Ein virtueller Laborassistent für die chemische Strukturanalyse mittels NMR-Spektren

Dissertation zur Erlangung des Grades einer Doktorin der Ingenieurwissenschaften (Dr.-Ing.)

der Technischen Fakultät der Universität Bielefeld

vorgelegt von

Michaela Hohenner

Dipl.-Inform. Michaela Hohenner
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
email:mhohenne@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung des akademischen Grades Doktor-Ingenieurin (Dr.-Ing.).

Der Technischen Fakultät der Universität Bielefeld am 3. Februar 2006 vorgelegt von Michaela Hohenner, am 13. November 2006 verteidigt und genehmigt.

Gutachter:

Dr. Sven Wachsmuth, Universität Bielefeld

Prof. Dr. Norbert Sewald, Universität Bielefeld

Prüfungsausschuss:

Prof. Dr. Ipke Wachsmuth, Universität Bielefeld

Dr. Sven Wachsmuth, Universität Bielefeld

Prof. Dr. Norbert Sewald, Universität Bielefeld

Prof. Dr. Gerhard Sagerer, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier °° ISO 9706

Danksagung

Ich möchte mich an dieser Stelle bei all denen bedanken, die mich während meiner Promotionszeit begleitet und unterstützt haben.

Mein Dank gilt als erstes Sven Wachsmuth und Gerhard Sagerer für die Betreuung innerhalb der Arbeitsgruppe sowie Norbert Sewald für die Begutachtung der Arbeit. Darüber hinaus geht ein herzlicher Dank an Frank Zöllner, Steffen Neumann, Kerstin Koch, Marko Tscherepanow und alle übrigen Mitglieder der Arbeitsgruppe Angewandte Informatik, die in allen Stadien der Arbeit von technischen Überlegungen über fachliche Gespräche bis zum Stressabbau eine sehr produktive Arbeitsatmosphäre geschaffen haben. Außerdem danke ich Anke Weinberger für ihre freundliche und engagierte Beratung in allen Formalfragen.

Nicht zuletzt gilt mein Dank auch der Deutschen Forschungsgesellschaft und dem Graduiertenkolleg Strukturbildungsprozesse, in dessen Rahmen meine Arbeit gefördert wurde. Ebenfalls danke ich Herrn Walter Maier, der es ermöglichte, dass mir von der BASF AG freundlicherweise die benötigten Spektraldaten zur Verfügung gestellt wurden.

Vor allem aber möchte ich meinem Mann Sascha danken, der mich die gesamte Zeit über unterstützt, mich immer wieder motiviert und mir in allen Höhen und Tiefen beigestanden hat.

Inhaltsverzeichnis

1 Einleitung

Unter *Strukturaufklärung* in der organischen Chemie versteht man das Aufdecken struktureller Eigenschaften organischer Moleküle. Diese Information ist von Interesse, da die Struktur einer Verbindung Einfluß auf ihr chemisches Verhalten sowie auf ihre physikalischen Eigenschaften wie Löslichkeit, Siedepunkt usw. hat und für die biologische bzw. biochemische oder physiologische Relevanz der Verbindung von Bedeutung ist.

Gegenstand der Strukturaufklärung sind dabei Substanzen vielfältigen Ursprungs: Die Untersuchung unbekannter Naturstoffe ist ebenso bedeutsam wie die Identifikation von Nebenprodukten der chemischen Synthese. Und auch beim Verständnis von Reaktionsmechanismen, im Labor genauso wie im lebenden Organismus, können Informationen über die Struktur der beteiligten Substanzen sowie der entstehenden Zwischenprodukte eine entscheidende Rolle spielen. Darüber hinaus sind in vielen Bereichen, von der pharmazeutischen Forschung oder Physiologie über Laborsicherheit und chemische Industrie bis hin zur Ökologie, zahlreiche Szenarien denkbar, in welchen die Aufklärung der Struktur einer unbekannten Verbindung eine Rolle spielt. Ebenso vielfältig sind jedoch die möglichen Ausgangsvoraussetzungen und die Ansprüche der unterschiedlichen Anwendungsbereiche. Jede Problemstellung bringt spezielles Vorwissen und ihre eigenen Anforderungen mit sich.

Dank moderner Methoden und automatisierter Verfahren kann Strukturaufklärung heutzutage vom technischen Standpunkt her als Routineaufgabe angesehen werden. Die wichtigste Methodengattung, die dabei zum Einsatz kommt, ist die NMR-Spektroskopie. Sie hat in den letzten zwei Jahrzehnten eine erhebliche Entwicklung durchlaufen, so daß heutzutage NMR-Spektren mit einem sehr geringen Aufwand an Zeit und Material aufgenommen werden können. Auch ermöglichen Weiterentwicklungen in den Bereichen der einzelnen speziellen Techniken nunmehr Zugang zu umfangreicher und sehr komplexer Information. Mit den immens großen, immer rascher produzierten Datenmengen hat sich somit der Engpaß des Strukturaufklärungsprozesses von der eigentlichen Untersuchung der Probe in den Bereich der Auswertung der dadurch gewonnenen Daten verlagert.

Computerprogramme zur Unterstützung oder Automatisierung des Auswertungsprozesses stellen daher ein wichtiges Ziel in der modernen organischen Chemie dar. Üblicherweise wird dabei heutzutage ein Ansatz gewählt, welcher demselben Grundprinzip folgt, das bereits Ende des 19. Jahrhunderts in den Anfängen der Strukturaufklärung Gültigkeit besaß: Im ersten Schritt werden alle infragekommenden Molekülstrukturen aufgelistet. Dann wird eine geeignete Substanzeigenschaft gewählt, um in einem zweiten Schritt die unbekannte Testsubstanz hinsichtlich dieser Eigenschaft mit den hypothetischen Strukturkandidaten zu vergleichen. Die gewählte Eigenschaft muß sowohl betreffend die unbekannte Substanz experimentell leicht zugänglich sein als auch mit Blick auf die Strukturhypothesen theoretisch vorhergesagt werden können.

Als Grundlage des ersten Schritts, zur Ermittlung der infragekommenden Strukturhypothesen, dient heute in der Regel die Summenformel: Sie gibt an, wie viele Atome welchen chemischen Elements jeweils an der Verbindung beteiligt sind, eine Basisinformation, die routinemäßig ermittelt werden kann. Als geeignete Substanzeigenschaft wird für den zweiten Schritt, die Auswahl der korrekten Strukturhypothese, das NMR-Spektrum herangezogen: Es kann nicht nur dank des technischen Fortschritts ebenfalls routinemäßig aufgenom-

men werden, der direkte Bezug zwischen spektralen und strukturellen Eigenschaften erlaubt außerdem die theoretische Berechnung der zugehörigen NMR-Spektren der im ersten Schritt ermittelten Strukturkandidaten.

Derartige Systeme erzielen trotz der Einfachheit des Ansatzes und nicht zuletzt dank der zunehmenden Rechenleistung heutiger Computer immer präzisere Ergebnisse. Gleichwohl ist zu bemerken, daß der Ansatz auch einige Nachteile in sich birgt: Erstens erfolgt der Abgleich der berechneten NMR-Spektren der Kandidaten mit dem experimentellen Spektrum nicht frei von Unsicherheiten. Die Spektrenaufnahme ist aufgrund der beschränkten Meßgenauigkeit und der Möglichkeit von Meßfehlern mit Unsicherheiten behaftet, und bei der Berechnung der theoretischen Spektren müssen Näherungen herangezogen werden, da eine exakte Berechnung aus quantenphysikalischen Gründen unmöglich ist.

Außerdem müssen sehr viele Vergleiche theoretischer Spektren mit dem experimentellen Spektrum durchgeführt werden, da im ersten Schritt allein aufgrund der Summenformel eine Zahl von Strukturkandidaten generiert wird, die bereits für kleine Moleküle mit nur wenigen Atomen nahezu unüberschaubar groß sein kann. Obwohl es Ansätze gibt, durch Auswertung von Zusatzinformationen, etwa durch Vorgabe bestimmter Strukturelemente, die in der Molekülstruktur vorkommen müssen oder die nicht vorkommen dürfen, die Gesamtmenge infragekommender Strukturen zu begrenzen, erscheint eine andere Idee vielversprechender: Die beiden Schritte der Generierung von Strukturhypothesen und der Auswahl der richtigen Hypothese sollten nicht unabhängig von einander durchgeführt, sondern integriert betrachtet oder mit einander verflochten werden. Insbesondere birgt das experimentelle NMR-Spektrum der unbekannten Substanz aufgrund des Zusammenhangs zwischen spektralen und strukturellen Eigenschaften Informationen in sich, welchen die alleinige Verwendung zu Vergleichszwecken kaum gerecht wird.

Neben diesen Ansatzpunkten für Verbesserungen sollte ein weiterer Gedanke bei der Entwicklung von Computersystemen Eingang finden: Das System dient einem menschlichen Experten als Werkzeug – es sollte daher den Menschen bei dessen Arbeit unterstützen und im besonderen seine Ergebnisse in einer Art und Weise präsentieren, die den Menschen in die Lage versetzt, den gewünschten Nutzen daraus zu ziehen, das heißt so viele Informationen wie möglich daraus zu erfassen.

Zu diesem Zweck erscheint es sinnvoll, sich bei der Entwicklung eines neuen Systems am menschlichen Vorgehen bei der betreffenden Aufgabe zu orientieren. Wenngleich der Ansatz des zweischrittigen Vorgehens heutiger Strukturaufklärungssysteme vom Vorgehen von Chemikern inspiriert wurde, würde man als Mensch jedoch niemals genau so wie diese Programme verfahren: Es wäre völlig unpraktikabel, zuerst alle Strukturen zu notieren, die zur gegebenen Summenformel passen, und diese dann eine nach der anderen mit Hilfe des NMR-Spektrums der unbekannten Substanz zu validieren. Viel naheliegender ist es, sich durch die Auswertung des gegebenen NMR-Spektrums der unbekannten Substanz die darin codierte Strukturinformation bereits zum Zeitpunkt der Ermittlung der Strukturkandidaten zunutze zu machen. Ein System, welches in derselben Weise vorgeht, könnte dabei mit dem Menschen in eine Art sachlichen Dialog eintreten und ihm nicht nur seine Einschätzung darlegen, sondern auch belegen, welche der beobachteten Fakten diese im Detail untermauern. Obwohl ein reales Fachgespräch zwischen Mensch und Computersystem sicherlich in den Bereich der Utopie fällt, so ist doch die Stichhaltigkeit der grundsätzlichen Idee eines solchen „virtuellen Laborassistenten“ nicht von der Hand zu weisen.

Der Zusammenhang zwischen Spektrum und Struktur ist jedoch sehr komplex. Abbildung 1.1 zeigt beispielhaft die Struktur des Koffeinmoleküls und das zugehörige NMR-Spektrum. Jedes Kohlenstoffatom des Moleküls ist im Spektrum durch einen Peak repräsentiert, dessen

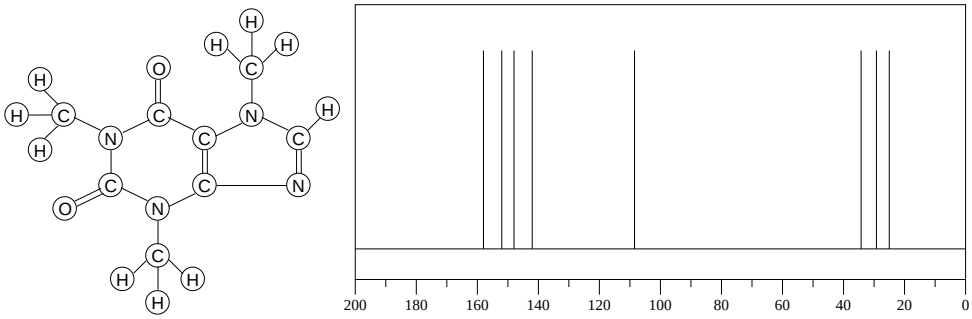


Abb. 1.1: *NMR-Spektrum und Struktur von Koffein (Summenformel $C_8H_{10}N_3O_2$). Jede Linie (Peak) im Spektrum entspricht einem Kohlenstoffatom (C) im Molekül.*

Lage im Zusammenhang mit der Umgebung des betreffenden Atoms im Molekül steht: Was für Bindungen hat es ausgebildet, welche Bindungspartner besitzt es, welche Atome befinden sich in der weiteren Nachbarschaft, und welche Einflüsse treten dadurch in Effekt?

Die Antwort auf diese Fragen ist eine Beschreibung der sogenannten Elektronenhülle des Moleküls. Ihre Gestalt wird in der Hauptsache von den Struktur der Bindungselektronen charakterisiert, die von den Nachbaratomen und der Art der Bindungen zu ihnen abhängt. Sie kann jedoch nicht allein lokal betrachtet werden, da auch entferntere Atome die Elektronendichte in der betreffenden Position beeinflussen. Daher enthält auch der einzelne Peak nicht nur Information über das ihn verursachende Atom und seine Bindungspartner, sondern auch über deren Bindungspartner und noch entferntere Atome; es besteht ein wechselweiser Zusammenhang zwischen den einzelnen Peaks sowie auch der in ihnen codierten Strukturinformation. Bei der Interpretation eines Spektrums sind sie somit immer auch in dessen Gesamtkontext zu betrachten. Bei der Realisierung eines Strukturaufklärungssystems muß dies ebenfalls berücksichtigt werden, wenn das System das menschliche Vorgehen nachahmen und daher eine Interpretation des NMR-Spektrums der unbekannten Substanz Teil der Verarbeitung sein soll.

Aufgrund der nicht nur oberflächlichen, sondern in größerem Umfang konzeptionellen Unterschiede der beschriebenen Idee zu aktuellen Systemen erscheint es am sinnvollsten, ein neues System von Grund auf zu entwickeln, und nicht ein bestehendes in das neue Konzept zu pressen. Zweifellos erfordert dies einiges an Aufwand, die Vorteile sind jedoch ebenso deutlich: Methoden und Systemaufbau können so gewählt werden, daß sie dem beschriebenen Ansatz optimal entsprechen. In der Realisierung können dabei einzelne Aspekte im Detail untersucht und mögliche Varianten mit einander verglichen werden, ohne daß Einflüsse, die aus den Gegebenheiten eines unter anderen Voraussetzungen entwickelten Systems resultieren, die Ergebnisse verschleiern.

Im Zentrum des Interesses steht bei der Entwicklung die Untersuchung des beschriebenen Ansatzes und der zu seiner Realisierung gewählten Methodik. Daher wird zunächst mit der Beschränkung auf aromatische Verbindungen (genauer auf Benzolderivate) eine überschaubare, zugleich aber nicht triviale Teilmenge der immensen Vielfalt organischer Verbindungen betrachtet. Zu erwarten, daß aus dem Nichts ein neues System erschaffen werden könnte, welches allen existierenden Strukturaufklärungssystemen von Anfang an überlegen ist, ist dabei sicherlich nicht realistisch. Vielmehr soll gezeigt werden, daß der stärker an der Herangehensweise des Menschen orientierte Ansatz nicht nur grundsätzlich für die Entwicklung eines Strukturaufklärungssystems geeignet ist, sondern daß auch die in diesem Rahmen

gewählten Methoden und Strategien das Potential besitzen, ein Werkzeug zu schaffen, das die Leistungsfähigkeit existierender Systeme erreichen oder sogar übertreffen kann. Derartige Untersuchungen und Analysen können in einem überschaubaren Rahmen, wie ihn die Betrachtung von Benzolderivaten bietet, besser durchgeführt werden als vor dem sehr weit gefaßten Hintergrund *aller* organischen Verbindungen.

