> H <sub>22</sub>	75	75	37
	C <sub>11</sub> H <sub>10</sub>	3614427	110	6



$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
143	CHN <sub>7</sub> O <sub>2</sub>	25160	0	0
	CH <sub>5</sub> NO <sub>7</sub>	153	0	0
	CH <sub>5</sub> N <sub>9</sub>	27189	0	0
	C <sub>2</sub> HN <sub>5</sub> O <sub>3</sub>	128212	0	0
	C <sub>2</sub> H <sub>5</sub> N <sub>7</sub> O	655227	1	0
	C <sub>3</sub> HN <sub>3</sub> O <sub>4</sub>	165781	0	0
	C <sub>3</sub> H <sub>5</sub> N <sub>5</sub> O <sub>2</sub>	3330976	14	2
	C <sub>3</sub> H <sub>9</sub> N <sub>7</sub>	496568	2	0
	C <sub>4</sub> HNO <sub>5</sub>	41232	0	0
	C <sub>4</sub> H <sub>5</sub> N <sub>3</sub> O <sub>3</sub>	4229478	48	0
	C <sub>4</sub> H <sub>9</sub> N <sub>5</sub> O	3729876	4	0
	C <sub>5</sub> H <sub>5</sub> NO <sub>4</sub>	1016168	38	1
	C <sub>5</sub> H <sub>9</sub> N <sub>3</sub> O <sub>2</sub>	5690451	124	1
	C <sub>5</sub> H <sub>13</sub> N <sub>5</sub>	504522	7	0
	C <sub>6</sub> HN <sub>5</sub>	3795824	3	0
	C <sub>6</sub> H <sub>9</sub> NO <sub>3</sub>	1530269	360	15
	C <sub>6</sub> H <sub>13</sub> N <sub>3</sub> O	1284174	88	4
	C <sub>7</sub> HN <sub>3</sub> O	8175534	0	0
	C <sub>7</sub> H <sub>13</sub> NO <sub>2</sub>	455946	872	38
	C <sub>7</sub> H <sub>17</sub> N <sub>3</sub>	53310	80	1
C <sub>8</sub> HNO <sub>2</sub>	2439749	0	0	
C <sub>8</sub> H <sub>5</sub> N <sub>3</sub>	70969521	50	4	
C <sub>8</sub> H <sub>17</sub> NO	34156	706	25	
C <sub>9</sub> H <sub>5</sub> NO	36456956	17	0	
C <sub>9</sub> H <sub>21</sub> N	507	142	11	
C <sub>10</sub> H <sub>9</sub> N	22467086	126	20	
144	CN <sub>6</sub> O <sub>3</sub>	6470	0	0
	CH <sub>4</sub> N <sub>8</sub> O	88451	0	0
	CH <sub>4</sub> O <sub>8</sub>	15	0	0
	C <sub>2</sub> N <sub>4</sub> O <sub>4</sub>	18233	0	0
	C <sub>2</sub> H <sub>4</sub> N <sub>6</sub> O <sub>2</sub>	877705	2	0
	C <sub>2</sub> H <sub>8</sub> N <sub>8</sub>	141375	2	0
	C <sub>3</sub> N <sub>2</sub> O <sub>5</sub>	11649	0	0
	C <sub>3</sub> H <sub>4</sub> N <sub>4</sub> O <sub>3</sub>	2185015	6	0
	C <sub>3</sub> H <sub>8</sub> N <sub>6</sub> O	1835635	2	0
	C <sub>4</sub> O <sub>6</sub>	932	1	0
	C <sub>4</sub> H <sub>4</sub> N <sub>2</sub> O <sub>4</sub>	1284622	15	3
	C <sub>4</sub> H <sub>8</sub> N <sub>4</sub> O <sub>2</sub>	5135621	43	2
	C <sub>4</sub> H <sub>12</sub> N <sub>6</sub>	361836	6	0
	C <sub>5</sub> N <sub>6</sub>	407003	0	0
	C <sub>5</sub> H <sub>4</sub> O <sub>5</sub>	85857	12	0
	C <sub>5</sub> H <sub>8</sub> N <sub>2</sub> O <sub>3</sub>	3223855	109	7
	C <sub>5</sub> H <sub>12</sub> N <sub>4</sub> O	1607685	21	0
	C <sub>6</sub> N <sub>4</sub> O	1489672	1	0
	C <sub>6</sub> H <sub>8</sub> O <sub>4</sub>	224720	259	21
	C <sub>6</sub> H <sub>12</sub> N <sub>2</sub> O <sub>2</sub>	1285303	301	12
	C <sub>6</sub> H <sub>16</sub> N <sub>4</sub>	82510	24	3
	C <sub>7</sub> N <sub>2</sub> O <sub>2</sub>	1000798	0	0
	C <sub>7</sub> H <sub>4</sub> N <sub>4</sub>	54741129	24	1
	C <sub>7</sub> H <sub>12</sub> O <sub>3</sub>	105625	1064	49
	C <sub>7</sub> H <sub>16</sub> N <sub>2</sub> O	114992	262	5
	C <sub>8</sub> O <sub>3</sub>	71079	0	0
	C <sub>8</sub> H <sub>4</sub> N <sub>2</sub> O	62428843	35	1
C <sub>8</sub> H <sub>16</sub> O <sub>2</sub>	13190	1062	91	
C <sub>8</sub> H <sub>20</sub> N <sub>2</sub>	1856	144	17	
C <sub>9</sub> H <sub>4</sub> O <sub>2</sub>	5541857	2	0	
C <sub>9</sub> H <sub>8</sub> N <sub>2</sub>	61680587	171	47	
C <sub>9</sub> H <sub>20</sub> O	405	175	48	
C <sub>10</sub> H <sub>8</sub> O	9693195	84	8	
C <sub>11</sub> H <sub>12</sub>	2135717	324	22	
C <sub>12</sub>	171886	0	0	
145	CH <sub>3</sub> N <sub>7</sub> O <sub>2</sub>	117237	0	0

$m$	$\beta$	$ \mathcal{M}_\beta^C $	$BS$	$MS$
	CH <sub>7</sub> N <sub>9</sub>	23255	0	0
	C <sub>2</sub> H <sub>3</sub> N <sub>5</sub> O <sub>3</sub>	599924	0	0
	C <sub>2</sub> H <sub>7</sub> N <sub>7</sub> O	563976	1	0
	C <sub>3</sub> H <sub>3</sub> N <sub>3</sub> O <sub>4</sub>	768517	0	0
	C <sub>3</sub> H <sub>7</sub> N <sub>5</sub> O <sub>2</sub>	2882173	5	0
	C <sub>3</sub> H <sub>11</sub> N <sub>7</sub>	178007	0	0
	C <sub>4</sub> H <sub>3</sub> NO <sub>5</sub>	186967	1	0
	C <sub>4</sub> H <sub>7</sub> N <sub>3</sub> O <sub>3</sub>	3682739	43	0
	C <sub>4</sub> H <sub>11</sub> N <sub>5</sub> O	1346659	3	0
	C <sub>5</sub> H <sub>7</sub> NO <sub>4</sub>	891604	63	0
	C <sub>5</sub> H <sub>11</sub> N <sub>3</sub> O <sub>2</sub>	2089121	104	5
	C <sub>5</sub> H <sub>15</sub> N <sub>5</sub>	89592	1	0
	C <sub>6</sub> H <sub>3</sub> N <sub>5</sub>	25665747	8	1
	C <sub>6</sub> H <sub>11</sub> NO <sub>3</sub>	575709	394	22
	C <sub>6</sub> H <sub>15</sub> N <sub>3</sub> O	233221	72	0
	C <sub>7</sub> H <sub>3</sub> N <sub>3</sub> O	55071818	4	0
	C <sub>7</sub> H <sub>15</sub> NO <sub>2</sub>	86195	554	20
	C <sub>7</sub> H <sub>19</sub> N <sub>3</sub>	4238	29	4
	C <sub>8</sub> H <sub>3</sub> NO <sub>2</sub>	16228009	1	0
	C <sub>8</sub> H <sub>7</sub> N <sub>3</sub>	97109499	144	16
C <sub>8</sub> H <sub>19</sub> NO	2876	249	10	
C <sub>9</sub> H <sub>7</sub> NO	49865161	148	21	
C <sub>10</sub> H <sub>11</sub> N	14778466	303	32	
146	CH <sub>2</sub> N <sub>6</sub> O <sub>3</sub>	76720	0	0
	CH <sub>6</sub> N <sub>8</sub> O	97234	0	0
	C <sub>2</sub> H <sub>2</sub> N <sub>4</sub> O <sub>4</sub>	216893	0	0
	C <sub>2</sub> H <sub>6</sub> N <sub>6</sub> O <sub>2</sub>	971399	1	0
	C <sub>2</sub> H <sub>10</sub> N <sub>8</sub>	57508	0	0
	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub> O <sub>5</sub>	137656	0	0
	C <sub>3</sub> H <sub>6</sub> N <sub>4</sub> O <sub>3</sub>	2429018	10	1
	C <sub>3</sub> H <sub>10</sub> N <sub>6</sub> O	749873	0	0
	C <sub>4</sub> H <sub>2</sub> O <sub>6</sub>	9986	1	0
	C <sub>4</sub> H <sub>6</sub> N <sub>2</sub> O <sub>4</sub>	1432731	22	0
	C <sub>4</sub> H <sub>10</sub> N <sub>4</sub> O <sub>2</sub>	2125930	33	1
	C <sub>4</sub> H <sub>14</sub> N <sub>6</sub>	68990	0	0
	C <sub>5</sub> H <sub>2</sub> N <sub>6</sub>	7055345	1	0
	C <sub>5</sub> H <sub>6</sub> O <sub>5</sub>	95870	28	2
	C <sub>5</sub> H <sub>10</sub> N <sub>2</sub> O <sub>3</sub>	1360645	153	9
	C <sub>5</sub> H <sub>14</sub> N <sub>4</sub> O	311390	6	0
	C <sub>6</sub> H <sub>2</sub> N <sub>4</sub> O	26123593	3	0
	C <sub>6</sub> H <sub>10</sub> O <sub>4</sub>	97394	345	25
	C <sub>6</sub> H <sub>14</sub> N <sub>2</sub> O <sub>2</sub>	257122	249	3
	C <sub>6</sub> H <sub>18</sub> N <sub>4</sub>	6742	7	2
	C <sub>7</sub> H <sub>2</sub> N <sub>2</sub> O <sub>2</sub>	17388955	0	0
	C <sub>7</sub> H <sub>6</sub> N <sub>4</sub>	96024197	94	10
	C <sub>7</sub> H <sub>14</sub> O <sub>3</sub>	22151	672	36
	C <sub>7</sub> H <sub>18</sub> N <sub>2</sub> O	9780	52	2
	C <sub>8</sub> H <sub>2</sub> O <sub>3</sub>	1187784	2	0
	C <sub>8</sub> H <sub>6</sub> N <sub>2</sub> O	109240025	177	14
	C <sub>8</sub> H <sub>18</sub> O <sub>2</sub>	1225	334	28
C <sub>9</sub> H <sub>6</sub> O <sub>2</sub>	9660231	45	4	
C <sub>9</sub> H <sub>10</sub> N <sub>2</sub>	46024195	411	22	
C <sub>10</sub> H <sub>10</sub> O	7288733	421	34	
C <sub>11</sub> H <sub>14</sub>	950064	450	52	
C <sub>12</sub> H <sub>2</sub>	3571212	1	0	
147	CHN <sub>5</sub> O <sub>4</sub>	24429	0	0
	CH <sub>5</sub> N <sub>7</sub> O <sub>2</sub>	176798	0	0
	CH <sub>9</sub> N <sub>9</sub>	10912	0	0
	C <sub>2</sub> HN <sub>3</sub> O <sub>5</sub>	37974	0	0
	C <sub>2</sub> H <sub>5</sub> N <sub>5</sub> O <sub>3</sub>	908888	0	0
	C <sub>2</sub> H <sub>9</sub> N <sub>7</sub> O	265965	1	0
	C <sub>3</sub> HNO <sub>6</sub>	10555	0	0