Das größte Gewicht kommt dabei dem expliziten Einbringen von Expertenwissen mit dem Ziel des Nachahmens einer menschlichen Vorgehensweise zu. Dieses Wissen betrifft die Betrachtung von Bindungen in organischen Molekülen, den Einfluß verschiedener Atomtypen auf die Bindungsstruktur und den Effekt all dessen auf die Elektronenverteilung innerhalb des Moleküls, die schließlich über die Lage der Peaks im NMR-Spektrum bestimmt. Statt durch Präsentation von Trainingsbeispielen wird dem System dabei Wissen aus dem Bereich, in dem es eingesetzt werden soll, explizit vorgegeben. Damit wird eine symbolische Ebene geschaffen, die dem System und dem Benutzer gleichermaßen zugänglich ist. Sie soll die Möglichkeit einer gemeinsamen Verständigungsbasis bieten, die das Strukturaufklärungssystem im Wortsinne zu einem „intelligenten Werkzeug“ macht.

Die vorliegende Arbeit dokumentiert die Schritte der Entwicklung sowie die Evaluierung eines solchen Forschungssystems. Zuerst wird hierzu eine Einführung in die Betrachtung von Molekülstrukturen in der organischen Chemie und deren Untersuchung mittels NMR-Spektroskopie benötigt. Diese gibt Kapitel 2 und liefert somit ein Fundament des nötigen Fachwissens und Vokabulars für die weiteren Überlegungen. Im Anschluß gibt Kapitel 3 einen Überblick über existierende Systeme, die in diesem Feld eingesetzten Methoden und Techniken und unterschiedlichen verfolgten Ansätze. Kapitel 4 schließt den ersten Teil der Arbeit mit Grundlagen aus dem Bereich der Informatik, genauer der Mustererkennung, ab.

Beginnend mit einem Überblick über den Aufbau des Systems in Kapitel 5 wird anschließend die Entwicklung im einzelnen dokumentiert, um eine solide Grundlage für dessen Evaluation und potentielle zukünftige Weiterentwicklungen zu bieten. Den Schwerpunkt bildet der Aspekt der expliziten Einbringung von Expertenwissen und damit die Entwicklung eines Kausalmodells der ursächlichen Zusammenhänge der Domäne, die in Kapitel 6 beschrieben wird. Es schließen sich Aspekte der Datenvorverarbeitung in Kapitel 7 an. Statistische Untersuchungen auf der Basis einer Stichprobe von NMR-Spektren und zugehörigen Molekülstrukturen sind Gegenstand von Kapitel 8 und dienen dazu, die kausalen Verknüpfungen innerhalb des Modells zu gewichten und dasselbe somit in ein *Bayes-Netz* zu überführen.

Die Generierung einer Strukturbeschreibung, also des letztlich gewünschten Resultats der Verarbeitung, sowie die Integration aller Systembestandteile zu einem grundsätzlich einsatzfähigen Gesamtsystem ist Gegenstand von Kapitel 9. An diesem können sodann praktische Untersuchungen durchgeführt werden, um die Eignung des Ansatzes im allgemeinen wie auch die des entwickelten Modells im speziellen zu untersuchen. Kapitel 10 beschäftigt sich mit der Evaluation vor dem Hintergrund der oben dargestellten Ideen, wiederum mit dem Schwerpunkt der expliziten Einbringung von Expertenwissen, das heißt der kausalen Modellierung. Die Ergebnisse der gesamten Arbeit werden anschließend in Kapitel 11 zusammengefaßt, und es wird ein Ausblick betreffend des Potentials des gewählten Ansatzes sowie möglicher Weiterentwicklungen gegeben.

2 Grundbegriffe der organischen Chemie

Um eine spezielle Fragestellung zu bearbeiten oder zu diskutieren ist es unerläßlich, grundlegende Begriffe und Definitionen des betreffenden Feldes zu kennen und ein entsprechendes Problemverständnis zu besitzen. Im folgenden sollen daher einige Begriffe aus der organischen Chemie eingeführt werden, die insbesondere für das Anwendungsfeld der Strukturaufklärung von Bedeutung sind.

Die *organische Chemie* ist die Chemie der Kohlenstoffverbindungen, das heißt derjenigen chemischen Verbindungen, deren Grundstruktur aus Kohlenstoffatomen aufgebaut ist. Etwa 800.000 anorganischen (nicht-Kohlenstoff-)Verbindungen stehen etwa 10 Millionen Kohlenstoffverbindungen gegenüber, und insbesondere sind die meisten Verbindungen mit biologischer Relevanz organisch.

Organische Substanzen sind in ihrem Reaktionsverhalten, aber auch in ihren physikalischen Eigenschaften durch die *kovalente Bindung* charakterisiert. Auf diese wird in Abschnitt 2.1 im Rahmen der Betrachtung organisch-chemischer Strukturen und ihrer Besonderheiten eingegangen. *Aromatische Verbindungen*, eine bestimmte Klasse organischer Verbindungen, die sich durch eine besondere Bindungsstruktur und daraus resultierende speziellen Eigenschaften auszeichnet, sind Gegenstand von Abschnitt 2.2. Bindungen und Bindungsstrukturen sind somit in Übereinstimmung mit dem Forschungsgebiet der Strukturaufklärung, dem sich diese Arbeit widmet, der Fokus der Betrachtungen. Abschnitt 2.3 gibt im Anschluß einen Einblick in die Thematik der Strukturaufklärung, wobei der Schwerpunkt auf der gewählten Methode der NMR-Spektroskopie als Informationsquelle liegt. Abschließend formuliert Abschnitt 2.4 vor diesem Hintergrund noch einmal die Ziele der gegenwärtigen Arbeit.

2.1 Organisch-chemische Strukturen

Die organische Chemie ist die Chemie der *Kohlenwasserstoffe* und ihrer *Derivate*. Ein Kohlenwasserstoff ist eine Substanz, die aus den Elementen Kohlenstoff (C) und Wasserstoff (H) aufgebaut ist. Ein Derivat einer organischen Verbindung ist eine Substanz, die durch geringfügige strukturelle Modifikation von der ursprünglichen Verbindung abgeleitet werden kann. Solche Abwandlungen betreffen oft den Einbau von Fremdatomen, sogenannten *Heteroatom*en, also anderen chemischen Elementen als Kohlenstoff und Wasserstoff. Die größte Bedeutung haben dabei Sauerstoff (O), Stickstoff (N), Schwefel (S) und die Halogene Fluor, Chlor, Brom und Iod (F, Cl, Br, I). Seltener kommen auch Metalle oder Phosphor (P) vor. Die Bindungen zwischen den einzelnen Atomen bestimmen die Struktur der Moleküle einer Verbindung, welche sie von den Molekülen anderer Verbindungen unterscheidet.

2.1.1 Die kovalente Bindung

Eine *chemische Bindung* ist eine feste Verknüpfung zweier Atome. Ein Atom besitzt in seinem Kern positiv geladene *Proton*en und ungeladene *Neutron*en. Negativ geladene *Elektron*en umgeben den Kern, so daß das Atom insgesamt neutral erscheint. In einem schalenartigen Aufbau sind äußere und innere Elektronen zu unterscheiden. Die äußeren oder *Valenzelektronen* sind diejenigen Elektronen, die an der Ausbildung von Bindungen betei-

ligt sind. Die inneren Elektronen sind dagegen nicht für eine Interaktion mit dem oder den Bindungspartnern zugänglich.

Die treibende Kraft hinter der Bindungsausbildung ist das Anstreben der sogenannten *Edelgaskonfiguration*, das heißt einer voll besetzten äußersten Elektronenschale. Edelgase sind sehr reaktionsträge chemische Elemente, die von sich aus eine voll besetzte äußerste Schale besitzen und sich somit bereits in einem energetisch günstigen Zustand befinden, welchen Atome anderer Elemente erst durch die Ausbildung von Bindungen anstreben. Man spricht in diesem Zusammenhang auch von der *Oktettregel*[1], da dieser Zustand nach dem *Schalenmodell* (RUTHERFORD-BOHR-SOMMERFELD-Atommodell, vgl. z.B. [Mor01] S. 65/66 sowie [BWe91] S. 18) in der Regel durch acht Elektronen erreicht wird. Eine der Ausnahmen ist Wasserstoff, der bereits mit zwei Elektronen in der äußersten Schale Edelgaskonfiguration erreicht.

Zu diesem Zweck können Atome Elektronen an ihre Bindungspartner abgeben oder Elektronen von ihren Bindungspartnern aufnehmen. Dadurch werden sie zu *Ionen*, das sind geladene Teilchen (Atome oder Moleküle) die nach außen nicht neutral erscheinen, da der eine Bindungspartner einen Überschuß an negativen Ladungen erhält, während der andere durch den Verlust von Elektronen positiv geladen wird. Eine solche Bindung wird *Ionenbindung* genannt und beruht auf der Anziehung der bei der Elektronenabgabe und Aufnahme entstehenden gegensätzlich geladenen Ionen.

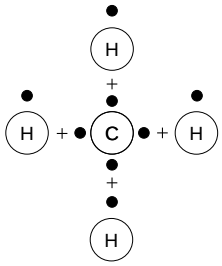
Voraussetzung für die Ionisierung ist eine genügend hohe Differenz der *Elektronegativität*en der Bindungspartner. Diese einheitenlose Größe bezeichnet die Neigung eines chemischen Elements, Valenzelektronen an sich zu ziehen. Sie ist eine Naturkonstante für jedes Element und ist im Allgemeinen um so größer, je mehr Valenzelektronen das betreffende chemische Element besitzt. Kohlenstoff besitzt jedoch mit vier Valenzelektronen im Sinne der Oktettregel eine genau halb besetzte Valenzschale. Somit ist nicht zu entscheiden, ob die Abgabe oder die Aufnahme von vier Elektronen der günstigere Weg zur Edelgaskonfiguration ist.

Tatsächlich ist die Besonderheit der Kohlenwasserstoffe das Ausbilden *kovalenter Bindungen*, das sind Bindungen zwischen chemischen Elementen mit annähernd gleicher Elektronegativität, bei welchen sich die Bindungspartner ein Bindungselektronenpaar gleichberechtigt teilen. Darüber hinaus spielen *polare Bindung*en eine Rolle: Unterscheidet sich die Elektronegativität der Bindungspartner, ohne daß es jedoch zu einer Ionisierung kommt, so wird das Bindungselektronenpaar mehr oder weniger stark zum elektronegativeren Bindungspartner hingezogen, welcher dadurch negativ polarisiert wird. Beide Arten von Bindungen, die beispielhaft in Abbildung 2.1 dargestellt sind, faßt man unter dem Begriff der *Elektronenpaarbindung* zusammen.

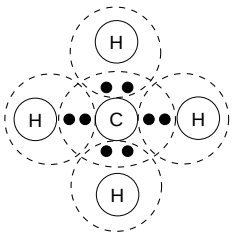
Die Eigenheiten kovalenter bzw. polarer Bindungen betreffen auch die physikalische Eigenschaften des Moleküls, die bei der in der Strukturaufklärung eingesetzten NMR-Spektroskopie[2] eine Rolle spielen: Dabei ist im besonderen nicht nur die Art der Bindung, sondern vor allem die Bindungsstruktur in Gestalt der Elektronenverteilung über das gesamte Molekül betrachtet von Bedeutung. Diese Verteilung läßt sich besser mit Hilfe des quantenmechanischen *Orbitalmodell*s beschreiben, das in Abschnitt 2.1.3 eingehender vorgestellt wird. Zuvor ist es jedoch wichtig, sich mit den Zusammenhängen strukturell ähnlicher Verbindungen vertraut zu machen, um eine systematische Unterscheidung anhand von Strukturcharakteristika zu ermöglichen.

[1] lat. *okta-*: „acht“

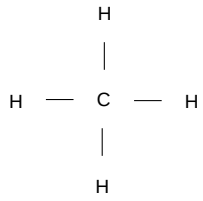
[2] Die NMR-Technik findet in der Medizin eine weitere wichtige, jedoch gänzlich anders ausgerichtete Anwendung, welche mit der Strukturaufklärung nicht vergleichbar ist.



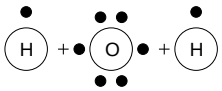
1 Atom Kohlenstoff bildet mit 4 Atomen Wasserstoff Methan.



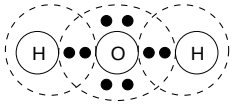
Da die Elektronegativitäten annähernd gleich sind, teilen sich die Bindungsparner die Elektronenpaare gleichberechtigt.



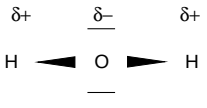
Kovalente Bindungen werden zur Verdeutlichung in der Regel in der Valenzstrichschreibweise notiert.



1 Atom Sauerstoff bildet mit 2 Atomen Wasserstoff Wasser.



Auch hier kommt es nicht zu einer Ionisierung, die Bindungs−partner teilen sich das Elektronenpaar



Die Keilschreibweise deutet die Polarisierung der Bindung an. Der Sauerstoff ist negativ polarisiert.

Abb. 2.1: *Die Bindungen in Methan (CH_4) sind nicht polar: Da sich die Elektronegativitäten von Kohlenstoff und Wasserstoff kaum unterscheiden, sind beide Valenzelektronen mit beiden Bindungspartnern gleich stark assoziiert. Wasser (H_2O) ist demgegenüber mit Wasserstoff und Sauerstoff aus zwei Elementen aufgebaut, deren Elektronegativitäten sich deutlich unterscheiden. Die Bindungen sind hier daher polar, wobei der Sauerstoff eine negative (δ−) und der Wasserstoff jeweils eine positive (δ+) Teilladung trägt.*

2.1.2 Unterscheidung strukturell ähnlicher Verbindungen

In der anorganischen Chemie ist es üblich, für den Aufbau einer Verbindung die *Summenformel* anzugeben. In dieser Formel werden die chemischen Elemente und ihre Multiplizitäten aufgezählt, z.B. ist H_2SO_4, Schwefelsäure, aus zwei Wasserstoffatomen, einem Schwefelatom und vier Sauerstoffatomen aufgebaut. In der organischen Chemie ist die Angabe der Summenformel jedoch in der Regel nicht ausreichend, da die aufgezählten Atome auf mehrere verschiedene Weisen mit einander verknüpft sein können.

*Strukturformel*n enthalten demgegenüber zusätzlich Information über die *Konnektivitäten*, das heißt die Bindungen im Molekül und damit über seine Struktur. Um sie wiederzugeben, wird die *Valenzstrichschreibweise* verwendet, wie bereits in Abbildung 2.1 (rechts oben) gezeigt. Jeweils ein Elektronenpaar wird dabei durch einen Strich dargestellt, Polarisierungen können durch Keile anstelle von Strichen verdeutlicht werden. Je nach Bedarf kommt auch eine verkürzte Schreibweise zur Anwendung, in der die Valenzstriche der Bindungen zu Wasserstoff weggelassen oder Kürzel für häufig vorkommende Gruppen verwendet werden.

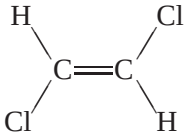
Durch eine solche Notation kann die Verschiedenheit von *Isomere*n erfaßt werden. Isomere sind Verbindungen, die dieselbe Summenformel, jedoch eine unterschiedliche Struktur haben. Trotz der Vorgaben der Summenformel bleiben viele Möglichkeiten offen, wie die Atome mit einander verbunden sein können. Als Beispiel sei Butan gegeben. Es hat die Summenformel C_4H_{10}. Mögliche Isomere sind:

- unverzweigtes Kohlenstoffskelett: $CH_3 - CH_2 - CH_2 - CH_3$
- verzweigtes Kohlenstoffskelett:

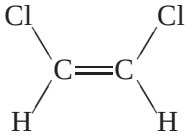
$$CH_3 - \underset{\displaystyle CH_3}{\overset{\displaystyle H}{\overset{|}{\underset{|}{C}}}} - CH_3$$

Die beiden dargestellten Strukturen sind *Konstitutionsisomere*, das heißt sie unterscheiden sich in Lage und Art der Bindungen innerhalb des Moleküls (Konstitution). Darüber hinaus unterscheidet die *Stereochemie*, die sich mit der relativen Stellung der Atome zueinander befaßt, *Konfigurationsisomere* sowie *Konformationsisomere*. Konfigurationsisomere besitzen dieselbe Konstitution, jedoch eine unterschiedliche relative Stellung der einzelnen Gruppen zu einander. Beispielsweise gibt es zwei Konfigurationsisomere des Dichlorethans; sie werden als *Z*- und *E*-Isomer bezeichnet („zueinander“, „entgegen“).

- *E-Dichlorethan*:



- *Z-Dichlorethan*:



Da die Doppelbindung zwischen den beiden Kohlenstoffatomen (anders als eine Einfachbindung) nicht frei drehbar ist, können die beiden genannten Isomere des Dichlorethans nicht durch Drehung in einander überführt werden. Es handelt sich um strukturell unterschiedliche Verbindungen mit unterschiedlichen physikalischen Eigenschaften, die im Zusammenhang physikalischer Methoden der Strukturaufklärung (wie der NMR-Spektroskopie) unterscheidbar sind.

Demgegenüber sind unterschiedliche Konformationen nicht als aufzählbare, unterscheidbare Varianten einer Grundstruktur zu betrachten. Sie bezeichnen vielmehr energetisch unterschiedliche Ausprägungen der relativen Anordnung von Teilen des Gesamtmoleküls. Aufgrund der Drehbarkeit von Bindungen sind hier die Übergänge fließend; anhand bestimmter Charakteristika kann der Gesamtraum möglicher Konformationen jedoch unterteilt werden. Man unterscheidet zum Beispiel die *ekliptische* und die *gestaffelte* Konformation von Seitengruppen, wie in Abbildung 2.2 dargestellt:

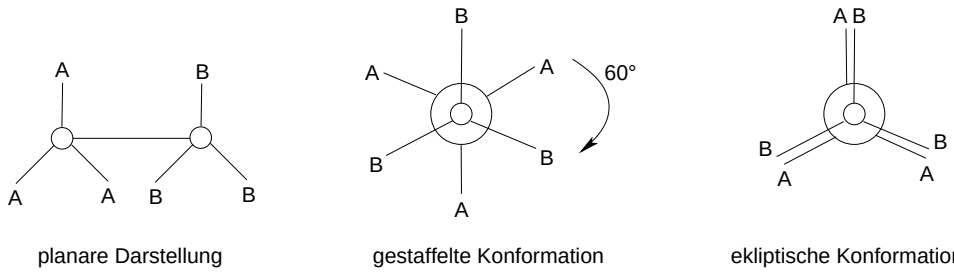


Abb. 2.2: *Ekliptische und gestaffelte Konformation von Seitengruppen sind durch Drehung um die Einfachbindung zwischen den Skelettatomen in einander überführbar.*

Es sind also vier Abstufungen des Detailgrads der Strukturbeschreibung zu unterscheiden:

1. Die Summenformel: Sie enthält lediglich Information über die beteiligten chemischen Elemente und deren Multiplizitäten.
2. Die Konstitution: Sie enthält zusätzlich Information über die Verknüpfung von Atomen und die Art dieser Bindungen.
3. Die Konfiguration: Sie unterscheidet außerdem die relative räumliche Anordnung von Teilstrukturen.
4. Die Konformation: Sie unterscheidet alle energetisch verschiedenen räumlichen Anordnungen der Molekülgestalt.

Neben der reinen Unterscheidung von Isomeren und Konformationen ist es im Zusammenhang mit strukturellen Eigenschaften jedoch ebenso wichtig, sich mit bestimmte Effekte hinsichtlich der Valenzelektronen zu beschäftigen. *Tautomerie* und *Mesomerie* sind solche Elektroneneffekte, die in der organischen Chemie eine wichtige Rolle spielen.

Tautomerie bezeichnet eine Umlagerung innerhalb des Moleküls, welche in Strukturformeln durch Verschieben von Elektronenpaaren dargestellt werden kann. Die dadurch unterscheidbaren *Tautomere* sind Isomere, die in einer Substanzprobe der betreffenden Verbindung vorliegen, jedoch in der Praxis nicht von einander getrennt werden können, da die Umlagerung spontan und unter sehr geringem Energieaufwand stattfindet (vgl. [BWe91] S. 294 ff.). Die Umgebungsbedingungen beeinflussen jedoch das Verhältnis, in welchem die einzelnen Tautomere vorliegen. Kommt eine Substanz beispielsweise mit einem Lösemittel in Kontakt, so kann ein Tautomer begünstigt werden, das unter normalen Umständen nur in einem sehr geringen Anteil vorkommt, und aufgrund dessen kann die Substanz in Lösung andere Eigenschaften zeigen als erwartet. Dies spielt auch bei der Aufnahme von NMR-Spektren gelöster Feststoffe eine Rolle: Ein „geeignetes“ Lösemittel sollte die Struktureigenschaften der untersuchten Substanz möglichst nicht beeinträchtigen.

Obwohl sich auch Mesomerie und Mesomerieeffekte bildhaft durch ein Verschieben von Elektronenpaaren in der Strukturformel verdeutlichen lassen, liegen in diesem Fall jedoch *keine* wirklichen, unterschiedlichen Isomere vor. Mesomerie bezeichnet vielmehr das Phänomen, daß die Bindungsverhältnisse einer Struktur nicht eindeutig angegeben, sondern nur durch mehrere sogenannte *mesomere Grenzformeln* umschrieben werden können. Wie der Begriff nahelegt[3], liegt die tatsächliche Elektronenverteilung jedoch *zwischen* diesen Grenzformeln und wird durch keine einzelne von ihnen komplett erfaßt. *Mesomeriestabilisierung* bezeichnet die Tatsache, daß der tatsächliche Elektronenzustand energetisch günstiger (stabiler) ist, als es jede einzelne der durch die Grenzformeln beschriebenen Strukturen wäre. Ein Beispiel für eine mesomeriestabilisierte Verbindung ist Benzol, das in Abschnitt 2.2 als klassischer Vertreter der Aromaten betrachtet wird.

Das Schalenmodell und die Valenzstrichschreibweise zur Beschreibung der (Bindungs-) Strukturen reichen zwar für die Beschreibung und Unterscheidung von Molekülstrukturen aus, zur Beschreibung der Struktur bzw. der Verteilung der Valenzelektronen ist jedoch das quantenmechanische *Orbitalmodell* besser geeignet. Es wird im folgenden eingeführt, da die Elektronenverteilung innerhalb des Moleküls bei den der NMR-Spektroskopie zugrundeliegenden Prinzipien von entscheidender Bedeutung ist.

2.1.3 Orbitaltheorie

Die zum Bereich der Quantenmechanik gehörige *Orbitaltheorie* bietet ein Modell, um die Bindungs- und Elektronenstruktur organischer Moleküle geeignet zu beschreiben, so daß ihre physikalischen Eigenschaften verständlich werden, welche für die Strukturaufklärung mittels NMR-Spektroskopie von Bedeutung sind. Es würde jedoch den Rahmen dieser Arbeit sprengen, im Detail auf die quantenmechanischen Grundlagen einzugehen; es sei daher auf die Fachliteratur (z.B. zur Einführung [Atk96], Kapitel 11–14) verwiesen. An dieser Stelle wird hauptsächlich auf ein grundlegendes Verständnis des *Orbital*-Begriffs wertgelegt.

Ausgangspunkt hierfür ist die sogenannte *Teilchen-Welle-Dualität*, das ist die Feststellung, daß kleine bewegte Objekte Welleneigenschaften besitzen, welche von verschiedenen Wissenschaftlern[4] aufgrund ihrer Beobachtungen in unterschiedlichen Experimenten gemacht

[3] griech. *meso*-: „Mitte“

[4] Davisson & Germer (1925), Thomson (1925), de Broglie (1924) – vgl. [Atk96]

wurde. Auf mikroskopischer Ebene können also bewegte Massepunkte, z.B. Elektronen, nicht nur als Teilchen, sondern auch durch eine (komplexe) Wellenfunktion Ψ beschrieben werden. Diese entspricht einer Wahrscheinlichkeitsamplitude, und ihr Quadrat $\Psi^*\Psi$ beschreibt die Wahrscheinlichkeit für das durch sie beschriebene System, sich zu einem bestimmten Zeitpunkt an einem bestimmten Ort zu befinden. Da diese Aufenthaltswahrscheinlichkeit die maximal verfügbare Information über das System im Raum ist, sagt man auch, die Wahrscheinlichkeitsdichte *ist* das System.

In der theoretischen Chemie wird diese Wahrscheinlichkeitsdichte Orbital genannt; man kann also sagen, das Orbital *ist* das Elektron. Üblicher ist jedoch die Formulierung, daß Elektronen bestimmte Orbitale *besetzen*. Durch Lösen der SCHRÖDINGER-Gleichung für ein Elektron in der Nähe eines positiv geladenen Atomkerns kann die Wellenfunktion Ψ und damit das Orbital exakt berechnet werden. Eine analytische Lösung ist nur für Ein-Elektronen-Systeme möglich, jedoch können numerisch auch Wellenfunktionen und deren zugehörige Energien für Systeme mit mehreren Elektronen berechnet werden. Jedes Orbital wird eindeutig durch Haupt-, Neben- und Magnetquantenzahl identifiziert (vgl. [Atk96] S. 399 ff.). Eine *Quantenzahl* ist ein Index, welcher mögliche Zustände von Systemen numeriert.

Jedes Atom besitzt grundsätzlich dieselben (unendlich vielen) Orbitale. Bei der Frage, welche davon abhängig von der Zahl der Elektronen besetzt sind, ist das zu jedem Orbital korrespondierende (diskrete) Energieniveau von Bedeutung. Elektronen besetzen stets das energetisch günstigste (niedrigste) noch nicht voll besetzte Orbital. Jedes Orbital kann dabei nicht nur ein, sondern zwei Elektronen aufnehmen, da Elektronen einen Spin (Drehimpuls) besitzen: Zwei durch dasselbe Orbital beschriebene Elektronen sind bei gegensätzlichen Spins dennoch unterscheidbar. *Gepaarte Elektronen*, das heißt mit zwei Elektronen voll besetzte Orbitale, anzustreben ist eine Triebkraft der Ausbildung von Bindungen. Sie sind aufgrund der Kompensationseffekte der gegensätzlichen Spins besonders günstig zu bewerten.

Für die Betrachtungen dieser Arbeit sind in erster Linie die für die Bindungsbildung relevanten Valenzelektronen interessant, da ihre Verteilung, im Gegensatz zu der der inneren Elektronen, die nicht an Bindungen beteiligt sind, sich mit der Molekülstruktur ändert. Dies sind für Kohlenstoff die Orbitale mit der Hauptquantenzahl 2, welche der zweiten Schale nach dem Schalenmodell entsprechen. Nach ihrer Form werden sie mit $2s$, $2p_x$, $2p_y$ und $2p_z$ bezeichnet. Ihre Gestalt ist schematisch in Abbildung 2.3 dargestellt: *s-Orbitale* sind kugelförmig, *p-Orbitale* sind hantelförmig mit einer Knotenebene in der Mitte. Es ist bei letzteren zu beachten, daß die beiden mit entgegengesetzten Vorzeichen versehenen Abschnitte beiderseits der Knotenebene *gemeinsam* das Orbital bilden, und daß drei verschiedene *p*-Orbitale nach ihrer Raumrichtung (*x*, *y* und *z*) zu unterscheiden sind. Energetisch liegen die drei $2p$-Orbitale oberhalb des $2s$-Orbitals (vgl.[Atk96] S. 399 ff., [BWe91] S. 18 ff.).

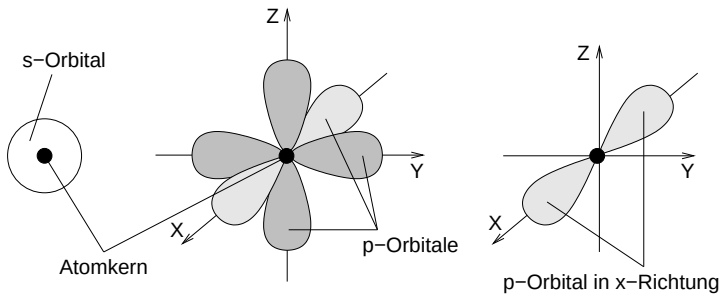


Abb. 2.3: *Grundsätzlicher Charakter von s- und p-Atomorbitalen. Das s-Orbital liegt immer auf einem niedrigeren Energieniveau als die p-Orbitale derselben Schale.*

Das einzige Elektron des Wasserstoffs besetzt das 1*s*-Orbital. Die Valenzelektronen des Kohlenstoffs besetzen dessen Orbitale der Hauptquantenzahl 2 vom untersten (günstigsten) Energieniveau nach oben. Das 2*s*-Orbital wird von zwei Elektronen besetzt, die verbleibenden beiden Elektronen besetzen je eines der 2*p*-Orbitale. Da die drei 2*p*-Orbitale alle auf demselben Energieniveau liegen, wird keines von ihnen bevorzugt, indem es mit einem zweiten Elektron besetzt wird.

Neben *Atomorbitale*n, die von Elektronen besetzt sind, welche klar mit einem bestimmten Atom assoziiert sind, gibt es auch *Molekülorbitale*. Sie entstehen durch Linearkombination von Atomorbitalen (LCAO, *linear combination of atomic orbitals*) und besitzen ein niedrigeres Energieniveau als die Atomorbitale im einzelnen („bindendes Molekülorbital“). Sie werden mit griechischen Buchstaben bezeichnet, die sich ebenfalls auf die Orbitalgestalt beziehen: Ein σ-*Orbital* bewahrt in etwa den Charakter von *s*-Atomorbitalen, ein π-*Orbital* hat demgegenüber einen Bezug zu den *p*-Orbitalen. Da das Bindungselektronenpaar einer kovalenten oder einer polaren Bindung zu beiden Bindungspartnern gehört, besetzt es ein solches Molekül- und kein Atomorbital.

Da bei der Ausbildung von Bindungen neben der Absenkung des Energieniveaus (energetisch günstigerer Zustand) auch das Anstreben gepaarter Elektronen eine Rolle spielt, beteiligten sich nur ungepaarte Elektronen an der Bindungsbildung, um anschließend paarweise die durch LCAO entstehenden Molekülorbitale zu besetzen. Da Kohlenstoff nur zwei ungepaarte Elektronen besitzt, wäre zu erwarten, daß er mit Wasserstoff zu CH_2 reagiert. Dadurch erreicht das Kohlenstoffatom jedoch nicht die Edelgaskonfiguration mit 8 Elektronen in der 2. Schale. In der Tat reagieren Kohlenstoff und Wasserstoff zu CH_4 (Methan), so daß alle beteiligten Atome Edelgaskonfiguration erreichen. Dazu müssen aber beim Kohlenstoff vier ungepaarte Elektronen vorliegen.

Dieser Zustand wird erreicht, indem eines der beiden 2*s*-Elektronen in das (energetisch höhere) noch unbesetzte 2*p*-Orbital angehoben wird. Die für diese sogenannte *Promotion* benötigte Energie ist vergleichsweise niedrig und wird durch den Energiegewinn bei der Bindungsbildung überkompensiert. Darüber hinaus findet eine *Hybridisierung* der Orbitale statt, das heißt eine intraatomare Linearkombination der *s*- und *p*-Orbitale, welche auf diese Weise „miteinander vermischt werden“ (vgl. [Atk96] S. 438 ff.), so daß vier gleichartige, energetisch äquivalente Orbitale entstehen. Die Hybridisierung erfordert keinen zusätzlichen Energieaufwand. Da nun vier gleichartige Atomorbitale für die Linearkombination mit 1*s* zur Verfügung stehen, können auch vier gleichartige Molekülorbitale bzw. Bindungen entstehen.