## 326 ANHANG E. ISOMERE NACH BRUTTOFORMEL UND MASSE

$m$	$\beta$	$ M_{\beta}^C $	$BS$	$MS$
	$C_3H_5N_3O_4$	1164356	2	0
	$C_3H_9N_5O_2$	1374370	4	0
	$C_3H_{13}N_7$	36852	0	0
	$C_4HN_7$	1022466	0	0
	$C_4H_5NO_5$	282310	9	0
	$C_4H_9N_3O_3$	1783871	41	0
	$C_4H_{13}N_5O$	281769	0	0
	$C_5HN_5O$	6233092	0	0
	$C_5H_9NO_4$	440821	105	7
	$C_5H_{13}N_3O_2$	448538	12	0
	$C_5H_{17}N_5$	7578	0	0
	$C_6HN_3O_2$	7967364	1	1
	$C_6H_5N_5$	61444491	37	2
	$C_6H_{13}NO_3$	128380	290	12
	$C_6H_{17}N_3O$	20368	7	0
	$C_7HNO_3$	1833789	0	0
	$C_7H_5N_3O$	131358449	91	9
	$C_7H_{17}NO_2$	8000	102	5
	$C_8H_5NO_2$	38484571	75	8
	$C_8H_9N_3$	83983472	235	11
	$C_8H_9NO$	43311373	429	50
	$C_{10}H_{13}N$	7122614	445	29
	$C_{11}HN$	6593791	1	0
148	$CN_4O_5$	3062	0	0
	$CH_4N_6O_3$	174687	0	0
	$CH_8N_8O$	54187	0	0
	$C_2N_2O_6$	2518	0	0
	$C_2H_4N_4O_4$	493211	0	0
	$C_2H_8N_6O_2$	547544	2	0
	$C_2H_{12}N_8$	13101	0	0
	$C_3N_8$	60490	0	0
	$C_3O_7$	217	0	0
	$C_3H_4N_2O_5$	310668	2	0
	$C_3H_8N_4O_3$	1387392	0	0
	$C_3H_{12}N_6O$	171753	0	0
	$C_4N_6O$	582583	0	0
	$C_4H_4O_6$	21975	3	1
	$C_4H_8N_2O_4$	831647	45	1
	$C_4H_{12}N_4O_2$	497254	0	0
	$C_4H_{16}N_6$	6153	0	0
	$C_5N_4O_2$	1310920	0	0
	$C_5H_4N_6$	25346382	9	1
	$C_5H_8O_5$	56687	52	4
	$C_5H_{12}N_2O_3$	328357	44	1
	$C_5H_{16}N_4O$	28301	0	0
	$C_6N_2O_3$	693682	0	0
	$C_6H_4N_4O$	93583559	38	5
	$C_6H_{12}O_4$	24562	267	16
	$C_6H_{16}N_2O_2$	24545	27	2
	$C_7O_4$	42867	0	0
	$C_7H_4N_2O_2$	61817403	27	0
	$C_7H_8N_4$	98791068	220	15
	$C_7H_{16}O_3$	2275	192	15
	$C_8H_4O_3$	4161969	7	2
	$C_8H_8N_2O$	112562582	391	16
	$C_9H_8O_2$	9990575	255	25
	$C_9H_{12}N_2$	24399762	397	20
	$C_{10}N_2$	3115390	1	0
	$C_{10}H_{12}O$	3916111	892	90
	$C_{11}O$	455822	0	0
	$C_{11}H_{16}$	323512	682	57
	$C_{12}H_4$	16079924	0	0
149	$CH_3N_5O_4$	99306	0	0
	$CH_7N_7O_2$	121630	0	0
	$CH_{11}N_9$	2766	0	0
	$C_2H_3N_3O_5$	152977	0	0
	$C_2H_7N_5O_3$	633408	0	0
	$C_2H_{11}N_7O$	67609	0	0
	$C_3H_3NO_6$	41580	0	0
	$C_3H_7N_3O_4$	822099	2	0
	$C_3H_{11}N_5O_2$	355574	0	0
	$C_3H_{15}N_7$	3483	0	0
	$C_4H_3N_7$	6505400	0	0
	$C_4H_7NO_5$	202072	10	0
	$C_4H_{11}N_3O_3$	473871	1	0
	$C_4H_{15}N_5O$	26983	0	0
	$C_5H_3N_5O$	39760215	5	0
	$C_5H_{11}NO_4$	121350	52	0
	$C_5H_{15}N_3O_2$	44621	0	0
	$C_6H_3N_3O_2$	50459744	2	0
	$C_6H_7N_5$	77737459	129	10
	$C_6H_{15}NO_3$	13539	23	2
	$C_7H_3NO_3$	11449751	3	0
	$C_7H_7N_3O$	166085562	222	7
	$C_8H_7NO_2$	48687255	265	13
	$C_8H_{11}N_3$	49755227	182	5
	$C_9H_{11}NO$	25895621	724	45
	$C_{10}H_{15}N$	2569697	558	40
	$C_{11}H_3N$	53109027	0	0
150	$CH_2N_4O_5$	31784	0	0
	$CH_6N_6O_3$	155356	0	0
	$CH_{10}N_8O$	15509	0	0
	$C_2H_2N_2O_6$	25361	0	0
	$C_2H_6N_4O_4$	443749	0	0
	$C_2H_{10}N_6O_2$	159347	0	0
	$C_2H_{14}N_8$	1341	0	0
	$C_3H_2N_8$	966328	0	0
	$C_3H_2O_7$	2113	0	0
	$C_3H_6N_2O_5$	282171	0	0
	$C_3H_{10}N_4O_3$	413244	0	0
	$C_3H_{14}N_6O$	17528	0	0
	$C_4H_2N_6O$	9630475	0	0
	$C_4H_6O_6$	20050	7	1
	$C_4H_{10}N_2O_4$	255379	0	0
	$C_4H_{14}N_4O_2$	52358	0	0
	$C_5H_2N_4O_2$	21728759	0	0
	$C_5H_6N_6$	41205407	88	8
	$C_5H_{10}O_5$	18092	34	3
	$C_5H_{14}N_2O_3$	36231	0	0
	$C_6H_2N_2O_3$	11366726	0	0
	$C_6H_6N_4O$	151838122	273	11
	$C_6H_{14}O_4$	2922	61	5
	$C_7H_2O_4$	674033	0	0
	$C_7H_6N_2O_2$	100082479	153	3
	$C_7H_{10}N_4$	66583863	105	1
	$C_8H_6O_3$	6717404	90	7
	$C_8H_{10}N_2O$	76307072	542	38
	$C_9H_{10}O_2$	6843602	667	71
	$C_9H_{14}N_2$	9459132	568	29
	$C_{10}H_2N_2$	65563828	0	0
	$C_{10}H_{14}O$	1548361	1938	150
	$C_{11}H_2O$	9414509	0	0
	$C_{11}H_{18}$	84051	762	49
	$C_{12}H_6$	34030905	12	0