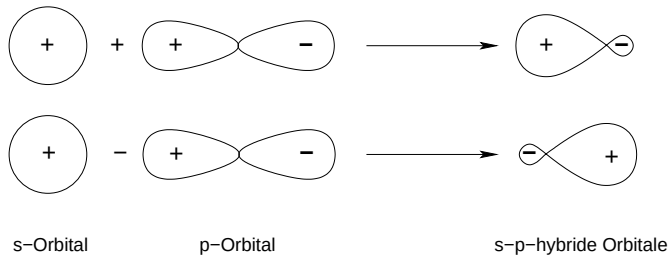
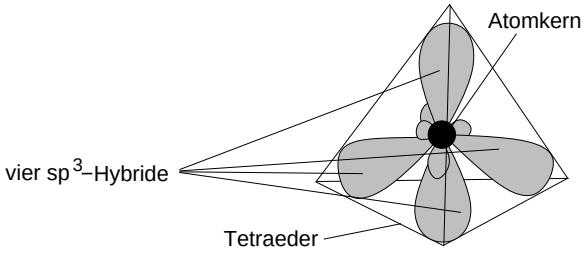


Abb. 2.4: *Hybridisierung von s- und p-Orbitalen im Zweidimensionalen: Durch die beiden möglichen Linearkombinationen der Orbitale entstehen zwei hybride Orbitale.*

In Abbildung 2.4 wird das Zustandekommen der Form von *s*-*p*-Hybridorbitalen im Zweidimensionalen dargestellt. Ihre Hantelform ist anders als bei den *p*-Orbitalen nicht zur Knotenebene symmetrisch. Im Dreidimensionalen ist darüber hinaus zu beachten, daß sowohl p_x als auch p_y und p_z in das resultierende sp^3-Orbital eingehen. Dadurch hat es einen stärkeren

p-Charakter (schlanke Hantelform), und auch seine räumliche Orientierung wird hierdurch bestimmt. Abbildung 2.5 zeigt die entstehende tetraedrische Anordnung der vier durch Linearkombination der *s*- und *p*-Orbitale entstehenden sp^3-Hybride.



Linearkombinationen:

$$s + p_x + p_y + p_z$$
$$s - p_x - p_y + p_z$$
$$s - p_x + p_y - p_z$$
$$s + p_x - p_y - p_z$$

Abb. 2.5: *Die vier sp^3-Hybridorbitale des Kohlenstoffs entstehen durch die nebenstehend aufgeführten vier möglichen Kombinationen der s- und p-Orbitale. Die sp^3-Hybride nehmen eine tetraedrische Anordnung an, in ihrer räumlichen Mitte befindet sich der Atomkern.*

Physikalische Untersuchungen zeigen, daß die vier C-H-Bindungen im Methan (CH_4) in der Tat alle äquivalent sind (vgl.[BWe91] S. 24 ff.), was für die sp^3-Hybridisierung der Atomorbitale spricht. Andernfalls müßte sich die Bindung, an welcher das 2*s*-Orbital beteiligt ist, von denen mit Beteiligung der 2*p*-Orbitale unterscheiden. Auch wegen der räumlichen Anordnung der sp^3-Hybridorbitale ist eine solche Elektronenstruktur besonders günstig: Durch den Tetraederwinkel der Bindungen nehmen die verschiedenen Bindungspartner den größtmöglichen Abstand voneinander ein, so daß sie sich so wenig wie möglich behindern.

Es kommt jedoch nur dann zu einer sp^3-Hybridisierung, wenn das betreffende Kohlenstoffatom, wie z.B. im Falle von Methan, vier äquivalente σ-Molekülorbitale (σ-Bindungen) bildet, das heißt wenn es ausschließlich *Einfachbindungen* besitzt. Besitzt das Atom dagegen nur drei Bindungspartner, so werden nur drei äquivalente sp^2-Orbitale benötigt, das verbleibende *p*-Orbital nimmt an der Hybridisierung nicht teil. Es steht senkrecht auf der Bindungsebene, in welcher die drei sp^2-Orbitale durch ihre trigonale Anordnung wiederum den räumlichen Abstand der verschiedenen Bindungspartner optimieren (vgl. Abbildung 2.6).

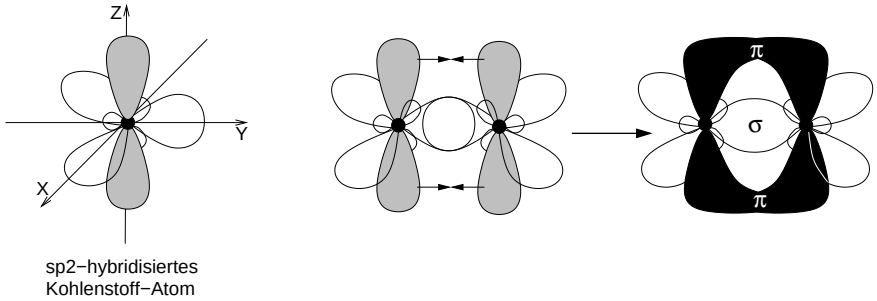


Abb. 2.6: *Die Doppelbindung zwischen zwei Kohlenstoffatomen besteht aus einem σ-Anteil, der durch Linearkombination von zwei sp^2-Hybriden gebildet wird, und einem π-Anteil. Er rührt von zwei p-Orbitalen her; wie auch bei diesen bilden oberer und unterer Abschnitt* zusammen *das Orbital.*

Auf diese Weise wird jedoch keine Edelgaskonfiguration des Kohlenstoffs erreicht. Da die ungepaarten Elektronen in den nicht hybridisierten *p*-Orbitalen der beiden Bindungspartner ebenfalls einen paarweisen Zustand anstreben, findet auch hier eine Linearkombination statt. Dadurch entsteht ein π-Orbital, welches die beiden einsamen *p*-Elektronen als Paar besetzen, also eine zweite Bindung parallel zu einer der σ-Bindungen. Auch dies ist Abbildung

2.6 zu entnehmen. In ähnlicher Weise, wie eine solche *Doppelbindung* entsteht, können *sp*-hybridisierte Kohlenstoffatome durch Ausbildung zweier aus unhybridisierten *p*-Orbitalen hervorgehender π-Bindungen *Dreifachbindung*en ausbilden.

Betreffend aromatische Verbindungen ist zudem ein Sonderfall zu betrachten, und zwar, daß *konjugierte Doppelbindungen* vorliegen, das heißt eine alternierenden Anordnung von Doppel- und Einfachbindungen. In dieser Situation besitzt jedes Atom nur drei Bindungspartner, ist also sp^2-hybridisiert. Da die σ-Bindungen jeweils äquivalent sind und die Durchmischung der unhybridisierten *p*-Orbitale, wie in Abbildung 2.7 dargestellt, in beiden Richtungen entlang des Kohlenstoffskeletts stattfinden kann, sind die Doppelbindungen nicht eindeutig zu lokalisieren. Daher wird dies als *delokalisiertes π-Elektronensystem* bezeichnet.

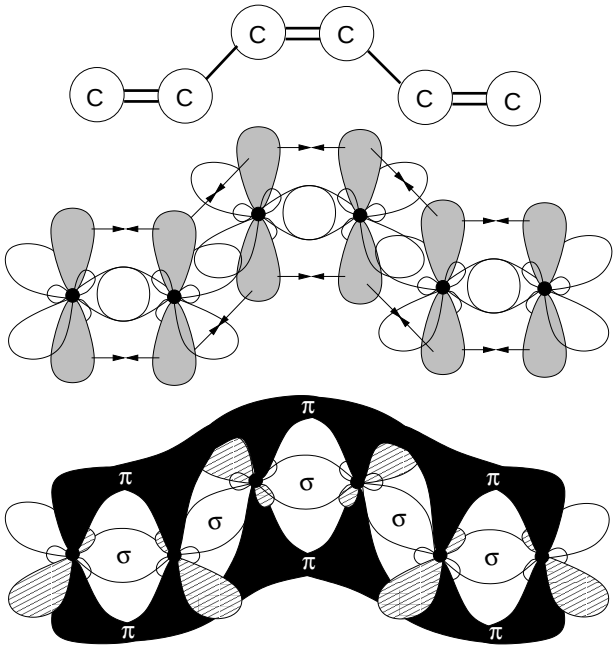


Abb. 2.7: *Im Falle konjugierter Doppelbindungen (oben) entsteht durch Mischung der p-Orbitale (Mitte) ein weitläufiges π-Elektronensystem (unten).*

Um eine π-Bindung auszubilden müssen die beteiligten *p*-Orbitale dieselbe räumliche Orientierung haben; für die Ausbildung eines π-Elektronensystems muß dies für *alle* beteiligten Orbitale gelten. Aufgrund der freien Drehbarkeit der σ-Bindungen geht diese Orientierung jedoch in kettenförmigen Verbindungen leicht verloren. Die Doppelbindung ist dann dort lokalisiert, wo die Orientierung der Bindungspartner dies zuläßt. Im Falle zyklischer Verbindungen (z.B. Benzol) allerdings ist die Konformation derart stabilisiert, daß die freie Drehbarkeit der σ-Bindungen aufgehoben ist, so daß sich ein π-Elektronensystem ausbilden kann, welches sich über den gesamten Ring erstreckt. Es wird auch *aromatisches System* genannt.

Abschnitt 2.2 beschäftigt sich eingehend mit *Aromaten*, das heißt Verbindungen, die durch ein aromatisches System charakterisiert werden. Zuvor sollen jedoch die Einflüsse von Heteroatomen auf die Elektronenstruktur betrachtet und Grundgedanken zur Kategorisierung organischer Verbindungen anhand struktureller Merkmale erläutert werden, um den allgemeinen Überblick der Strukturbetrachtung organischer Moleküle abzuschließen.

2.1.4 Funktionelle Gruppen

Wenngleich sich die Elektronegativitäten von Kohlenstoff und Wasserstoff kaum von einander unterscheiden, so unterscheiden sie sich doch deutlich von der etwaiger Heteroatome im Molekül. Diese bewirken somit eine Polarisierung der betreffenden Bindungen, wodurch sich die entsprechenden Positionen von der Umgebung abheben. Sie wirken bezüglich chemischer Eigenschaften (z.B. des Reaktionsverhaltens) wie Markierungen im Molekül und sind dadurch oft mit einer bestimmten Funktionalität assoziiert (*funktionelle Gruppen*). In vielen Fällen spiegelt die Benennung einer organischen Substanz das Vorhandensein derartiger Gruppen wider.

Bei der Kategorisierung organischer Verbindungen in Stoffgruppen spielt gleichwohl die Sichtweise des jeweiligen Anwendungsfeldes eine große Rolle: Im Kontext der Laborsynthese sind beispielsweise andere Aspekte von Interesse als für biologische Fragestellungen. Da sich diese Arbeit mit Strukturaufklärung beschäftigt, werden im folgenden Einordnungen anhand struktureller Merkmale (im Gegensatz zu chemischen, biologischen oder physikalischen Eigenschaften) vorgenommen.

Statt funktioneller Gruppen spricht man in diesem Zusammenhang eher von *Substituenten*, da die betreffenden Atome Wasserstoff in der Ausgangsverbindung ersetzen[5]. Für die Strukturaufklärung durch NMR-Spektroskopie sind insbesondere die Einflüsse der Substituenten auf die Elektronenstruktur bzw. die Elektronendichte im Molekül interessant: Durch diese beeinflussen sie quantenphysikalische Eigenschaften, die für die NMR-Spektroskopie von Bedeutung sind (vgl. Abschnitt 2.3). Man kann im wesentlichen zwei Arten von Einflüssen unterscheiden: *mesomere Effekte* und *induktive Effekte*.

Induktive Substituenteneffekte beruhen auf der unterschiedlichen Elektronegativität von Kohlenstoff und Heteroatom, durch welche Bindungen polarisiert werden. Die entstehende partielle Ladung an den Bindungspartnern ruft elektrostatische Feldwirkungen hervor, welche eine Veränderung der Elektronendichte in der direkten, aber auch in der weiteren Umgebung der polaren Bindung verursachen. Da die Auswirkungen nicht ausschließlich lokal sind, haben induktive Effekte eine nicht unerhebliche Bedeutung, im besonderen für die NMR-Spektroskopie, da hier die Elektronendichte eine entscheidende Rolle spielt.

Die in der klassischen organischen Chemie am häufigsten vorkommenden Heteroatome haben eine höhere Elektronegativität als Kohlenstoff: Die Elektronendichte ist in der Umgebung des Heteroatoms erhöht; es selbst wird negativ, Kohlenstoff positiv polarisiert. Man spricht hierbei von einem *negativen induktiven Effekt* oder $-I$*-Effekt*. Der entgegengesetzte $+I$*-Effekt* wird in erster Linie von Metallatomen verursacht. Darüber hinaus gibt es jedoch einige Sonderfälle. Insbesondere besitzen reine Kohlenwasserstoffgruppen im Zusammenhang mit den für diese Arbeit wichtigen Aromaten einen $+I$-Effekt.

Der zweite Typ von Einflüssen, die mesomeren Effekte, kann zwar in nichtaromatischen Verbindungen meist vernachlässigt werden, spielt aber an aromatischen Systemen eine bedeutendere Rolle. Voraussetzung ist ein freies Elektronenpaar (wie z.B. beim Stickstoff oder Sauerstoff) in einem an das aromatische System gebundenen Heteroatom, oder Elektronen in Doppelbindungen im betreffenden Substituenten. Interferieren solche Elektronen mit dem aromatischen System, indem sie in die mesomeren Grenzformeln der Verbindung Eingang finden (vgl. Abschnitt 2.1.2) so spricht man von einem mesomeren Effekt. Wie der induktive Effekt verändert auch der mesomere Effekt die Elektronendichte und läßt Partialladungen an einzelnen Positionen entstehen. Mit Hilfe der Valenzstrichschreibweise kann man sich dies durch „Umklappen“ von Elektronenpaaren veranschaulichen.

[5] lat. *substituere*: „ersetzen“

Neben der Bedeutung derartiger Effekte und wegen ihres darauf beruhenden Einflusses auf das chemische Verhalten spielen funktionelle Gruppen, wie oben erwähnt, auch bei der Einteilung organischer Verbindungen in Substanzklassen eine Rolle. Es handelt sich bei diesen Klassen um Oberbegriffe, unter welchen Substanzen mit ähnlichen Eigenschaften zusammengefaßt werden, jedoch nicht notwendigerweise Klassen im Sinne disjunkter Kategorien. Einige grundlegende Substanzklassen, welche insbesondere auch unter strukturellen Gesichtspunkten interessant sind, sollen hier kurz beschrieben werden.

Heterosubstituierte Kohlenwasserstoffe sind anders als *reine Kohlenwasserstoffe* solche Verbindungen, die Heteroatome enthalten. Sie können nach den enthaltenen Elementen und deren Einbindung in das Kohlenwasserstoffgerüst mehrfach weiter unterteilt werden. *Halogenierte Verbindungen* beispielsweise sind solche, in denen Halogene (Fluor, Chlor, Brom und Iod) ein oder mehrere Wasserstoffatome der Ausgangsverbindung ersetzen, wie in Abbildung 2.8 dargestellt. Dies ist vor allem oftmals für das Reaktionsverhalten der Substanz interessant.