# Literaturverzeichnis

- [1] ALLINGER, N. L.: *MM2. A Hydrocarbon Force Field Utilizing  $V_1$  and  $V_2$  Torsional Terms*. J. Am. Chem. Soc., 99:8127–8134, 1977.
- [2] AUGUSTIN, V.: *Computerunterstützte Berechnung von Symmetrien unscharfer Strukturen*. Diplomarbeit, Universität Bayreuth, 2004.
- [3] BADERTSCHER, M., A. KORYTKO, K.-P. SCHULZ, M. MADISON, M. E. MUNK, P. PORTMAN, M. JUNGHANS, P. FONTANA und E. PRETSCH: *Assemble 2.0: A Structure Generator*. Chemom. Intel. Lab. Syst., 51:73–79, 2000.
- [4] BALABAN, A. T.: *Highly Discriminating Distance-Based Topological Index*. Chem. Phys. Lett., 89:399–404, 1982.
- [5] BALABAN, A. T.: *Topological Indices Based on Topological Distances in Molecular Graphs*. Pure Appl. Chem., 55:199–206, 1983.
- [6] BARNARD, J. M. und G. M. DOWNS: *Computer Representation and Manipulation of Combinatorial Libraries*. Band 7/8 der Reihe *Perspectives in Drug Discovery and Design*, Seiten 13–30. Kluwer Academic Publishers, 1997.
- [7] BARNARD, J. M. und G. M. DOWNS: *Use of Markush Structure Techniques to Avoid Enumeration in Diversity Analysis of Large Combinatorial Libraries*. <http://www.daylight.com/meetings/mug97/Barnard/970227JB.html>, 1997.
- [8] BARNARD, J. M., G. M. DOWNS und A. v. SCHOLLEY-PFAB: *Use of Markush Structure Analysis Techniques for Descriptor Generation and Clustering of Large Combinatorial Libraries*. J. Mol. Graph. & Mod., 18:452–463, 2000.
- [9] BASAK, S. C.: *Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach*. Med. Sci. Res., 15:605–609, 1987.

- [10] BASAK, S. C.: *Information Theoretic Indices of Neighborhood Complexity and their Applications*. In: DEVILLERS, J. und A. T. BALABAN (Herausgeber): *Topological Indices and Related Descriptors in QSAR and QSPR*, Kapitel 12. Gordon and Breach, Amsterdam, 1999.
- [11] BAUERSCHMIDT, S.: *Repräsentation von Molekülstrukturen zur computergestützten Behandlung chemischer Reaktionen*. Doktorarbeit, Friedrich–Alexander–Universität Erlangen–Nürnberg, 1997.
- [12] BENECKE, C.: *Objektorientierte Darstellung und Algorithmen zur Klassifizierung bewerteter endlicher Strukturen*. Doktorarbeit, Universität Bayreuth, 1997.
- [13] BENECKE, C., R. GRUND, R. HOHBERGER, R. LAUE, A. KERBER und T. WIELAND: *MOLGEN+, a Generator of Connectivity Isomers and Stereoisomers for Molecular Structure Elucidation*. *Anal. Chim. Acta*, 314:141–147, 1995.
- [14] BENECKE, C., T. GRÜNER, A. KERBER, R. LAUE und T. WIELAND: *Molecular Structure Generation with MOLGEN, new Features and Future Developments*. *Fresenius J. Anal. Chem.*, 358:23–32, 1997.
- [15] BENKÖ, G.: *A Toy Model of Chemical Reaction Networks*. Diplomarbeit, Universität Wien, 2002.
- [16] BENKÖ, G., C. FLAMM und F. STADLER: *A Graph–Based Toy Model of Chemistry*. *J. Chem. Inf. Comput. Sci.*, 43:1085–1093, 2003.
- [17] BIEBL, J.: *Computerunterstützte Clusteranalyse mit fuzzy Clusteralgorithmen und Anwendungen auf chemische Moleküle*. Diplomarbeit, Universität Bayreuth, 1999.
- [18] BÖCKER, J.: *Spektroskopie*. Vogel Buchverlag, Würzburg, 1997.
- [19] BRAUN, J.: *Topologische Indizes und ihre computerunterstützte Anwendung in der Chemie*. Diplomarbeit, Universität Bayreuth, 1999.
- [20] BRAUN, J., R. GUGISCH, A. KERBER, R. LAUE, M. MERINGER und C. RÜCKER: *MOLGEN–CID, A Canonizer for Molecules and Graphs Accessible through the Internet*. *J. Chem. Inf. Comput. Sci.*, 44:542–548, 2004.
- [21] BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN und C. J. STONE: *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, 1984.

- [22] BUTINA, D. und J. M. R. GOLA: *Modeling Aqueous Solubility*. J. Chem. Inf. Comput. Sci., 43:837–841, 2003.
- [23] BYVATOV, E., U. FECHNER, J. SADOWSKI und G. SCHNEIDER: *Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification*. J. Chem. Inf. Comput. Sci., 43:1882–1889, 2003.
- [24] CARELL, T., E. A. WINTNER, A. BASHIR-HASHEMI und J. REBEK JR.: *Neuartiges Verfahren zur Herstellung von Bibliotheken kleiner organischer Moleküle*. Angew. Chemie, 106:2159–2161, 1994.
- [25] CARELL, T., E. A. WINTNER, A. J. SUTHERLAND, J. REBEK JR. und Y. M. DUNAYEVSKIY: *New Promise in Combinatorial Chemistry: Synthesis, Characterization, and Screening of Small-Molecule Libraries in Solution*. Chem. & Biol., 2:171–183, 1995.
- [26] CHANG, C.-C. und C.-J. LIN: *LIBSVM — A Library for Support Vector Machines*. <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>, 2002.
- [27] CHRISTIE, B. D. und M. E. MUNK: *The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation*. J. Am. Chem. Soc., 113:3750–3757, 1991.
- [28] COLBORN, C. J. und R. C. READ: *Orderly Algorithms for Generating Restricted Classes of Graphs*. J. Graph Theory, 3:187–195, 1979.
- [29] CONNOLLY, M. L.: *Molecular Surface and Volume*. In: SCHLEYER, P. v. R. (Herausgeber): *Encyclopedia of Computational Chemistry*, Seiten 1698–1703. Wiley, Chichester, 1998.
- [30] CORTES, C. und V. VAPNIK: *Support-Vector Network*. Machine Learning, 20:1–25, 1995.
- [31] DUDA, R. O. und P. E. HART: *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.
- [32] DUDLEY, H. J. und I. FLEMMING: *Spektroskopische Methoden zur Strukturaufklärung*. Georg Thieme Verlag, Stuttgart, 1975.
- [33] DUGUNDJI, J. und I. UGI: *An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs*. Topics In Current Chemistry, 39:19–64, 1973.

- [34] ELYASHBERG, M. E., K. A. BLINOV und E. R. MARTIROSIAN: *A New Approach to Computer-Aided Molecular Structure Elucidation: The Expert System Structure Elucidator*. Lab. Autom. Inf. Man., 34:15–30, 1999.
- [35] ELYASHBERG, M. E., Y. Z. KARASEV, E. R. MARTIROSIAN, H. THIELE und H. SOMBERG: *Expert Systems as a Tool for the Molecular Structure Elucidation by Spectral Methods. Strategies of Solution to the Problems*. Anal. Chim. Acta, 348:443–463, 1997.
- [36] ELYASHBERG, M. E., E. R. MARTIROSIAN, Y. Z. KARASEV, H. THIELE und H. SOMBERG: *X-PERT: A User Friendly Expert System for Molecular Structure Elucidation by Spectral Methods*. Anal. Chim. Acta, 337:265–286, 1997.
- [37] ELYASHBERG, M. E., V. V. SEROV, E. R. MARTIROSIAN, L. A. ZLATINA, Y. Z. KARASEV, V. N. KOLDASHOV und Y. Y. YAMPOLSKIY: *An Expert System for the Molecular Structure Elucidation Based on Spectral Data*. J. Mol. Struct., 230:191–205, 1991.
- [38] FARADZHEV, I. A.: *Constructive Enumeration of Combinatorial Objects*. Problèmes Combinatoires et Théorie des Graphes, 260:131–135, 1978. Colloq. Internat. CNRS, University of Orsay, Orsay 1976.
- [39] FARADZHEV, I. A.: *Generation of Nonisomorphic Graphs with a Given Degree Sequence*, Seiten 11–19. Algorithmic Studies in Combinatorics. NAUKA, Moskau, 1978. In Russisch.
- [40] FISCHER, D.: *Verwendung graphentheoretischer Netzwerkalgorithmen bei Zuordnungsproblemen, insbesondere bei der automatischen Analyse von Spektren*. Diplomarbeit, Universität Bayreuth, 1996.
- [41] FISCHER, M.: *Algorithmen zur Erkennung von Aromatizitäts- und Tautomerie-Überlagerungen in organischen Strukturen*. Diplomarbeit, Universität Bayreuth, 1996.
- [42] FUJITA, S.: *Computer-Oriented Representation of Organic Reactions*. Yoshioka Shoten Publishing Company, Kyoto, 2001.
- [43] FUNATSU, K., N. MIYABAYASKI und S. SASAKI: *Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS*. J. Chem. Inf. Comput. Sci., 28:18–28, 1988.

- [44] FUNATSU, K. und SASAKI S.: *Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates.* J. Chem. Inf. Comput. Sci., 36:190–204, 1996.
- [45] FURNIVAL, G. und R. WILSON: *Regression by Leaps and Bounds.* Technometrics, 16:499–511, 1974.
- [46] GASTEIGER, J., W. HANEBECK und K.-P. SCHULZ: *Prediction of Mass Spectra from Structural Information.* J. Chem. Inf. Comput. Sci., 32:264–271, 1992.
- [47] GASTEIGER, J., W. HANEBECK, K.-P. SCHULZ, S. BAUERSCHMIDT und R. HÖLLERING: *Automatic Analysis and Simulation of Mass Spectra.* Band 4 der Reihe *Computer-Enhanced Analytical Spectroscopy*, Seiten 97–133. Kluwer Academic Publishers, 1993.
- [48] GIBSON, K. und H. SCHERAGA: *Exact Calculation of the Volume and the Surface Area of Fused Hard Sphere Molecule with Unequal Atomic Radii.* Mol. Phys., 62:1247–1265, 1987.
- [49] GRUBER, B.: *Eine lineare Algebraische Repräsentation für Objekte der Synthesplanung*, Band 9 der Reihe *Software-Entwicklung in der Chemie*, Seiten 99–111. Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1995.
- [50] GRUND, R.: *Konstruktion schlichter Graphen mit gegebener Gradpartition.* Bayreuther Mathematische Schriften, 44:73–104, 1993.
- [51] GRUND, R.: *Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten.* Bayreuther Mathematische Schriften, 49:1–113, 1995.
- [52] GRÜNER, T.: *Ein neuer Ansatz zur rekursiven Erzeugung von schlichten Graphen.* Diplomarbeit, Universität Bayreuth, 1995.
- [53] GRÜNER, T.: *Strategien zur Konstruktion diskreter Strukturen.* Doktorarbeit, Universität Bayreuth, 1998.
- [54] GRÜNER, T.: *Strategien zur Konstruktion diskreter Strukturen und ihre Anwendung auf molekulare Graphen.* MATCH — Commun. Math. Comput. Chem., 39:39–126, 1999.

- [55] GRÜNER, T., A. KERBER, R. LAUE, M. LIEPELT, M. MERINGER, K. VARMUZA und W. WERTHER: *Bestimmung von Summenformeln aus Massenspektren durch Erkennung überlagerter Isotopenmuster*. MATCH — Commun. Math. Comput. Chem., 37:163–177, 1998.
- [56] GRÜNER, T., A. KERBER, R. LAUE und M. MERINGER: *MOLGEN 4.0*. MATCH — Commun. Math. Comput. Chem., 37:205–208, 1998.
- [57] GRÜNER, T., A. KERBER, R. LAUE und M. MERINGER: *Mathematics for Combinatorial Chemistry*, Band II der Reihe *Scientific Computing in Chemical Engineering*, Seiten 74–81. Springer Verlag, 1999.
- [58] GRÜNER, T., A. KERBER, R. LAUE, M. MERINGER, K. VARMUZA und W. WERTHER: *MASSMOL*. MATCH — Commun. Math. Comput. Chem., 38:173–180, 1998.
- [59] GUGISCH, R., A. KERBER, R. LAUE, M. MERINGER und J. WEIDINGER: *MOLGEN-COMB, a Software Package for Combinatorial Chemistry*. MATCH — Commun. Math. Comput. Chem., 41:189–203, 2000.
- [60] GUTMAN, I., B. RUŠČIĆ, N. TRINAJSTIĆ und C. F. WILCOX JR.: *Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes*. J. Chem. Phys., 62:3399–3405, 1975.
- [61] HAGER, R., A. KERBER, R. LAUE, D. MOSER und W. WEBER: *Construction of Orbit Representatives*. Bayreuther Mathematische Schriften, 35:157–169, 1991.
- [62] HANEBECK, W.: *Simulation und Rekonstruktion von Reaktionen im Massenspektrometer*. Doktorarbeit, Technische Universität München, 1991.
- [63] HARARY, F.: *Graphentheorie*. R. Oldenbourg Verlag, München, Wien, 1974.
- [64] HASTIE, T., R. TIBSHIRANI und J. FRIEDMAN: *The Elements of Statistical Learning*. Springer Verlag, New York, Berlin, Heidelberg, 2001.
- [65] HAWKINS, D. M.: *The Problem of Overfitting*. J. Chem. Inf. Comput. Sci., 44:1–12, 2004.
- [66] HESSE, M., H. MEIER und B. ZEEH: *Spektroskopische Methoden in der Organischen Chemie*. Georg Thieme Verlag, Stuttgart, New York, 1991.



- [67] HEUERDING, S. und T. CLERC: *Simple Tools for the Computer-Aided Interpretation of Mass Spectra*. Chemom. Intel. Lab. Syst., 20:57–69, 1993.
- [68] HÖLLERING, R.: *Simulation von Massenspektren und Entwicklung eines Systems zur Reaktionsvorhersage*. Doktorarbeit, Friedrich–Alexander–Universität Nürnberg–Erlangen, 1998.
- [69] IHAKA, R. und R. GENTLEMAN: *R: A Language for Data Analysis and Graphics*. J. Comput. Graph. Stat., 5:299–314, 1996.
- [70] IRTH, H., S. LONG und T. SCHENK: *High-Resolution Screening in an Expanded Chemical Space*. Current Drug Discovery, Seiten 19–23, January 2004.
- [71] KATRITZKY, A. R., V. S. LOBANOV und M. KARELSON: *CODESSA: Reference Manual, Version 2*. University of Florida, 1994.
- [72] KERBER, A.: *Algebraic Combinatorics via Finite Group Actions*. Wissenschaftsverlag, Berlin, Heidelberg, New York, 2. Auflage, 1999.
- [73] KERBER, A. und A. KOHNERT: *Online Applications of SYMMETRICA to the Enumeration of Permutational Isomers and to the Enumeration of Certain Combinatorial Libraries*. MATCH — Commun. Math. Comput. Chem., 38:163–172, 1998.
- [74] KERBER, A., R. LAUE und M. MERINGER: *An Application of the Structure Generator MOLGEN to Patents in Chemistry*. MATCH — Commun. Math. Comput. Chem., 47:169–172, 2003.
- [75] KERBER, A., R. LAUE, M. MERINGER und C. RÜCKER: *MOLGEN-QSPR, a Software Package for the Search of Quantitative Structure Property Relationships*. MATCH — Commun. Math. Comput. Chem., 51:187–204, 2004.
- [76] KERBER, A., R. LAUE, M. MERINGER und K. VARMUZA: *MOLGEN-MS: Evaluation of Low Resolution Electron Impact Mass Spectra with MS Classification and Exhaustive Structure Generation*, Band 15 der Reihe *Advances in Mass Spectrometry*, Seiten 939–940. Wiley, 2001.
- [77] KIER, L. B. und HALL L. H.: *The Nature of Structure-Activity Relationships and their Relation to Molecular Connectivity*. Eur. J. Med. Chem., 12:307–312, 1977.

- [78] KIER, L. B. und HALL L. H.: *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Chichester, 1986.
- [79] KIER, L. B., W. J. MURRAY, M. RANDIĆ und L. H. HALL: *Molecular Connectivity V: Connectivity Series Applied to Density*. J. Pharm. Sci., 65:1226–1230, 1976.
- [80] KONSTANTINOVA, E. V. und V. A. SKOROBOGATOV: *Molecular Hypergraphs: The New Representation of Nonclassical Molecular Structures with Polycentric Delocalized Bonds*. J. Chem. Inf. Comput. Sci., 35:472478, 1995.
- [81] KONSTANTINOVA, E. V. und V. A. SKOROBOGATOV: *Application of Hypergraph Theory in Chemistry*. Discrete Math., 235:365–383, 2001.
- [82] KRUGLINSKI, D., S. WINGO und G. SHEPHERD: *Inside Visual C++ 6.0*. Microsoft-Press, Redmond, Washington, 1998.
- [83] KUMAR, K. und A. G. MENON: *Computer-Assisted Determination of Elemental Composition of Fragments in Mass Spectra*. Rapid Comm. Mass Spec., 6:585–591, 1992.
- [84] LAUE, R.: *Construction of Combinatorial Objects — A Tutorial*. Bayreuther Mathematische Schriften, 43:53–96, 1993.
- [85] LAUE, R., T. GRÜNER, M. MERINGER und A. KERBER: *Constrained Generation of Molecular Graphs*. DIMACS Series in Discrete Mathematics And Theoretical Computer Science. American Mathematical Society. (In Druck).
- [86] LEBEDEV, K. S. und D. CABROL-BASS: *New Computer Aided Methods for Revealing Structural Features of Unknown Compounds Using Low Resolution Mass Spectra*. J. Chem. Inf. Comput. Sci., 38:410–419, 1998.
- [87] LINDSAY, R. K., B. G. BUCHANAN, E. A. FEIGENBAUM und J. LEDERBERG: *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. McGraw-Hill Book Company, New York, St. Louis, San Francisco, 1980.
- [88] LUINGE, H. J.: *EXSPEC: A Knowledge-Based System for Structure Analysis of Organic Molecules from Combined Spectral Data*. Doktorarbeit, Universität Utrecht, 1989.