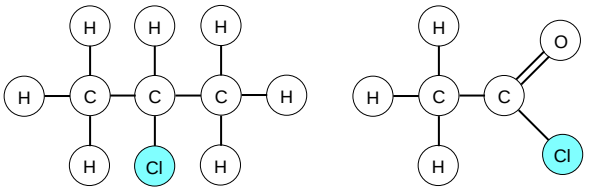


Abb. 2.8: *Beispiele halogenierter Verbindungen, links 2-Chlorpropan, ein Halogenalkan, rechts Essigsäurechlorid, ein Carbonsäurehalogenid*

Größer ist die Vielfalt der Substanzklassen bei anderen Heteroatomen, welche mehr als einbindig sind. Sauerstoff als zweibindiges Atom etwa kann zum einen als Bindeglied zwischen dem Kohlenwasserstoffgerüst und einer mehratomigen Endgruppe fungieren (z.B. *Alkohole*, *Cyanate*, vgl. Abbildung 2.9). Zum anderen kann er aber auch als Bindeglied zwischen zwei Teilen des Kohlenstoffgrundgerüsts der Verbindung auftreten (*Ether*, *Ester*). Schließlich besteht noch die Möglichkeit einer Doppelbindung zu einem Kohlenstoffatom (Ester, *Carbonsäuren*, *Amide*) und der Beteiligung an aromatischen Systemen (z.B. in dem in Abbildung 2.10 dargestellten Furan).

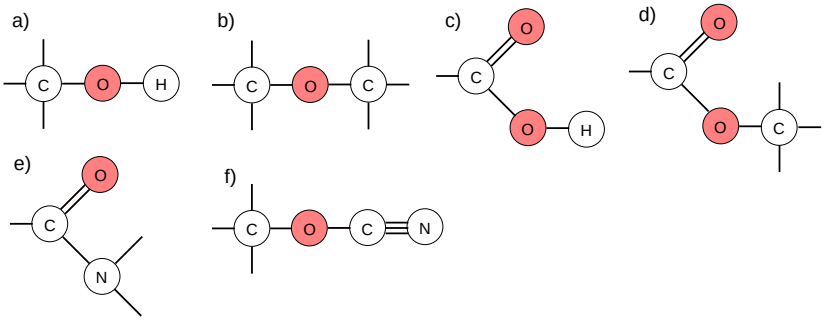


Abb. 2.9: *Beispiele sauerstoffhaltiger Verbindungen: a) Alkohole b) Ether c) Carbonsäuren d) Ester e) Amide f) Cyanate*

Schwefel verhält sich in vieler Hinsicht ähnlich wie Sauerstoff, besitzt aber eine leicht niedrigere Elektronegativität. Sein Vorkommen wird durch die Silbe *thio-* bezeichnet (z.B. Thioether, Thiocyanate, Thioalkohole). Gegenüber anderen Heteroatomen spielt er eine we-

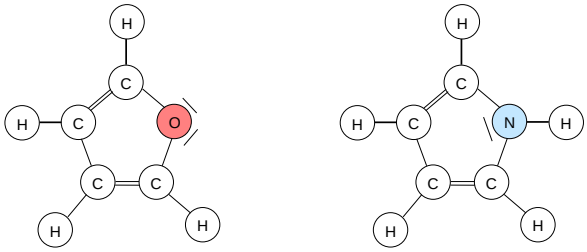


Abb. 2.10: *Furan (links) und Pyrrol (rechts) sind heteroaromatische Verbindungen: Über ihre freien Elektronenpaare sind die Heteroatome am aromatischen System beteiligt.*

niger wichtige Rolle und ist vorwiegend in sehr spezialisierten Teilgebieten von Interesse. Sein Hauptcharakteristikum ist die Möglichkeit zur *Valenzschalenerweiterung* (vgl. [Mor01] S. 126). Details sind der speziellen Fachliteratur (z.B. [VVD$^+$87]) zu entnehmen. Es genügt an dieser Stelle zu wissen, daß Schwefel dadurch mehr als die üblichen acht Valenzelektronen und somit mehr als vier kovalente Bindungen (in der Regel bis zu sechs) besitzen kann.

Beispiele stickstoffhaltiger Gruppen sind Abbildung 2.11 zu entnehmen. Der dreibindige Stickstoff kann verzweigend, nicht verzweigend oder endständig innerhalb des Kohlenstoffskeletts einer organischen Verbindung auftreten (*Amine*) und über sein freies Elektronenpaar auch an aromatischen Systemen beteiligt sein (z.B. Pyrrol in Abbildung 2.10). Ähnlich dem Kohlenstoff kann er Mehrfachbindungen ausbilden (*Imine*, *Cyanide*) und in verschiedenen mehratomigen funktionellen Gruppen im Inneren oder am Ende der Kohlenstoffkette vorkommen (Amide, Cyanate, *Azoverbindungen*, *Nitroverbindungen*, *Nitrosoverbindungen*). Für die Nitro-Gruppe (–NO_2) sind hier mesomere Grenzformeln angegeben. Durch eine komplexe Durchmischung der freien *p*-Orbitale aller drei Atome entsteht ein Elektronenzustand, welcher sich in der Valenzstrichschreibweise nicht wiedergeben läßt.

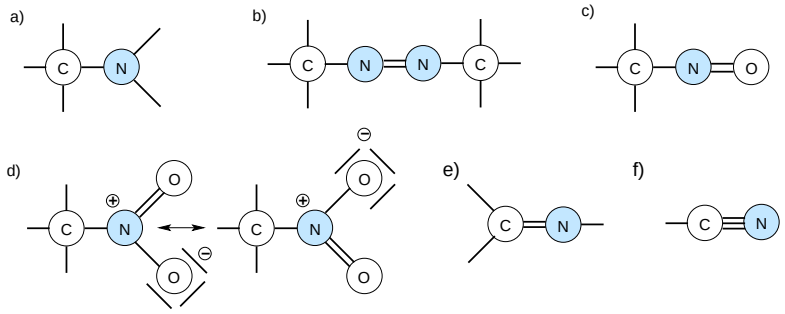


Abb. 2.11: *Beispiele stickstoffhaltiger Verbindungen: a) Amine b) Azoverbindungen c) Nitrosoverbindungen d) Nitroverbindungen e) Imine f) Cyanide*

Auch einige *Ester anorganischer Säuren* spielen in der Organik eine Rolle. Die betreffenden Säurereste können als endständige Substituenten an ein organisches Molekül gebunden sein (vgl. Abbildung 2.12) oder als mehratomige, verzweigende Brücke mehrere organische Teilstrukturen mit einander verbinden.

Legt man weniger Gewicht auf das Vorhandensein von Heteroatomen oder bestimmten funktionellen Gruppen, so können organische Verbindungen auch hinsichtlich ihres Kohlenstoffskeletts unterteilt werden. *Gesättigte Verbindungen* besitzen ausschließlich Einfachbindungen und heißen *Alkane*, sofern sie keine Heteroatome enthalten; Substituenten, die Alkanen gleichen, werden *Alkylreste* genannt und abgekürzt mit R bezeichnet. Verbindungen mit

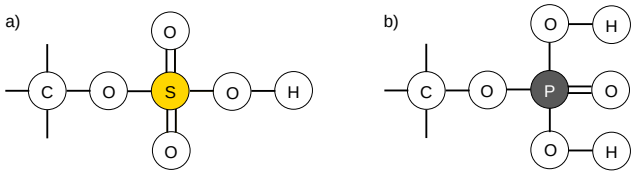


Abb. 2.12: *Zwei Beispiele anorganischer Säurereste: a) Ester der Schwefelsäure (H_2SO_4) b) Phosphate, Ester der Phosphorsäure (H_3PO_4)*

Doppel- oder Dreifachbindungen heißen *ungesättigte Verbindungen* oder *Olefine*. Alternativ zur Unterscheidung nach dem Vorhandensein oder Nichtvorhandensein von Mehrfachbindungen ist auch eine Unterscheidung von *Aliphaten* und *Cycloverbindungen* möglich. Aliphaten sind verzweigte oder unverzweigte kettenförmige Verbindungen, Cycloverbindungen haben ringförmige Gestalt.

Aromaten sind dagegen, wie bereits erwähnt, durch ein delokalisiertes π-Elektronensystem charakterisiert, und damit grenzt ihr Elektronenzustand sie sowohl von den ungesättigten als auch von den gesättigten Kohlenwasserstoffen ab. Auch ihr chemisches Verhalten unterscheidet sie von beiden. Sie werden im folgenden Abschnitt eingehender betrachtet.

2.2 Aromatische Verbindungen

Die Gruppe der aromatischen Verbindungen wird in der organischen Chemie gesondert betrachtet, da diese sich, aufgrund ihrer besonderen Elektronenstruktur in ihren chemischen und physikalischen Eigenschaften von den übrigen Substanzklassen abheben. Die Bezeichnung Aromaten leitet sich von einem auffallenden „aromatischen“ Geruch der ersten in der Aromatenchemie betrachteten Verbindungen her. Es gibt jedoch keinen systematischen Zusammenhang zwischen dieser frühen Beobachtung und der Aromatizität im chemischen Sinne.

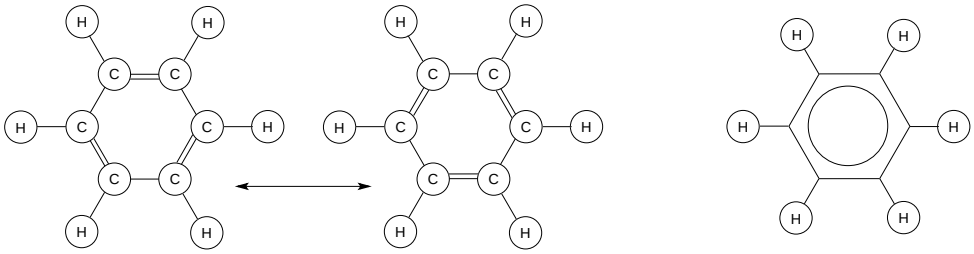


Abb. 2.13: *Die beiden mesomeren Grenzformeln des Benzols (links) suggerieren das Vorliegen von alternierenden Doppel- und Einfachbindungen, was jedoch nicht der tatsächlichen Elektronenverteilung entspricht. Die Kreisschreibweise (rechts) verdeutlicht den wahren Elektronenzustand.*

Als klassischer Vertreter aromatischer Verbindungen soll exemplarisch das in Abbildung 2.13 dargestellte Benzol betrachtet werden. Die Besonderheit dieser Verbindung ist, daß sie ein sich über den gesamten Ring erstreckendes delokalisiertes π-Elektronensystem besitzt: Aufgrund der Planarität des Sechsringes und der Bindungswinkel der Ringglieder befinden sich die unhybridisierten *p*-Orbitale der sp^2-hybridisierten Kohlenstoffatome in einer günstigen Orientierung, welche, wie in Abschnitt 2.1.3 beschrieben, deren Kombination zu π-Molekülorbitalen erlaubt. Anders als bei einer isolierten Doppelbindung ist dies jedoch zu

beiden Seiten jedes Kohlenstoffatoms der Fall, so daß die betreffenden Elektronen nicht in einzelnen Doppelbindungen lokalisiert sind, sondern sich in einer Art „Wolke“ über den gesamten Ring verteilen. Es handelt sich also nicht, wie durch die beiden mesomeren Grenzformeln suggeriert, um eine Struktur mit jeweils drei Doppel- und Einfachbindungen. Physikalische Untersuchungen zeigen, daß statt dessen in der Tat sechs äquivalente Bindungen vorliegen.

Die Energiedifferenz zwischen dem Elektronenzustand abwechselnder Doppel- und Einfachbindungen und einem aromatischen System wird als *Delokalisierungsenergie* bezeichnet. Bei nichtaromatischen Verbindungen müßte dieser Energiebetrag zugeführt werden, um eine Delokalisierung der vorhandenen π-Elektronen zu erreichen, das heißt eine geeignete Konformation zu forcieren. Bei aromatischen Verbindungen dagegen ist die Delokalisierungsenergie negativ, das bedeutet, der aromatische Elektronenzustand ist energetisch günstiger und wird daher spontan eingenommen. Er prägt die Eigenschaften aromatischer Verbindungen und wird aus diesem Grund auch zur Definition des Begriffs der Aromatizität herangezogen. Im folgenden soll er daher näher charakterisiert werden.

2.2.1 Elektronenzustand aromatischer Verbindungen

Weit verbreitet ist die Definition aromatischer Systeme über die HÜCKEL-Regel[6]: Demnach sind alle planaren, monozyklischen Ringe mit $4n+2$ π-Elektronen in zyklischer Anordnung aromatisch. Diese Regel gilt im besonderen mit $n=1$ für Benzol, für das Cyclopentadienyl-Anion und Pyrrol (vgl. Abbildung 2.14). Im Fall von Pyrrol ist mit Stickstoff ein Heteroatom am Ring beteiligt, weshalb man es zur Gruppe der *Heteroaromaten* zählt. Das Beispiel des Cyclopentadienyl-Anions zeigt, daß auch geladene Teilchen aromatisch sein können.

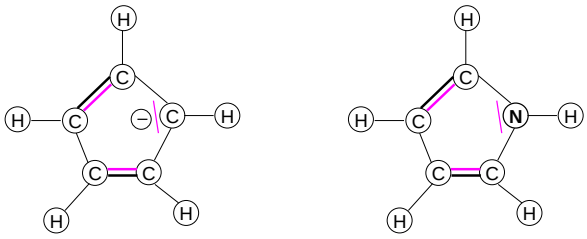


Abb. 2.14: *Das Cyclopentadienyl-Anion (links) besitzt ein 6-π-Elektronensystem, an welchem das freie Elektronenpaar (negative Ladung) beteiligt ist. Analoges gilt für Pyrrol (rechts); hier stammt das freie Elektronenpaar vom Stickstoff.*

Die Definition der Aromaten über die HÜCKEL-Regel ist jedoch nicht erschöpfend, es gibt auch aromatische Verbindungen, auf welche diese Regel nicht zutrifft. Insbesondere zählen auch solche Verbindungen zu den Aromaten, deren Grundgerüst aus mehreren, mit einander verbundenen Benzolringen besteht, so daß ein noch weitläufigeres System delokalisierter π-Elektronen entsteht. Ein Beispiel für eine solche Verbindung ist das in Abbildung 2.15 dargestellte Phenanthren. Anstelle von Benzol können auch andere Aromaten als Bausteine in dem Ringsystem vorkommen. Entscheidendes Kriterium bleibt die auf einer günstigen Orientierung der *p*-Orbitale beruhende Ausbildung eines delokalisierten π-Elektronensystems.

Bei solchen *polyzyklischen* Verbindungen ist es in der Regel anders als bei Benzol nicht möglich, eine geeignete Notation anzugeben, welche dem aromatischen Charakter der Verbindung gerecht wird; es sei daher noch einmal auf die nötige Aufmerksamkeit für Systeme

[6] E. Hückel, 1931, vgl. z.B. [Bud90], S. 309

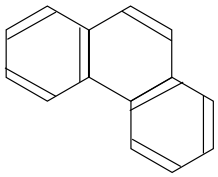


Abb. 2.15: *Phenanthren ist eine aromatische Verbindung: Es besteht quasi aus drei mit einander verschmolzenen aromatischen Benzolringen. Es entsteht ein gemeinsames, sich über die gesamte Struktur erstreckendes π-Elektronensystem.*

konjugierter Doppelbindungen hingewiesen. Zugleich spielt jedoch auch die Molekülgeometrie (z.B. die Bindungswinkel an Heteroatomen oder die Anordnung der Benzoleinheiten in komplizierten Ringsystemen) eine Rolle, da nur bei entsprechender Ausrichtung der unhybridisierten *p*-Orbitale ein delokalisiertes π-Elektronensystem entstehen kann.