- [89] LUINGE, H. J. und VAN DER MAAS J. H.: *Artificial Intelligence for the Interpretation of Combined Spectral Data. Design and Development of a Spectrum Interpreter*. Anal. Chim. Acta, 223:135–147, 1989.
- [90] LUKS, E.: *Isomorphism of Graphs of Bounded Valence Can be Tested in Polynomial Time*. J. Computer Syst. Sci., 25:42–65, 1982.
- [91] MCLAFFERTY, F. W. und F. TURECEK: *Interpretation of Mass Spectra*. University Science Books, Mill Valley, California, 4. Auflage, 1993.
- [92] MEILER, J. und M. MERINGER: *Ranking MOLGEN Structure Proposals by  $^{13}\text{C}$  NMR Chemical Shift Prediction with ANALYZE*. MATCH — Commun. Math. Comput. Chem., 45:85–108, 2002.
- [93] MEILER, J., E. SANLI, J. JUNKER, R. MEUSINGER, T. LINDEL, M. WILL, W. MAIER und M. KÖCK: *Validation of Structural Proposals by Substructure Analysis and  $^{13}\text{C}$  NMR Chemical Shift Prediction*. J. Chem. Inf. Comput. Sci., 42:241–248, 2002.
- [94] MEILER, J. und M. WILL: *Automated Structure Elucidation of Organic Molecules from  $^{13}\text{C}$  NMR Spectra using Genetic Algorithms and Neural Networks*. J. Chem. Inf. Comput. Sci., 41:1535–1546, 2001.
- [95] MERINGER, M.: *Erzeugung regulärer Graphen*. Diplomarbeit, Universität Bayreuth, 1996.
- [96] MERINGER, M.: *Fast Generation of Regular Graphs and Construction of Cages*. J. Graph Theory, 30:137–146, 1999.
- [97] MEYER, D.: *Support Vector Machines*. R-News, 1/3, 2001. Teil des R-Pakets e1071.
- [98] MILLER, A. J.: *Subset Selection in Regression*. Chapman and Hall, London, New York, Tokyo, Melbourne, Madras, 1990.
- [99] MOLODTSOV, S. G.: *The Generation of Molecular Graphs with Obligatory, Forbidden and Desirable Fragments*. MATCH — Commun. Math. Comput. Chem., 37:157–162, 1998.
- [100] MUN, I. K., R. VENKATARAGHAVAN und F. W. MCLAFFERTY: *Computer Prediction of Molecular Weights from Mass Spectra*. Anal. Chem., 53:179–182, 1981.

- [101] MUN, I. K., R. VENKATARAGHAVAN und F. W. MCLAFFERTY: *Molecular Weight Parity Predicted from the Parity of Mass Spectral Peaks*. *Organic Mass Spectroscopy*, 16:82–84, 1981.
- [102] NEUDERT, R. und M. PENK: *Enhanced Structure Elucidation*. *J. Chem. Inf. Comput. Sci.*, 36:244–248, 1996.
- [103] NIKOLIĆ, S., G. KOVAČEVIĆ, A. MILIČEVIĆ und N. TRINAJSTIĆ: *The Zagreb Indices 30 Years After*. *Croat. Chem. Acta*, 76:113–127, 2003.
- [104] OTTO, M.: *Chemometrie: Statistik und Computereinsatz in der Analytik*. VCH Verlagsgesellschaft, Weinheim, 1997.
- [105] PENCHEV P. N., ANDREEV G. N., VARMUZA K.: *Automatic Classification of Infrared Spectra Using a Set of Improved Expert-Based Features*. *Anal. Chim. Acta*, 388:145–159, 1999.
- [106] PÓLYA, G.: *Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen*. *Acta Mathematica*, 68:145–253, 1937.
- [107] POZO, R.: *Template Numerical Toolkit*. <http://math.nist.gov/tnt>, 2002.
- [108] PRETSCH, E., P. BÜHLMANN, C. AFFOLTER und M. BADERTSCHER: *Spektroskopische Daten zur Strukturaufklärung organischer Verbindungen*. Springer Verlag, Berlin, Heidelberg, New York, 2001.
- [109] PRETSCH, E. und J. T. CLERC: *Spectra Interpretation of Organic Compounds*. VCH Verlagsgesellschaft, Weinheim, New York, Basel, Cambridge, Tokyo, 1997.
- [110] QUINLAN, J. R.: *Learning with Continuous Classes*. In: ADAMS und STERLING (Herausgeber): *Proc. AI '92*, Seiten 343–348, 1992.
- [111] RANDIĆ, M.: *On Characterization of Molecular Branching*. *J. Am. Chem. Soc.*, 97:6609–6615, 1975.
- [112] READ, R. C.: *Everyone a Winner*. *Annals of Discrete Mathematics*, 2:107–120, 1978.
- [113] RENAU, T. E., J. P. SANCHEZ, J. W. GAGE, J. A. DEVER, M. A. SHAPIRO, S. J. GRACHECK und J. M. DOMAGALA: *Structure–Activity*

- Relationships of the Quinolone Antibacterials against Mycobacteria: Effect of Structural Changes at N-1 and C-7.* J. Med. Chem., 39:729–735, 1996.
- [114] RINNE, H.: *Taschenbuch der Statistik*. Wissenschaftlicher Verlag Harry Deutsch, Frankfurt am Main, 3. Auflage, 2003.
- [115] RIPLEY, B. D.: *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.
- [116] ROBERTS, A. P. und M. A. KNACKSTEDT: *Structure–Property Correlations in Model Composite Materials*. Phys. Review E, 54:2313–2328, 1996.
- [117] RUCH, E., W. HÄSSELBARTH und B. RICHTER: *Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung*. Theor. Chim. Acta, 19:288–300, 1970.
- [118] RUCH, E. und D. J. KLEIN: *Double Cosets in Chemistry and Physics*. Theor. Chim. Acta, 63:447–472, 1983.
- [119] RÜCKER, C., J. BRAUN, A. KERBER und R. LAUE: *The Molecular Descriptors Computed with MOLGEN*. <http://www.mathe2.uni-bayreuth.de/molgenqspr>, 2003.
- [120] RÜCKER, C. und M. MERINGER: *How Many Organic Compounds are Graph–Theoretically Nonplanar?* MATCH — Commun. Math. Comput. Chem., 45:153–172, 2002.
- [121] RÜCKER, C. und G. RÜCKER: *Counts of All Walks as Atomic and Molecular Descriptors*. J. Chem. Inf. Comput. Sci., 33:683–695, 1993.
- [122] RÜCKER, C. und G. RÜCKER: *Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules*. J. Chem. Inf. Comput. Sci., 40:99–106, 2000.
- [123] SADOWSKI, J. und J. GASTEIGER: *From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders*. Chem. Reviews, 93:2567–2581, 1993.
- [124] SADOWSKI, J., J. GASTEIGER und G. KLEBE: *Comparison of Automatic Three-dimensional Model Builders Using 639 X-Ray Structures*. J. Chem. Inf. Comput. Sci., 34:1000–1008, 1994.

- [125] SARDANA, S. und A. K. MADAN: *Application of Graph Theory: Relationship of Antimycobacterial Activity of Quinolone Derivatives with Eccentric Connectivity Index and Zagreb Group Parameters*. MATCH — Commun. Math. Comput. Chem., 45:35–53, 2002.
- [126] SCHITTKOWSKI, K.: *NLPQL; A FORTRAN Subroutine Solving Constrained Nonlinear Programming Problems*. Annals of Operations Research, 5:485–500, 1985.
- [127] SCHMALZ, B.: *Verwendung von Untergruppenleitern zur Bestimmung von Doppelnebenklassen*. Bayreuther Math. Schr., 31:109–143, 1993.
- [128] SCHULTZ, H. P.: *Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes*. J. Chem. Inf. Comput. Sci., 29:227–228, 1989.
- [129] SCHULTZ, H. P. und T. P. SCHULTZ: *Topological Organic Chemistry. 6. Graph Theory and Molecular Topological Indices of Cycloalkanes*. J. Chem. Inf. Comput. Sci., 33:240–244, 1993.
- [130] SCHULZ, K.-P.: *Computergestützte Untersuchungen über Zusammenhänge zwischen Struktur und Massenspektrum*. Doktorarbeit, Technische Universität München, 1991.
- [131] SCSIBRANY, H. und K. VARMUZA: *ToSiM: PC-Software for the Investigation of Topological Similarities in Molecules*, Band 8 der Reihe *Software-Entwicklung in der Chemie*, Seiten 235–249. Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1994.
- [132] SEEBAS, B. und E. PRETSCH: *Automated Compatibility Tests of the Molecular Formulas or Structures of Organic Compounds with their Mass Spectra*. J. Chem. Inf. Comput. Sci., 39:705–717, 1999.
- [133] SHELLEY, C. A.: *Heuristic Approach for Displaying Chemical Structures*. J. Chem. Inf. Comput. Sci., 23:61–65, 1983.
- [134] SIMS, C. C.: *Computation with permutation groups*. In: PETRICK, S. R. (Herausgeber): *Proceedings of the Second Symposium on Symbolic and Algebraic Manipulation*, Seiten 23–28, New York, 1971.
- [135] SONNTAG, M.: *Eine Anwendung des Algorithmus von Ford und Fulkerson bei der Interpretation von Infrarotspektren*. MATCH — Commun. Math. Comput. Chem., 30:37–51, 1994.

- [136] STAPLETON, J. H.: *Linear Statistical Models*. Wiley, New York, Chichester, Brisbane, Toronto, Singapore, 1995.
- [137] STEIN, S. E. ET AL.: *NIST '98 Mass Spectral Library User's Guide*. NIST Mass Spectrometry Data Center, Gaithersburg, 1998.
- [138] STROUSTRUP, B.: *Die C++ Programmiersprache*. Addison-Wesley, Bonn, Reading, 1998.
- [139] STRUPPE, C.: *Anwendungsmöglichkeiten der Methodenkombination Gaschromatographie-Atomemissionsspektroskopie in der Umweltanalytik*. Doktorarbeit, Universität Leipzig, 1998.
- [140] TEMKIN, O. N., A. V. ZEIGARNIK und D. BONCHEV: *Chemical Reaction Networks: A Graph-Theoretical Approach*. CRC Press, Boca Raton, New York, London, Tokyo, 1996.
- [141] TENHOSAARI, A.: *Computer-Assisted Composition Analysis of Unknown Compounds by Simultaneous Analysis of the Intensity Ratios of Isotope Patterns of the Molecular Ion and Daughter Ions in Low-Resolution Mass Spectra*. *Organic Mass Spectrometry*, 23:236-239, 1988.
- [142] TODESCHINI, R. und V. CONSONNI: *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, 2000.
- [143] TOPLISS, J. G. und R. P. EDWARDS: *Chance Factors in Studies of Quantitative Structure-Activity Relationships*. *J. Med. Chem.*, 22:1238, 1979.
- [144] TRINAJSTIĆ, N.: *Chemical Graph Theory*. CRC Press, Boca Raton, 2. Auflage, 1992.
- [145] UGI, I., A. DÖMLING, B. GRUBER, C. HEILINGBRUNNER, C. HEISS und W. HÖRL: *Formale Unterstützung bei Multikomponentenreaktionen: Automatisierung der Synthesechemie*, Band 9 der Reihe *Software-Entwicklung in der Chemie*, Seiten 113-128. Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1995.
- [146] VAPNIK, V.: *The Nature of Statistical Learning Theory*. Springer Verlag, New York, Berlin, Heidelberg, 1995.
- [147] VAPNIK, V.: *Statistical Learning Theory*. Wiley, New York, 1998.

- [148] VARMUZA, K.: *Pattern Recognition in Chemistry*. Springer Verlag, Berlin, 1980.
- [149] VARMUZA, K., P. HE und K.-T. FANG: *Boosting Applied to Classification of Mass Spectra*. *J. Data Sci.*, 1:391–404, 2003.
- [150] VARMUZA, K., U. JORDIS und G. WOLF: *Database Mining for Heterocycles: are Structures of Small Heterocycles Generated by a Computer Program Present in Databases*. <http://www.ch.ic.ac.uk/ectoc/echet96/papers/014/>, 1996.
- [151] VARMUZA, K., P. PENCHEV, F. STANCL und W. WERTHER: *Systematic Structure Elucidation of Organic Compounds by Mass Spectra Classification*. *J. Mol. Struct.*, 408/409:91–96, 1997.
- [152] VARMUZA, K. und H. SCSIBRANY: *Cluster Analysis of Chemical Structures Based on Binary Molecular Descriptors and Principal Component Analysis*, Band 9 der Reihe *Software-Entwicklung in der Chemie*, Seiten 81–90. Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1995.
- [153] VARMUZA, K. und H. SCSIBRANY: *Substructure Isomorphism Matrix*. *J. Chem. Inf. Comput. Sci.*, 40:308–313, 2000.
- [154] VARMUZA, K. und W. WERTHER: *Mass Spectral Classifiers for Supporting Systematic Structure Elucidation*. *J. Chem. Inf. Comput. Sci.*, 36:323–333, 1996.
- [155] VARMUZA, K., W. WERTHER, F. STANCL, A. KERBER und R. LAUE: *Computer-Assisted Structure Elucidation of Organic Compounds, based on Mass Spectra Classification and Exhaustive Isomer Generation*, Band 10 der Reihe *Software-Entwicklung in der Chemie*, Seiten 303–314. Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1996.
- [156] VARMUZA, K. ET AL.: *MSclass. Software for Chemical-Structure-Structure-Oriented Classification of Mass Spectra. Reference and User Guide*. Applied ChemoMetrics, Technische Universität Wien, 1996.
- [157] VARMUZA, K. ET AL.: *MSclass. Software for Chemical-Structure-Structure-Oriented Classification of Mass Spectra. Classifier Guide*. Applied ChemoMetrics, Technische Universität Wien, 1996.
- [158] WAGENER, M. und V. J. VAN GEERESTEIN: *Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features*. *J. Chem. Inf. Comput. Sci.*, 40:280–292, 2000.



- [159] WARR, W. A. und C. SUHR: *Chemical Information Management*. VCH Verlagsgesellschaft, Weinheim, New York, Basel, Cambridge, 1992.
- [160] WERTHER, W.: *Einsatz von Methoden der explorativen Datenanalyse zur Interpretation und Klassifikation von Massenspektren*. Doktorarbeit, Technische Universität Wien, 1992.
- [161] WERTHER, W.: *Überblick über die automatische Interpretation von EI(LR)-Massenspektren*. Unveröffentlichtes Manuskript, 1996.
- [162] WERTHER, W.: *Versuch einer Systematik der Reaktionsmöglichkeiten in der Elektronenstoß-Massenspektrometrie (EI-MS)*. Unveröffentlichtes Manuskript, 1996.
- [163] WERTHER, W., H. LOHNINGER, F. STANCL und K. VARMUZA: *Classification of Mass Spectra: A Comparison of Yes/No Classification Methods for the Recognition of Simple Structural Properties*. Chemom. Intel. Lab. Syst., 22:63–76, 1994.
- [164] WIELAND, T.: *Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung*. Doktorarbeit, Universität Bayreuth, 1996.
- [165] WIENER, H.: *Structural Determination of Parafin Boiling Point*. J. Am. Chem. Soc., 69:17–20, 1947.
- [166] WILL, M., W. FACHINGER und RICHERT J. R.: *Fully Automated Structure Elucidation – A Spectroscopists Dream Comes True*. J. Chem. Inf. Comput. Sci., 36:221–227, 1996.
- [167] WRIGHT JR., R. S. und M. SWEET: *OpenGL Superbible*. Waite Group Press, Corte Madera, 1996.
- [168] ZUPAN, J. und J. GASTEIGER: *Neural Networks for Chemists*. VCH Verlagsgesellschaft, Weinheim, New York, Basel, Cambridge, Tokyo, 1993.



# Abkürzungsverzeichnis

AAS	Atomabsorptionsspektroskopie, 168
AB	Aromatic Bond, 141
ABA	Antimycobakterielle Aktivität, 153
AES	Atomemissionsspektroskopie, 168
ANN	Artificial Neural Network, 85
BP	Boiling Point, 118
BSS	Best Subset Selection, 82
CART	Classification and Regression Trees, 88
CASE	Computer Aided Structure Elucidation, i
CSC	Closed Shell Chemistry, 19
CT	Classification Tree, 88
CV	Crossvalidation, 77
DB	Double Bond, 141
DBE	Double Bond Equivalent, 25
EEI	Even Electron Ion, 195
EI	Electron Impact, 173
FR	Fisher Ratios, 81
HN	Hidden Neuron, 85
HRS	High Resolution Screening, 170

HTS	High Throughput Screening, 95
IC	Integral Chemistry, 19
IR	Infrarot(spektroskopie), 168
KNN	$k$ nächste Nachbarn, 90
LDA	Lineare Diskriminanzanalyse, 85
LM	Lineares Modell, 83
LOO	Leave One Out, 77
LR	Low Resolution, 173
LS	Least Squares, 196
MC	Multicenter Chemistry, 19
MCE	Mean Classification Error, 75
MIC	Minimum Inhibitory Concentration, 153
MLR	Multiple lineare Regression, 83
MMD	Minimale Massendifferenzen, 263
MMG	Mehrdeutiger molekularer Graph, 27
MS	Massenspektrometrie, Massenspektrum, 168
MV	Matchvalue, 192
NMR	Nuclear Magnetic Resonanz (Spectroscopy), 168
OEI	Odd Electron Ion, 195
OLS	Ordinary Least Squares, 84
PCA	Principal Component Analysis, 164
PCR	Principal Component Regression, 84
PD	Physikalische Dichte, 140
QDA	Quadratische Diskriminanzanalyse, 162
QSAR	Quantitative Structure Activity Relationship, 69