Die Bedeutung der Molekülgeometrie wird an dem relativ einfachen Beispiel von [10]Annulen (vgl. Abbildung 2.16) deutlich: Obwohl sie der HÜCKEL-Regel mit $n = 2$ zu genügen scheint, verhält sich diese Verbindung nicht aromatisch. Dies rührt daher, daß zwei der Ringglieder, welche ins Innere des Ringes gerichtete Wasserstoffatome tragen, einander aufgrund der sterischen Hinderung ausweichen und so die Planarität des Ringes aufheben. Im Gegensatz dazu wird im in derselben Abbildung gezeigten sogenannten VOGEL-Aromaten[7] die planare Konformation durch ein brückenartiges CH_2-Glied fixiert, so daß ein aromatisches System entstehen kann.

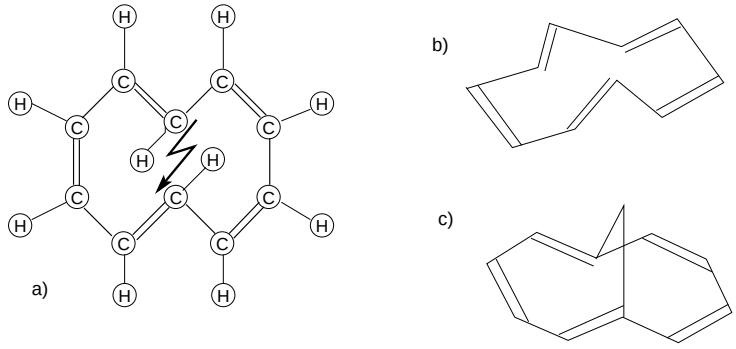


Abb. 2.16: *Im [10]Annulen behindern sich die beiden nach innen gerichteten Wasserstoffatome (a), so daß zwei Ringatome aus der Ebene heraus ausweichen (b). Im Vogel-Aromaten fixiert eine CH_2-Brücke die Konformation (c).*

Während die Einordnung einer Verbindung als Aromat also teilweise recht schwierig sein kann, können jedoch viele Betrachtungen betreffend die Elektroneneigenschaften oder Strukturcharakteristika von Aromaten im allgemeinen gut anhand einfacherer, leicht einzuordnender Verbindungen angestellt werden. Benzol als klassisches Beispiel einer aromatischen Verbindung sowie seine Derivate sind in der Geschichte der Aromatenchemie eingehend untersucht und wohl dokumentiert worden. Im folgenden soll daher, trotz der Vielfalt der Aromaten insgesamt, diese Teilmenge für die vorliegende Arbeit verwendet werden.

Benzol wurde in Abbildung 2.13 bereits vorgestellt; Benzolderivate sind Verbindungen,

[7] E. VOGEL zeigte 1964 den aromatischen Charakter dieser Verbindung mit dem systematischen Namen 1,6-Methano-cyclodecapentaen

die sich durch Substitution von einem oder mehreren der sechs Wasserstoffatome strukturell auf Benzol zurückführen lassen. Diese Verbindungen und eine systematische Charakterisierung ihrer Struktur sind gegenstand des folgenden Abschnitts.

2.2.2 Substitutionsmuster an Benzolderivaten

Oft werden Benzolderivate nach ihrem *Substitutionsmuster* unterschieden, das heißt nach Zahl (*Substitutionsgrad*), Art und relativer Stellung der Substituenten zu einander. Werden jedoch sehr viele unterschiedliche Substituenten betrachtet, so entsteht auch eine sehr breit gefächerte Unterteilung. In vielen Fällen werden daher die möglichen Substituenten oder Substituentenkombinationen wiederum zu Gruppen mit ähnlichen Eigenschaften zusammengefaßt, um eine für den Arbeitsbereich sinnvolle Hierarchie zu erhalten.

Bezüglich der relativen Stellung von Substituenten ist es wichtig, die einzelnen Positionen des Benzolringes zu unterscheiden. In *monosubstituiert*en Benzolderivaten, Verbindungen mit genau einer von Wasserstoff verschiedenen Gruppe an einem der Ringatome, wird das Atom, an welches diese Gruppe gebunden ist, als *ipso*-Position bezeichnet. In *disubstituiert*en Benzolderivaten unterscheidet man drei Fälle: direkt benachbarte Gruppen heißen *ortho*-ständig, befindet sich genau ein unsubstituiertes Kohlenstoffatom zwischen ihnen, so sind die Substituenten *meta*-ständig, und sind sie einander gegenüberliegend angeordnet, so spricht man von *para*-ständigen Substituenten.

Die Begriffe *ipso, ortho, meta* und *para* können darüber hinaus in analoger Weise zur Bezeichnung der relativen Positionen innerhalb des Ringes unabhängig von der dort gebundenen Gruppe verwendet werden. Die *ipso*-Position wird als Fokuspunkt festegelegt, Abbildung 2.17 verdeutlicht, wie die *ortho*-, *meta*- und *para*-Positionen sich entsprechend ergeben.

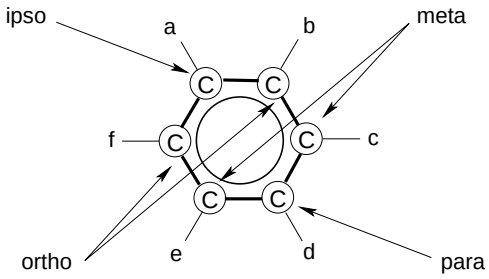


Abb. 2.17: *Bezeichnung der sechs Positionen des Benzolringes nach ihrer relativen Stellung. Die einzelnen Substituenten sind schematisch mit a) – f) bezeichnet.*

Damit sind nun die theoretischen Betrachtungen von Bindungsstrukturen, und damit unmittelbar verbunden auch von Elektronenstrukturen innerhalb eines Moleküls, im allgemeinen wie auch im speziellen Falle aromatischer Verbindungen abgeschlossen. Mit diesem Vorwissen gibt nun der folgende Abschnitt im Rahmen der Thematik der Strukturaufklärung eine Einführung in die spezielle Methodik der NMR-Spektroskopie. Zwischen der Elektronenverteilung innerhalb des Moleküls und den konkreten Untersuchungsbefunden besteht dabei ein direkter systematischer Zusammenhang.

2.3 Strukturaufklärung in der organischen Chemie

Wie bereits in Kapitel 1 erwähnt hat es die Strukturaufklärung in der organischen Chemie zum Ziel, den Aufbau der Moleküle einer unbekannten Verbindung aufzudecken. Die grundlegendste Information hierzu liefern die sogenannte qualitative und quantitative Elementaranalyse, welche Aufschluß über die beteiligten chemischen Elemente sowie deren jeweilige Multiplizitäten, also die Summenformel der Verbindung, geben (vgl. z.B. [BWe91], Kapitel 1). Diese Information bildet die Basis für alle weiteren Untersuchungen und ist leicht zugänglich, so daß sie in der Regel als gegeben vorausgesetzt werden kann.

In Analogie zu der auf Seite 8 beschriebenen Hierarchie betreffend den Detailgrad der Strukturbeschreibung können anschließend hinsichtlich der strukturbezogenen Information über das Molekül nach und nach weitere Details ermittelt werden. Um als nächstfeinere Information die Konstitution des Moleküls zu ermitteln, können verschiedene spektroskopische Methoden eingesetzt werden. Bei der Wahl geeigneter Verfahren ist die Erkennung von Substituenten anhand charakteristischer Befunde im Spektrum möglich, andere Untersuchungen, vorrangig aus dem Bereich der *NMR-Spektroskopie,* geben Aufschluß über größere Teilstrukturen, etwa die Nachbarschaft bestimmter Strukturfragmente.

Die ermittelten Teilstrukturen können schließlich wie Puzzleteile zusammengesetzt werden. Ist es nicht möglich, die Konstitution des Moleküls auf diese Weise eindeutig zu ermitteln, müssen gezielt zusätzliche Untersuchungen herangezogen werden. Auch Vorwissen, das sich z.B. aus der Herkunft der Substanzprobe oder ihren chemischen Eigenschaften ableiten läßt, ist in diesem Zusammenhang von Bedeutung. Beispielsweise könnte bekannt sein, daß das Molekül zu einer bestimmten Substanzklasse gehört und demzufolge bestimmte Strukturbausteine enthalten muß oder andere nicht enthalten darf.

Darüber hinaus gilt, daß mit fortschreitendem Detailgrad der Strukturbeschreibung zunehmend anspruchsvollere Untersuchungen nötig sind. Das Ermitteln der Konfiguration etwa ist mit Hilfe der Analyse bestimmter Parameter aus speziellen spektroskopischen Experimenten zweifellos möglich, geht jedoch mit einem deutlich höheren Aufwand in der Spektrenaufnahme wie auch in der Auswertung der aufgenommenen Daten einher.

Für die gegenwärtige Arbeit soll die Frage der Konstitution (Detailgrad 2 im Sinne der Hierarchie auf Seite 8) Gegenstand der Betrachtung sein. Bereits in diesem Bereich sind umfangreiche Aufgaben zu lösen, wie der folgende Einblick in die NMR-Spektroskopie und die Auswertung von NMR-Spektren zeigen wird. Erst wenn die Konstitutionsinformation gegeben ist, können Verfahren zur Untersuchung noch detaillierterer Struktureigenschaften darauf aufsetzen; dies geht jedoch deutlich über den Rahmen der vorliegenden Arbeit hinaus.

2.3.1 Grundlagen der NMR-Spektroskopie

Spektroskopische Methoden, insbesondere die NMRSpektroskopie, haben sich in den vergangenen Jahren und Jahrzehnten stark entwickelt. Heute ist mit vergleichsweise wenig technischem und finanziellem Aufwand ein hohes Maß an Information zugänglich, so daß die NMRSpektroskopie als *die* Methode der Wahl zur Strukturaufklärung bezeichnet werden kann. Ihre allgemeinen Grundlagen werden im folgenden beschrieben.

Spektroskopie ist die Untersuchung der quantisierten Wechselwirkung von Strahlungsenergie mit Materie. Bei der Einstrahlung von Energie kann ein Molekül diese absorbieren, um einen *Zustandsübergang* zu vollziehen: Eine Triebkraft der Chemie ist, wie bereits erwähnt, das Anstreben eines möglichst energiearmen (energetisch günstigen) Zustandes. Energiereichere Zustände können nur angenommen werden, wenn entsprechend mehr Energie zur Ver-

fügung steht, wie es durch die Energieeinstrahlung beim spektroskopischen Experiment der Fall ist.

Die Intensität der absorbierten Strahlung ist im *Spektrum* gegen eine Energieachse aufgetragen. Von besonderem Interesse sind dabei die Absorptionsmaxima (*Peaks*), welche durch ihre Lage auf der Energieachse bestimmte Zustandsübergänge im untersuchten Molekül charakterisieren: Jedem möglichen Zustand ist ein diskretes Energieniveau zugeordnet. Ein Zustandsübergang kann nicht durch Einstrahlung eines beliebigen Energiebetrags erreicht werden, sondern die zur Verfügung gestellte Energie muß genau der Differenz der beiden Energieniveaus entsprechen. Wird nun ein geeigneter Energiebetrag eingestrahlt, so kann aus seiner Absorption auf den Vollzug des Zustandsüberganges geschlossen werden. Die Teile des Moleküls, welche für die Absorption verantwortlich sind und somit im Spektrum indirekt sichtbar sind, werden *Chromophore* genannt.

Die Energie der zugeführten Strahlung ist abhängig von ihrer Frequenz: Der Energiebetrag ΔE entspricht dem Produkt der Frequenz ν mit dem PLANCK'schen Wirkungsquantum h, das eine Konstante ist.

$$\Delta E = h \cdot \nu \tag{2.1}$$

Damit wird auch ersichtlich, weshalb nicht nur ein Teil der Energie absorbiert werden kann: Sie ist direkt proportional zur Frequenz, und eine nur teilweise Aufnahme würde eine Änderung der Frequenz, also eine physikalische Veränderung der Strahlung, erfordern.

NMR-Spektroskopie ist eine spezielle spektroskopische Methode. Das Kürzel NMR steht für *nuclear magnetic resonance* (Kernmagnetresonanz) und gibt somit bereits wieder, wodurch sich die Methode auszeichnet: Bezüglich der erwähnten Zustandsübergänge werden Atomkerne betrachtet, und bei der eingestrahlten Energie handelt es sich um elektromagnetische Energie. Im NMR-Spektrum wird die Reaktion der Kerne auf ihre sogenannten *Resonanzfrequenzen* sichtbar gemacht.

In einem Atomkern, welcher einen Spin besitzt (dieser ist einem mechanischen Drehimpuls vergleichbar), induzieren die enthaltenen Protonen als bewegte Ladungen ein magnetisches Moment μ. In einem statischen Magnetfeld B_0 parallel zur z-Achse des Raumes ist μ ein im Koordinatenursprung verankerter Vektor, welcher mit der Frequenz ν_0 um die Richtung des äußeren Feldes B_0 präzessiert. Abbildung 2.18 veranschaulicht dies. Die Frequenz ν_0 wird *Larmorfrequenz* genannt. Sie bestimmt über den Zusammenhang zwischen Energie und Frequenz (vgl. Gleichung (2.1)) das Energieniveau des zugehörigen *Spinzustands*.

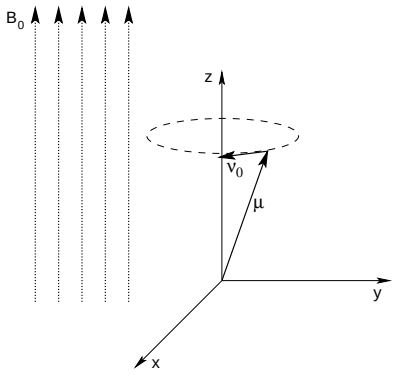


Abb. 2.18: *Atomkerne mit einem Spin besitzen ein magnetisches Moment μ, welches in einem statischen Magnetfeld B_0 mit der Larmor-Frequenz ν_0 um die z-Richtung des Raumes präzessiert.*

Der Betrag des Spins läßt sich mit Hilfe der SCHRÖDINGER-Gleichung berechnen; für

die NMR-Spektroskopie ist jedoch die Anzahl unterschiedlicher Spinzustände eines Kerns in einem angelegten Magnetfeld von größerem Interesse. Diese kann mit Hilfe der *Spinquantenzahl I* bestimmt werden:

$$\text{Anzahl Spinzustände} = 2I + 1 \tag{2.2}$$

Eine Quantenzahl ist allgemein ein Index, welcher mögliche Zustände von Systemen numeriert (vgl. S. 10). Die Spinquantenzahl nimmt diese Indizierung durch Festlegen der Richtung des Spins relativ zur Richtung des äußeren Magnetfeldes vor; genau genommen wird die *z*-Komponente der Richtung festgelegt, wobei die *z*-Achse durch die Richtung des äußeren Feldes gegeben ist. Ihr Wert hängt von der *Ladungszahl* (Zahl der Protonen) und von der *Massenzahl* (Zahl der Protonen und Neutronen zusammen) ab. Hinsichtlich der Kompensationseffekte der Einzelspins dieser Kernbestandteile sind unterschiedliche Typen von Kernen zu unterscheiden, und zwar danach, ob Ladungs- und Massenzahl jeweils gerade oder ungerade sind.

Sind beide Zahlen gerade, so hat I den Wert 0, und es gibt nur einen einzigen Spinzustand. Da demzufolge keine Zustandsübergänge vollzogen werden können, sind solche sogenannten (g,g)-Kerne für die NMR-Spektroskopie ungeeignet. Nicht völlig ungeeignet, jedoch problematisch sind (g,u)-Kerne mit gerader Massen- und ungerader Ladungszahl. Für solche Kerne ist I eine natürliche Zahl, deren Betrag vom jeweiligen Atomtypus abhängt. Sie besitzen somit nach Gleichung (2.2) mindestens drei verschiedene Zustände. Damit besteht das Problem, daß ein und derselbe Energiebetrag geeignet ist, unterschiedliche Zustandsübergänge anzuregen, wie Abbildung 2.19 verdeutlicht:

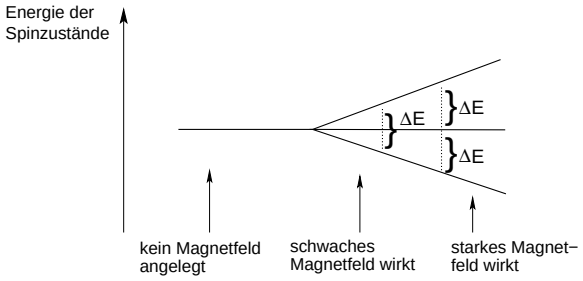


Abb. 2.19: *Je stärker das auf einen Kern wirkende Magnetfeld ist, desto größer wird die Differenz der Energieniveaus. Die drei markierten Zustandsübergänge haben alle dieselbe Energiedifferenz ΔE und werden somit alle durch dieselbe Energieeinstrahlung angeregt.*

Ohne ein wirkendes Magnetfeld ist der Energieunterschied der Spinzustände Null (links in der Grafik). Wirkt ein schwaches Magnetfeld (Mitte), so liegt ein Betrag von ΔE zwischen dem energieärmsten und dem energiereichsten Zustand. Wirkt jedoch ein stärkeres Magnetfeld (rechts), so liegt derselbe Betrag ΔE zwischen dem energieärmsten und dem mittleren sowie zwischen dem mittleren und dem energiereichsten Zustand. Bei Absorption des betreffenden Energiebetrags ist also nicht eindeutig zu sagen, welcher der drei Zustandsübergänge stattgefunden hat, so daß sie mit einander verwechselt werden können. NMR-Spektren von (g,g)-Kernen sind daher schwierig zu interpretieren, weshalb man, sofern alternative Methoden zur Verfügung stehen, NMR-Untersuchungen an ihnen vermeidet.

Am besten für die NMR-Spektroskopie geeignet sind Kerne mit ungerader Massenzahl ((u,g)- und (u,u)-Kerne, *Fermionen*): Für sie ist die Spinquantenzahl I ein ganzzahliges positives Vielfaches von $\frac{1}{2}$. Insbesondere gehören zu diesen Kernen der Wasserstoffkern ^{1}H und

der Kohlenstoffkern ^{13}C: Für beide ist $I = \frac{1}{2}$, somit besitzen sie nach Gleichung (2.2) beide zwei Spinzustände. Für diesen Fall ist der Zusammenhang zwischen Energiebetrag und Zustandsübergang eindeutig. NMR-Spektren dieser beiden Kerne werden darum am häufigsten aufgenommen; zudem sind Kohlenstoff und Wasserstoff die zentralen und bei weitem am häufigsten vorkommenden Elemente in organischen Verbindungen.

Anschaulich kann man sich Kerne mit zwei Spinzuständen als kleine Stabmagneten vorstellen, um sich den energetischen Unterschied ihrer beiden Spinzustände zu veranschaulichen. Ein solcher Stabmagnet kann innerhalb eines Magnetfeldes zwei mögliche Ausrichtungen haben: die parallele und die antiparallele. Sie korrespondieren zu den beiden Spinzuständen des Kerns. Aufgrund der Abstoßung gleichartiger Pole erfordert es Anstrengung, den Stabmagneten in die antiparallele Ausrichtung zu bringen und in dieser Orientierung festzuhalten, während die parallele spontan eingenommen wird. Die parallele Ausrichtung entspricht also dem energetisch günstigeren, die antiparallele dem angeregten Spinzustand. Die nötige „Kraft“ für das Umdrehen des parallel ausgerichteten Stabmagneten repräsentiert die nötige Energie für den Übergang in den angeregten Spinzustand.

Liegt überhaupt kein Magnetfeld an, so macht es keinen Unterschied, in welcher Orientierung sich der Stabmagnet befindet, die Energiedifferenz ist Null. Mit wachsender Feldstärke wächst jedoch auch für Kerne mit zwei Spinzuständen die Differenz der Energieniveaus, wie für Kerne mit drei Spinzuständen bereits in Abbildung 2.19 angedeutet. Genau dies ist der entscheidende Punkt für die Gewinnung von Information aus NMR-Spektren: Neben der angelegten Feldstärke hat die umgebende Elektronenhülle einen Einfluß auf die effektiv wirkende Feldstärke, da sie den Kernspin abschirmt, indem sie ein dem äußeren Magnetfeld entgegengerichtetes Feld erzeugt. Der Kernspin erfährt dadurch nicht die volle Stärke des anliegenden Feldes, und seine Larmorfrequenz verringert sich entsprechend, da sie von der tatsächlich auf ihn wirkenden und nicht von der angelegten Feldstärke abhängt.

Dies erklärt auch, weshalb die in Abbildung 2.19 gekennzeichneten Zustandsübergänge mit derselben Differenz ihrer Energieniveaus alle mit einander verwechselt werden können, obwohl sie bei unterschiedlichen Feldstärken auftreten: Auch bei gleicher angelegter Feldstärke kann durch unterschiedlich starke Abschirmung auf verschiedene Kerne desselben Moleküls effektiv ein unterschiedlich starkes Feld wirken.

Die Abhängigkeit der Larmorfrequenz eines Kernspins von dessen chemischer Umgebung (Elektronendichte), das heißt die Lage der Peaks auf der *x*-Achse, wird als *chemische Verschiebung* bezeichnet. Dies ist die essentielle spektroskopische Information, anhand welcher bei der Auswertung eines NMR-Spektrums Rückschlüsse auf die Struktur des Moleküls gezogen werden. Näheres zur Auswertung der in der chemischen Verschiebung codierten Information wird im nächsten Abschnitt erläutert. Zuvor schließen einige Bemerkungen zur Skalierung der Energieachse im NMR-Spektrum die allgemeine Einführung ab.

Neben der chemischen Verschiebung ist der Zusammenhang zwischen Feldstärke und Differenz der Energieniveaus auch für die Vergleichbarkeit von NMR-Experimenten von erheblicher Bedeutung: Im angelegten Magnetfeld ist die Präzessionsfrequenz des Kernspins (Larmorfrequenz) zur Stärke des Feldes proportional. Da auch Energie und Frequenz proportional sind, rücken die Energieniveaus verschiedener Spinzustände um so weiter auseinander, je stärker das auf den Kern wirkende Magnetfeld ist, da damit auch die Larmorfrequenz wächst. Damit sind basierend auf absoluten Energiebeträgen betrachtete Spektren nicht mit einander vergleichbar, wenn sie bei unterschiedlich starkem Magnetfeld aufgenommen wurden.

Um eine Vergleichbarkeit herzustellen bedient man sich einer relativen Energieskala mit Bezug zu einer standardisierten Referenzsubstanz, deren Larmorfrequenz in Abhängigkeit von der Feldstärke bekannt ist. Im NMR-Experiment wird nun keine absolute Frequenz, sondern die Frequenzdifferenz zu diesem Standardsignal gemessen. Das Standardsignal für ^{1}H-

und ^{13}C-Kerne liefert jeweils TMS (Tetramethylsilan): Der zu untersuchenden Probe wird etwas von dieser Substanz zugesetzt. TMS ist leicht flüchtig, so daß es nach der Untersuchung rasch wieder verfliegt und die Probe so nicht dauerhaft verunreinigt wird. Zwischen seiner Larmorfrequenz (in der Größenordnung von einigen MHz) und der Frequenzdifferenz anderer Signale zu ihr (Einheit Hz) ergibt sich ein Einheitenverhältnis von $1 : 10^6$. Daher wird in der Regel die Frequenzachse in *ppm* skaliert[8]. Es gilt:

$$\text{Lage in ppm} = \frac{10^6 \cdot \text{Abstand von TMS in [Hz]}}{\text{Spektrometer-Frequenz in [Hz]}} \tag{2.3}$$

Da nicht die Frequenz normiert, sondern eine relative Frequenzachse zur Herstellung der Vergleichbarkeit verwendet wird, ist es außerdem möglich, die höhere Auflösung, welche ein stärkeres Magnetfeld mit sich bringt, zu nutzen, wenn aufgrund der benötigten Informationen der damit verbundene hohe technische Aufwand gerechtfertigt ist. Die erhöhte Auflösung beruht auf der Beeinflussung des Anteils von Kernen, welche sich im energetisch günstigeren Grundzustand befinden und die somit für den Übergang in den angeregten Zustand zur Verfügung stehen. Sei α der angeregte und β der Grundzustand eines Atomkerns mit zwei Spinzuständen, dann ist die Anzahl N der Kerne in jedem der zugehörigen Spinzustände durch eine BOLTZMANN-Verteilung gegeben (vgl. [Atk96], Kapitel 18):

$$\frac{N_\beta}{N_\alpha} = e^{\frac{\Delta E}{R \cdot T}} \tag{2.4}$$

R und T sind dabei jeweils Konstanten für ein Experiment, so daß das Verhältnis der Zahl von Kernen in jedem der beiden Zustände allein von der Energiedifferenz ΔE abhängt. Ohne ein angelegtes Feld geht sie gegen Null, so daß beide Niveaus annähernd gleich besetzt sind. Je stärker jedoch das wirkende Feld, desto größer ist wegen der wachsenden Larmorfrequenz auch die Energiedifferenz. Für das Verhältnis $\frac{N_\beta}{N_\alpha}$ bedeutet dies, daß ein Überschuß im günstigeren Energieniveau β geschaffen wird, der mit der Stärke des angelegten Feldes wächst. Somit wird die Beobachtungsgenauigkeit (Absorptionsintensität) erhöht, da mehr Kerne zur Verfügung stehen, um den Übergang in den angeregten Zustand zu vollziehen.

2.3.2 Auswertung von ^{13}C-NMR-Spektren

Der Schwerpunkt der Betrachtungen liegt im folgenden auf der Gewinnung von Konstitutionsinformation aus NMR-Spektren, in erster Linie also bei der Aufklärung der Grundstruktur eines unbekannten Moleküls. Hierzu sind ^{13}C-NMR-Spektren besonders geeignet, da die in diesen untersuchten Kohlenstoffatome das Skelett eines organischen Moleküls bilden und somit seine grundsätzliche Struktur charakterisieren. ^{1}H-NMR-Spektren werden an dieser Stelle nicht näher betrachtet.

Wie bereits erwähnt ist die chemische Verschiebung (Lage des Peaks auf der x-Achse) die essentielle spektroskopische Information, durch deren Auswertung Rückschlüsse auf die Struktur des Moleküls gezogen werden können. Sie hängt von der Elektronendichte ab, welche auf Bindungsstrukturen, Hybridisierung, die Differenz der Elektronegativitäten einzelner Atome usw. zurückgeht, wie in Abschnitt 2.1 beschrieben. Man spricht in diesem Zusammenhang von *abgeschirmt*en und *entschirmt*en Kernen, je nachdem, ob die Elektronendichte in ihrer Umgebung erhöht oder verringert ist. Abgeschirmte Kerne erfahren durch eine erhöhte Elektronendichte eine starke Abschirmung; sie besitzen durch ein *hohes Abschirmfeld*

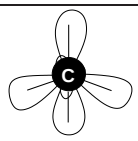
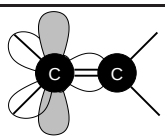
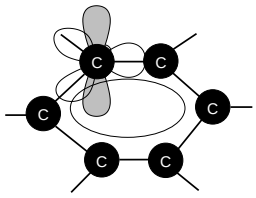
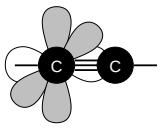
[8] engl. *parts per million*: „Millionstel“. Die Standardgröße δ, welche sich direkt auf die relative Frequenzdifferenz bezieht und die sich von der Lage eines Peaks in ppm um den Faktor 10^6 unterscheidet, findet dagegen kaum Verwendung.

eine kleine Verschiebung. Entschirmte Kerne haben demgegenüber ein kleines oder *tiefes Abschirmfeld* und besitzen eine große Verschiebung, da sie durch eine verringerte Elektronendichte weniger stark abgeschirmt sind. Entsprechend dem hohen oder tiefen Abschirmfeld spricht man auch von *hochfeldverschobenen* oder *tieffeldverschobenen* Absorptionen.

Dem ^{13}C-NMR-Spektrum ist zunächst die Zahl äquivalenter Kerne zu entnehmen. Es handelt sich dabei um *chemische Äquivalenz*, welche genau dann vorliegt, wenn sich zwei Atomkerne in gleicher chemischer Umgebung befinden. Normalerweise entspricht die Zahl der Peaks der Zahl der Kohlenstoffatome in der Verbindung, liegen jedoch Symmetrien im Molekül vor, so findet man entsprechend weniger Peaks: Die Absorptionen derjenigen ^{13}C-Kerne, die sich aufgrund der Symmetrie in einer gleichen chemischen Umgebung befinden, liegen aufeinander, so daß nur ein einziger Peak sichtbar ist.

Zum zweiten besitzen bestimmte Typen von Kohlenstoffkernen charakteristische Absorptionsbereiche. Diese „Typen“ sind entscheidend von der Elektronenstruktur in ihrer Umgebung geprägt, welche über den Grad der Abschirmung des betreffenden Kerns vom äußeren Feld entscheidet. Den größten Einfluß haben hier die Valenzelektronen, welche durch ihre Beteiligung an den Bindungen im Molekül anders als die inneren Elektronen einer erheblichen Varianz unterworfen sind. Damit ergibt sich ein direkter Zusammenhang zwischen Bindungsstruktur und Erscheinung eines bestimmten Atomkerns im NMR-Spektrum.

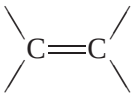
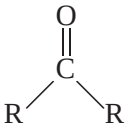
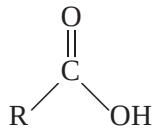
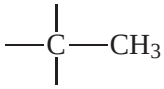
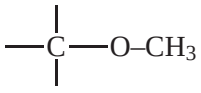
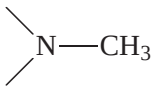
Die Gestalt der Molekülorbitale (vgl. Abschnitt 2.1.3) prägt die Elektronenstruktur in grundlegendem Maß, so daß unterschiedlich hybridisierte Kohlenstoffatome ihre Absorptionen in typischen Bereichen des Spektrums zeigen. Tabelle 2.1 stellt diese typischen Absorptionsbereiche dar. Betrachtet werden dabei nur Atome in reinen Kohlenwasserstoffen, welche ausschließlich an Einfachbindungen (Alkane), an einer Doppelbindung (*Alkene*), einer Dreifachbindung (*Alkine*) bzw. an einem aromatischen System beteiligt sind.

Verbindungsklasse	Hybridisierung		Absorptionsbereich
Alkane	sp^3		0 – 50 ppm
Alkene	sp^2		90 – 140 ppm
Aromaten	sp^2		95 – 155 ppm
Alkine	sp		70 – 85 ppm

Tab. 2.1: *Charakteristische Absorptionsbereiche der Grundtypen von Kohlenstoffatomen. In den Grafiken sind Hybridorbitale weiß und unhybridisierte p-Orbitale grau dargestellt.*

Die beschriebenen Intervalle werden jedoch verlassen bzw. verschoben, wenn Heteroatome im Molekül vorkommen. Mesomere und induktive Effekte (vgl. Abschnitt 2.1.4) führen zu Veränderungen der Elektronenstruktur und somit auch zu einer Veränderung des Abschirmfeldes, was die Lage der betreffenden Peaks beeinflußt. Dadurch ist es möglich, bestimmte funktionelle Gruppen anhand von ^{13}C-NMR-Absorptionen zu erkennen oder zu unterscheiden. Dies beruht jedoch nicht auf einer Sensitivität gegenüber funktionellen Gruppen (Molekülabschnitten mit einer bestimmten chemischen Funktionalität), sondern auf den entsprechenden Substituenteneinflüssen auf die Elektronenstruktur.