QSPR	Quantitative Structure Property Relationship, i
RBF	Radiale Basisfunktion, 241
RC	Restricted Chemistry, 19
RRP	Relative Ranking-Position, 201
RSS	Residual Sum of Squares, 72
RT	Regression Tree, 88
SAR	Structure Activity Relationship, 69
SB	Single Bond, 141
SC	Substructure Count, 117
SE	sterische Energie, 43
SI	Soft Ionisation, 176
SP	Structural Property, 218
SVD	Singular Value Decomposition, 84
SVM	Support-Vektor-Maschine, 86
TB	Triple Bond, 141
TCE	Total Classification Error, 73
TI	Topologischer Index, 110
UV	Ultraviolett(spektroskopie), 168



# Index

- 2D-Platzierung, 41
- 3D-Platzierung, 41
- Abbildung
  - fusionierende, 5
- abhängige Variable, 71
- Abhängigkeit
  - affine, 81
- Abstand, 10
- abstrakte Basisklasse, 27
- adjazent, 8
- Adjazenzmatrix, 7
- affine Abhängigkeit, 81
- Ångström, 114
- arithmetischer Index, 108
- arithmetisches Mittel, 74
- aromatische Bindung, 38
- aromatischer Ring, 38
- Aromatizität
  - strukturelle Restriktion, 52
- Atom, 16
- Atom-Anzahl
  - Bruttoformel-Restriktion, 50
- atomares Profil, 141
- Atomic Mass Unit, 109
- Atommasse
  - mittlere, 262
- Atomtyp, 27
  - Any, 28
  - Element, 28
  - MS, 28
  - Multi, 28
  - Standard, 28
- Atomzustände
  - Bruttoformel-Restriktion, 51
- Atomzustand, 16
  - gültiger, 17
- Ausgangsschicht, 86
- Autokorrelations-Deskriptor, 220
- Automorphismengruppe, 6
- Autoskalierung, 79
- Badlist, 31
  - permanente, 242
- Bahn, 4
- Basiserweiterung, 79
- Basisklasse
  - abstrakte, 27
- Basispeak, 178
- Baum, 10
  - binärer, 88
- Beobachtung, 71
- Bereichsskalierung, 78
- Bestimmtheitsmaß, 74
- Bias, 86
- Bias-Neuron, 85
- Bibliothek
  - kombinatorische, 95
  - reale, 96
  - virtuelle, 96
- bimolekulare Reaktion, 32
- binäre Klassifikation, 73
- binärer Baum, 88
- Bindung
  - aromatische, 38
  - kovalente, 16
- Bindungen
  - Bruttoformel-Restriktion, 51

- Bindungsänderung, 33
- Bindungsänderungsgraph
  - einer chemischen Reaktion, 33
  - eines Reaktionsschemas, 35
- bipartiter Graph, 58
- Boxplot, 236
- Bruttoformel, 23
  - einer chem. Verbindung, 24
  - eines molekularen Graphen, 23
  - kompatible, 50
  - weiche, 50
- Bruttoformel–Restriktion
  - Atom–Anzahl, 50
  - Atomzustände, 51
  - Bindungen, 51
  - DBE, 51
  - Heteroatom–Anzahl, 50
  - Konnektivität, 51
  - Ladung, 51
  - Molekülmasse, 51
  - Radikalstellen, 51
  - Wasserstoff–Verteilung, 51
- Chemie, 19
  - abgeschlossener Schalen, 19
  - eingeschränkte, 19
  - ganzzahlige, 19
  - kombinatorische, 95
- chemische Reaktion, 32
- chemisches Element, 16
- chemisches Reaktionsnetzwerk, 58
- Cholesky–Zerlegung, 84
- Clusteranalyse, 164
- Decan, 118
- Dendrogramm, 164
- Deskriptor
  - Autokorrelations–, 220
  - Ionenserien–, 220
  - molekularer, 108
  - MS–, 218
  - Spektrrentyp–, 221
- Dichte
  - physikalische, 140
  - topologische, 143
- diskrete Struktur, 6
- diskrete Variable, 71
- Diskriminanzfunktion, 73
- Distanz
  - Substruktur–Restriktion, 30
- Diversität, 179
- Doppelbindungsäquivalent, 25
- Doppelnebenklasse, 5
- drittes Quartil, 202
- Edukt, 32
- Eduktgraph, 32
- Eigenschaft
  - strukturelle, 218
- Einbettung, 29
  - als molekulare Substruktur, 31
  - als molekulare Teilstruktur, 31
  - von Graphen, 13
- einfache Regression, 72
- Eingangsschicht, 85
- eingeschränkte Chemie, 19
- Einkomponentenreaktion, 32
- Elektron, 16
  - ungepaartes, 16
- Element
  - chemisches, 16
  - fusionierendes, 5
  - schweres, 28
- empirische Formel, 25
  - einer chem. Verbindung, 26
- empirischer  $F$ –Wert, 75
- endliche weiche Bruttoformel, 50
- Entscheidungsbaum, 88
- Entscheidungsregel, 89
- erstes Quartil, 202
- Erzeugung
  - ordnungstreue, 53



- euklidische Norm, 79
- Expertensystem, 170
- F*-Wert
  - empirischer, 75
- Faltung, 184
- Feature
  - MS-, 220
- Feedforward-Netz, 85
- Fehler erster Art, 159
- Fehler zweiter Art, 159
- Fragment, 175
- Fragmentierung, 175
- freies Elektronenpaar, 16
- Freiheitsgrad, 75
- fusionierende Abbildung, 5
- fusionierendes Element, 5
  
- gültiger Atomzustand, 17
- ganzzahlige Chemie, 19
- generische Strukturformel, 49
- geometrischer Index, 108
- gerichteter Graph, 58
- Gesamt-Klassifikationsfehler, 73
- geschlossener Kantenzug, 9
- geschlossener Subgraph, 12
- Gewicht, 102
- Goodlist, 31
- Grad
  - eines Knotens, 8
- Gradpartition, 8
- Graph, 7
  - bipartiter, 58
  - gerichteter, 58
  - molekularer, 20
  - nicht-nummerierter, 7
  - schlichter, 7
  - zusammenhängender, 10
- graphentheoretische Planarität, 56
- Grenzstruktur
  - mesomere, 37
  
- Grundzustand, 16
  
- Hauptkomponentenanalyse, 164
- Heteroatom, 23
- Heteroatom-Anzahl
  - Bruttoformel-Restriktion, 50
- Hitliste, 169
- hochaufgelöste Isotopenmasse, 261
- Hochdurchsatz-Screening, 95
- Homogenität, 179
- Homomorphieprinzip, 49
- Hybridisierung, 9
  - Substruktur-Restriktion, 30
- Hyperebene
  - separierende, 86
  
- Index
  - arithmetischer, 108
  - geometrischer, 108
  - rein arithmetischer, 109
  - rein topologischer, 110
  - topologischer, 108
- induzierter Subgraph, 12
- induzierter Teilgraph, 12
- Infrarotspektroskopie, 168
- inkonsistente Restriktion, 55
- innerer Knoten, 88
- Intensität, 175
- Intensitätsverhältnis
  - logarithmisches, 220
- Invariante, 6
  - MS-, 220
- inverse QSAR, 69
- inverse QSPR, 69
- inzident, 8
- Ion
  - primäres, 175
  - sekundäres, 175
- Ionenserien-Deskriptor, 220
- Ionisation, 173
- Isomer, 23

- isomorphe Graphen, 7
- isomorphe molekulare Graphen, 22
- Isomorphieklasse
  - molekularer Graphen, 22
  - von Abbildungen, 6
  - von Graphen, 7
- Isotope, 181
- Isotopenmasse, 181
  - hochaufgelöste, 261
- Isotopenmuster, 183
  - theoretisches, 185
- Isotopenverteilung
  - natürliche, 181
- kanonische Form
  - für Graphen, 7
  - für molekulare Graphen, 22
- kanonische Nummerierung, 7
- Kante, 8
- Kantenvielfachheit, 8
- Kantenzug, 9
  - geschlossener, 9
  - offener, 9
- Kernel-Funktion, 87
- Kernresonanzspektroskopie, 168
- Klassengleichung, 4
- Klassifikation, 72
  - binäre, 73
- Klassifikationsbaum, 88
- Klassifikationsfehler
  - Gesamt-, 73
  - mittlerer, 75
- Klassifikator
  - MS-, 218
- Knoten, 8
  - innerer, 88
  - terminaler, 88
  - verbindbare, 10
- Knotendistanzgrad, 111
- Knotenvalenzgrad, 112
- kombinatorische Bibliothek, 95
- kombinatorische Chemie, 95
- kompatible Bruttoformel, 50
- Konfiguration, 48
- Konformation, 48
- Konnektivität
  - Bruttoformel-Restriktion, 51
- konsistente Restriktion, 55
- Konstitution, 47
- Konstitutionsisomer, 23
- kontinuierliche Variable, 71
- Korrelationskoeffizient
  - multipler, 74
- korreliert
  - vollständig, 80
- Kostenfunktion, 72
  - Null-Eins-, 73
- kovalente Bindung, 16
- Kreis, 10
- Kreuzvalidierung, 77
- Kroneckersche Deltafunktion, 73
- Länge
  - eines Kantenzugs, 10
- Ladung, 16
  - Bruttoformel-Restriktion, 51
- Lernen
  - überwachtes, 70
  - unüberwachtes, 164
- Lernkriterium, 55
- Lernsatz, 76
- lexikographische Ordnung, 53
- Ligand, 57
- Linksnebenklasse, 5
- Lorenzkraft, 175
- Makro
  - strukturelle Restriktion, 52
- Markush-Formel, 271
- Masse, 178
- Masse-Ladungs-Verhältnis, 175
- Massenspektrometrie, 168

- Massenspektrum, 178
- Median, 202
- Mehrzentren–Chemie, 19
- mesomere Grenzstruktur, 37
- Minimalitätstest, 55
- Missklassifikationsrate, 75
- Mittel
  - arithmetisches, 74
- mittlere Atommasse, 262
- mittlerer Klassifikationsfehler, 75
- Molekülion, 173
- Molekülmasse, 47
  - Bruttoformel–Restriktion, 51
- molekulare Strukturaufklärung, 69
- molekulare Substruktur, 31
- molekulare Teilstruktur, 31
- molekularer Deskriptor, 108
  - binärer, 117
- molekularer Graph, 20
- MS–Deskriptor, 218
- MS–Feature, 220
- MS–Invariante, 220
- MS–Klassifikator, 218
- Multigraph, 7
- multigraphische Partition, 9
- multiple Regression, 72
- Mycobacterium fortuitum*, 153
- Nachbarschaft
  - Substruktur–Restriktion, 30
- Nachbarschaftsliste
  - reduzierte, 277
  - vollständige, 277
- natürliche Isotopenverteilung, 181
- Neutralteilchen, 175
  - primäres, 175
  - sekundäres, 175
- Neutron, 180
- nicht–nummerierter Graph, 7
- nominale Masse
  - einer Bruttoformel, 186
  - eines Elements, 181
- Norm
  - euklidische, 79
- Null–Eins–Kostenfunktion, 73
- Nummerierung
  - kanonische, 7
- offener Kantenzug, 9
- Oktettregel, 19
- Operation, 3
- Ordnung
  - lexikographische, 53
- ordnungstreue Erzeugung, 53
- Ordnungszahl, 16
- Overfitting, 79
- Partition, 8
  - multigraphische, 9
- Partitionierung
  - rekursive, 88
- Peak, 178
- Peakcluster, 179
- permanente Badlist, 242
- physikalische Dichte, 140
- planare Platzierung, 41
- Planarität
  - graphentheoretische, 56
- Platzierung
  - 2D–, 41
  - 3D–, 41
  - planare, 41
- primäres Ion, 175
- primäres Neutralteilchen, 175
- Produkt, 32
- Produktgraph, 32
- Profil
  - atomares, 141
- Proton, 16
- QR–Zerlegung, 84
- QSAR
  - inverse, 69

- QSPR
  - inverse, 69
- Quantil, 199
- Quartil
  - drittes, 202
  - erstes, 202
- R-Gruppe, 49
- Radikalstelle, 16
- Radikalstellen
  - Bruttoformel-Restriktion, 51
- Rand, 86
- Ranking, 170
  - von Bruttoformeln, 176
  - von Strukturformeln, 176
- Rankingfunktion, 192
- Reaktand, 32
- Reaktion
  - bimolekulare, 32
  - chemische, 32
  - unimolekulare, 32
- Reaktionsänderungsgraph, 32
- Reaktionsgraph, 33
- Reaktionsnetzwerk
  - chemisches, 58
- Reaktionsschema, 35
  - bimolekulares, 35
  - unimolekulares, 35
- Reaktionsstruktur, 35
- Reaktionszentrum, 33
- Reaktionszentrumsgraph, 34
- reaktive Stelle, 57
- reale Bibliothek, 96
- Rechtsnebenklasse, 5
- Regression, 72
  - einfache, 72
  - multiple, 72
- Regressionsbaum, 88
- rein arithmetischer Index, 109
- rein topologischer Index, 110
- Reinsubstanz, 167
- rekursive Partitionierung, 88
- Residuum, 72
- Restriktion, 49
  - inkonsistente, 55
  - konsistente, 55
- Resubstitution, 74
- Ring, 10
  - aromatischer, 38
  - Substruktur-Restriktion, 30
- Ring-Substruktur, 53
- Schicht
  - verborgene, 85
- schlichter Graph, 7
- schweres Element, 28
- Schwerpunkt
  - eines Massenspektrums, 221
- Screening, 95
  - virtuelles, 98
- sekundäres Ion, 175
- sekundäres Neutralteilchen, 175
- Selektion
  - von Bruttoformeln, 176
  - von Strukturformeln, 176
- Semikanonizität, 55
- separierende Hyperebene, 86
- sequentieller Zugriff, 277
- Sims-Kette, 55
- Singulärwert, 84
- Spektreninterpretation, 170
- Spektrensimulation, 170
- Spektrientyp-Deskriptor, 221
- Spektrvergleich, 169, 170
- Spektroskopie, 167
- Standardabweichung, 79
- Standardfehler, 75
- Standardvalenz, 19
- Stoffgemisch, 167
- Struktur
  - diskrete, 6
- Struktur-Verifikation, 170

- Strukturaufklärung
  - molekulare, 69
- strukturelle Eigenschaft, 218
- strukturelle Restriktion
  - Aromatizität, 52
  - Makro, 52
  - Substruktur, 52
  - Symmetrie, 52
- Strukturformel, 22
  - generische, 49
- Strukturgenerierung
  - bruttoformelbasierte, 49
  - reaktionsbasierte, 49
- Strukturraum, 49
- Substruktur–Vielfachheit, 117
- Subgraph, 12
  - geschlossener, 12
  - induzierter, 12
  - mehrdeutiger molekularer, 29
- Substruktur, 27
  - molekulare, 31
  - strukturelle Restriktion, 52
- Substruktur–Restriktion, 30
  - Distanz, 30
  - Hybridisierung, 30
  - Nachbarschaft, 30
  - Ring, 30
- Substruktursuche, 31
- Summe molekularer Graphen, 21
- Summenformel, 23
- Summenformel–Substruktur, 52
- Support–Vektor, 87
- Symmetrie
  - strukturelle Restriktion, 52
- Symmetriefunktion
  - eines Massenspektrums, 221
- Synthesereaktion, 32
- Taillenkreis, 11
- Taillenweite, 11
- Teilgraph, 12
  - induzierter, 12
  - mehrdeutiger molekularer, 29
- Teilmengenrelation
  - für Bruttoformeln, 25
- Teilstruktur
  - molekulare, 31
- terminaler Knoten, 88
- Testsatz, 76
- Teststichprobe, 76
- theoretisches Isotopenmuster, 185
- Tiefe
  - für Reaktanden, 62
  - für Reaktionsschemata, 60
- topologische Dichte, 143
- topologischer Index, 108
- Transversale, 4
- überwachtes Lernen, 70
- Ultraviolett-spektroskopie, 168
- Umlagerungsreaktion, 32
- unüberwachtes Lernen, 164
- unabhängige Variable, 71
- ungepaartes Elektron, 16
- unimolekulare Reaktion, 32
- unkorreliert, 80
- Valenz, 16
- Valenzelektron, 16
- Van der Waals Radius, 114
- Van der Waals Volumen, 45
- Variable
  - abhängige, 71
  - diskrete, 71
  - kontinuierliche, 71
  - unabhängige, 71
- verbindbare Knoten, 10
- verborgene Schicht, 85
- Vergleichswert, 192
  - für Bruttoformeln, 176
  - für Strukturformeln, 176
- Verzweigthheit, 111

- Vielfachheit
  - von Subgraphen, 14
- virtuelle Bibliothek, 96
- virtuelles Screening, 98
- vollständig korreliert, 80
- Vorhersagefähigkeit, 75
- Vorhersagefunktion, 71
- Vorhersagevariable, 71
  
- Wasserstoff-Verteilung
  - Bruttoformel-Restriktion, 51
- Weg, 9
- weiche Bruttoformel, 50
  - endliche, 50
- Wurzelbaum, 88
  
- Zentralkmolekül, 57
- Zentrierung, 78
- Zerfallsreaktion, 32
- Zielvariable, 71
- Zufallszugriff, 277
- Zugriff
  - sequentieller, 277
- zusammenhängender Graph, 10
- Zusammenhangskomponente, 11
  - triviale, 11
- Zustandsänderung, 32
- Zustandsänderungsverteilung
  - einer chemischen Reaktion, 33
  - eines Reaktionsschemas, 35
- Zustandsverteilung, 20
- Zweikomponentenreaktion, 32