Tabelle 2.2 zeigt die typischen Absorptionsbereiche einiger kohlenstoffhaltiger Gruppen sowie jeweils die chemische Verschiebung eines Kohlenstoffatoms derselben Hybridisierung in einem reinen Kohlenwasserstoff. Dadurch wird die große Bedeutung der jeweiligen Substituenteneinflüsse noch einmal verdeutlicht. Sie ist auf die polarisierende Wirkung von Heteroatomen zurückzuführen, welche die Elektronendichte im Molekül beeinflußt. In Extremfällen (Mehrfachbindung zum Heteroatom) kann der Einfluß des Heteroatoms die Elektronenstruktur sogar völlig dominieren, so daß die zugehörige Absorption weitab von der eines unsubstituierten Kohlenstoffatoms erscheint.

Gruppe		typische Absorption
Alken		90 – 140 ppm
Carbonyl (Keton)		190 – 220 ppm
Carbonyl (Carbonsäure, auch Derivate)		160 – 180 ppm
Alkin	–C≡C–	70 – 85 ppm
Cyanid	–C≡N	110 – 120 ppm
Methyl (Alkan)		0 – 50 ppm
Methyl (Ether)		50 – 60 ppm
Methyl (Amin)		25 – 45 ppm

Tab. 2.2: *Typische Absorptionsbereiche einiger kohlenstoffhaltiger funktioneller Gruppen.*

Da sich Elektroneneinflüsse über das Gesamtsystem der Bindungen („*Elektronenhülle*“) fortpflanzen, so daß prinzipiell jeder Kern innerhalb des Moleküls mehr oder minder stark beeinträchtigt wird, kann ein Peak nie isoliert für sich betrachtet werden, sondern steht immer im Kontext des Gesamtspektrums. Er beinhaltet neben Informationen über das Kohlenstoffatom, das ihn verursacht, auch Information über dessen chemische Umgebung, das heißt benachbarte Atome. Diese komplexen Zusammenhänge sind die Ursache der Herausforderung bei der Interpretation eines ^{13}C-NMR-Spektrums, begründen aber auch den Umfang der Information, die auf diese Weise gewonnen werden kann.

Den Einfluß von Substituenten, vor allem, daß dieser nicht lokal ist und dadurch unter Umständen das ganze ^{13}C-NMR-Spektrum verändern kann, verdeutlicht das folgende Beispiel. Darin werden Propan und Chlorpropan mit einander verglichen und somit die Auswirkung eines Chloratoms bei dessen Einbringung in die Molekülstruktur betrachtet. Propan (vgl. Abbildung 2.20) hat die Summenformel C_3H_8. Die beiden äußeren Kohlenstoffatome sind äquivalent, das mittlere unterscheidet sich geringfügig von ihnen. Im NMR-Spektrum[9] sind zwei Absorptionen mit beinahe identischer chemischer Verschiebung zu beobachten.

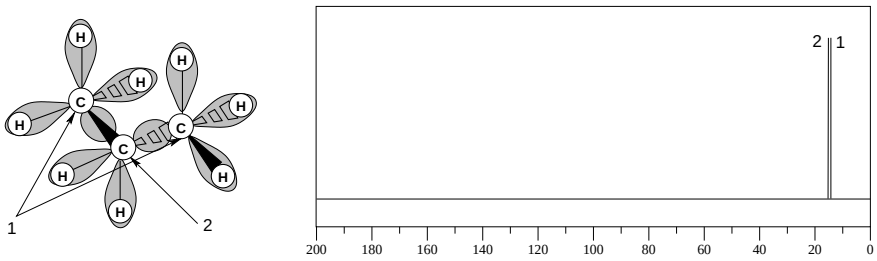


Abb. 2.20: *Struktur (links) und Spektrum (rechts) des Propans. In der Struktur sind die Bindungen zur Veranschaulichung der Elektronendichte als „Elektronenwolken“ angedeutet.*

Wird Chlor als Substituent an einem der äußeren Kohlenstoffatome eingebracht, so verursacht dessen deutlich höhere Elektronegativität gegenüber Kohlenstoff einen Elektronensog. Die betreffende Bindung wird polarisiert, so daß eine erhöhte Elektronendichte auf der Seite des Chloratoms entsteht. Der zugleich auftretende Elektronenmangel am Kohlenstoffatom führt zu einem Elektronensog an dessen übrigen Bindungen. Dies wiederum läßt bei den Bindungspartnern ebenfalls einen, wenn auch schwächeren, Elektronenmangel entstehen. Abbildung 2.21 verdeutlicht dies:

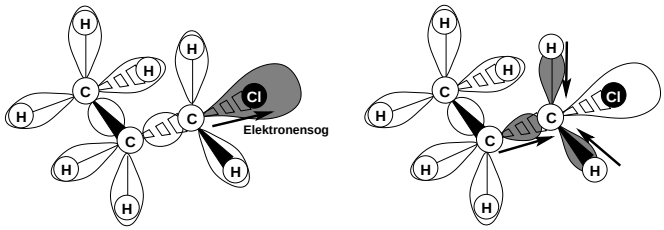


Abb. 2.21: *Chlor verursacht durch seine höhere Elektronegativität einen Elektronensog (links). Um den dadurch entstehenden Elektronenmangel am benachbarten Kohlenstoffatom auszugleichen, werden Elektronen aus den übrigen Bindungen angezogen (rechts).*

[9] Es sei an dieser Stelle darauf hingewiesen, daß alle in dieser Arbeit dargestellten Spektren exemplarische Zeichnungen sind. Die Lage der Peaks entstammt der Literatur oder digital vorliegenden Spektren, die freundlicherweise von der BASF AG, Ludwigshafen, zur Verfügung gestellt wurden.

Da sich auch die Elektronenumgebung der beiden unsubstituierten Kohlenstoffatome verändert hat, weisen sie im NMR-Spektrum nicht mehr dieselbe chemische Verschiebung auf wie in der unsubstituierten Verbindung (vgl. Abbildung 2.22). Außerdem sind nun die beiden äußeren Methylgruppen nicht mehr äquivalent, da die eine nun das Chloratom trägt. Daher sind für 1-Chlorpropan drei Peaks im NMR-Spektrum zu erkennen.

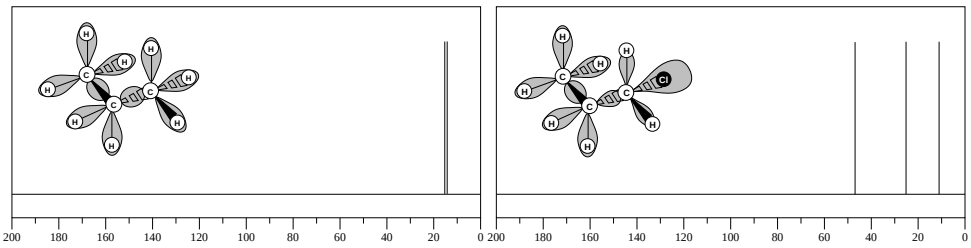


Abb. 2.22: *Die Spektren von Propan und 1-Chlorpropan im Vergleich. Aufgrund der unterschiedlichen Elektronenstruktur erscheinen auch die beiden unsubstituierten Kohlenstoffatome an veränderten Positionen.*

Für den Zusammenhang zwischen chemischer Verschiebung und Elektronenstruktur gilt allgemein, daß ein negativer induktiver Substituenteneffekt ($-I$-Effekt) zu einer Erhöhung der Verschiebung des α-*Kohlenstoff*s führt. Mit α-Kohlenstoff bezeichnet man dasjenige Kohlenstoffatom, welches den Substituenten trägt, sein direkter Nachbar wird mit β bezeichnet, das nächstentferntere Kohlenstoffatom mit γ und so weiter. Die chemische Verschiebung des α-Kohlenstoffs wird proportional zur Stärke des $-I$-Effekts erhöht. Auch der β-Kohlenstoff erfährt, wie in obigem Beispiel angedeutet, eine (gleichwohl geringere) Beeinflussung, welche in Alkanen wie Propan zu einer Erhöhung der chemischen Verschiebung führt. Der γ-Kohlenstoff hingegen weist eine Verringerung seiner chemischen Verschiebung auf; dies wird mit dem Begriff γ-*Effekt* bezeichnet. Er hängt mit einer sterisch induzierten Polarisierung innerhalb des Moleküls zusammen.

Entferntere Kohlenstoffatome als die γ-Position bleiben in der Regel unbeeinflußt. Eine Ausnahme bilden aromatische Systeme, da sie aufgrund ihrer Elektronenstruktur Substituenteneinflüsse besonders gut propagieren können. Mesomere und induktive Substituenteneffekte haben über das aromatische System einen deutlichen Einfluß auf die chemische Verschiebung aller sechs Benzolringatome. Dies erlaubt im besonderen die Erkennung von Substitutionsmustern an Benzolderivaten.

Schließlich ist nun noch zu erwähnen, daß es innerhalb der NMR-Spektroskopie unterschiedliche Methoden (vgl. [Bre92] Abschnitt 2.2) gibt, welche auf bestimmte Anwendungen und das Gewinnen dementsprechender Informationen spezialisiert sind. Die hier verwendeten Spektren sind *protonenbreitbandentkoppelte* Spektren. Bei ihrer Aufnahme werden gezielt alle *Kopplungen* zwischen ^{13}C- und ^{1}H-Kernen ausgeschaltet.

Unter dem Begriff der Kopplung versteht man, daß die Zustände des betrachteten Kerns und der Kerne in seiner Umgebung sich gegenseitig beeinflussen, sofern sie, wie im Falle von ^{1}H und ^{13}C, dieselbe Spinquantenzahl haben. Durch die Wechselwirkung wird die Resonanzfrequenz des betrachteten Kohlenstoffkerns um einige Hz verschoben; dadurch entsteht eine Aufspaltung der Absorptionen in Liniengruppen, sogenannte *Multipletts*. Die Zahl der Multiplettlinien hängt von der Zahl gebundener Wasserstoffatome ab: Ein Kohlenstoffatom ohne benachbarte Wasserstoffatome erscheint als *Singlett* (einzelne Linie ohne Aufspaltung), ein CH-Fragment verursacht ein *Dublett* (Zweiergruppe), ein CH_2-Fragment ein *Triplett* (Dreiergruppe) und ein CH_3-Fragment ein *Quadruplett* (Vierergruppe).

Die Kopplung der Kerne benachbarter Kohlenstoffatome, die natürlich ebenfalls dieselbe

Spinquantenzahl haben, spielt dagegen in der Praxis keine Rolle: Nur etwa 1,1% des natürlich vorkommenden Kohlenstoffs gehört zu dem *Isotop* ^{13}C. Isotope sind Varianten eines chemischen Elements mit einer unterschiedlichen Anzahl von Neutronen im Kern. Das bei weitem am häufigsten vorkommende Isotop ^{12}C hat die Massenzahl 12 und die Ladungszahl 6, es besitzt als (g,g)-Kern also nur einen einzigen Spinzustand (vgl. Abschnitt 2.3.1) und ist somit für die NMR-Spektroskopie nicht nutzbar. ^{13}C-Kohlenstoff besitzt durch sein zusätzliches Neutron eine ungerade Massenzahl und somit als (u,g)-Kern mehrere unterscheidbare Spinzustände. Seine Verbreitung von 1,1% in der Natur genügt zwar, um die Absorptionen der betreffenden Kerne im NMR-Experiment aufzuzeichnen, jedoch ist die Wahrscheinlichkeit, zwei benachbarte ^{13}C-Kerne in ausreichend vielen Molekülen der Probe vorzufinden, um Kopplungen zwischen ihnen zu beobachten, beliebig gering.

Die Kopplung von Kernen und damit auch die Aufspaltung von Peaks in Multipletts kann jedoch vermieden werden. Bei Einstrahlung geeigneter Frequenzen ist es möglich, selektiv eine bestimmte Protonensorte (z.B. nur solche in CH_3-Fragmenten) zu entkoppeln. Geeignet ist eine Frequenz genau dann, wenn sie in Resonanz mit der Larmorfrequenz desjenigen Kerns ist, dessen Einfluß ausgeschaltet werden soll. Durch sie werden die betreffenden Kerne, deren Einfluß eliminiert werden soll, sehr schnell immer wieder angeregt, so daß infolge des andauernden Zustandswechsels zu keinem Zeitpunkt genau bestimmt ist, in welchem Zustand sie sich geradebefinden. Dies macht eine Kopplung unmöglich. Bei der Protonenbreitbandentkopplung in der ^{13}C-NMR-Spektroskopie wird ein Frequenzband eingestrahlt, welches den gesamten Bereich der ^{1}H-Verschiebungen umfaßt.

In Abbildung 2.23 ist ein breitbandentkoppeltes ^{13}C-NMR-Spektrum dargestellt, bei welchem die ursprünglichen Multiplettstrukturen als Buchstaben annotiert sind. Betrachtet man insbesondere den Bereich der zahlreichen nahe beieinanderliegenden Absorptionen in der Mitte, so wird deutlich, daß im Falle aufgespaltener Liniengruppen deren Zuordnung durch die Durchmischung der Multipletts sehr schwierig ist.

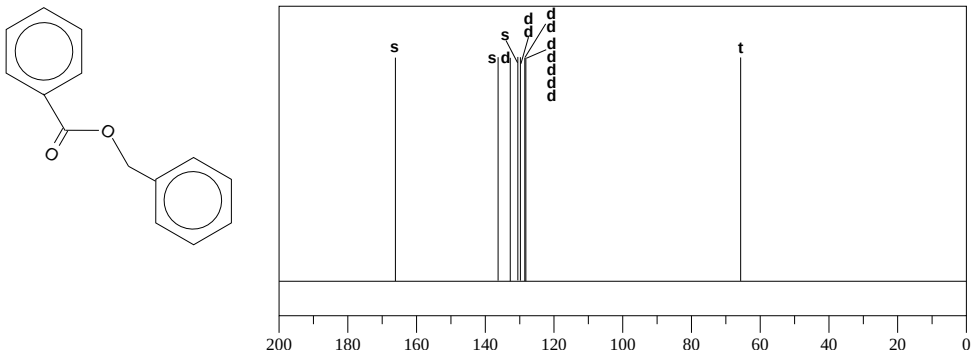


Abb. 2.23: *Struktur und Spektrum von Benzoesäurebenzylester. Im breitbandentkoppelten Spektrum sind die Multiplettstrukturen der einzelnen Peaks als Buchstaben annotiert: s=Singlett, d=Dublett, t=Triplett.*

Wenngleich natürlich bei der Breitbandentkopplung die Information über benachbarte Wasserstoffkerne verlorengeht, so bleibt doch mit der chemischen Verschiebung der einzelnen Kohlenstoffkerne (Lage der Peaks) die essentielle spektroskopische Information erhalten. Ihre Zugänglichkeit wird sogar verbessert, da es nicht zu einer Durchmischung von Multiplettlinien kommt.

Darüber hinaus hat die Entkopplung einen technischen Vorteil: Sie hat einen günstigen Einfluß auf die *Relaxationszeiten* der ^{13}C-Kerne. *Relaxation* bezeichnet die allmähliche Wie-

derherstellung der im Ruhezustand für die Besetzung der Energieniveaus geltenden BOLTZMANN-Verteilung (vgl. Gleichung (2.4)), welche durch die gezielte Anregung der Kernspins im NMR-Experiment aufgehoben wird. Die unterschiedlichen Typen von ^{13}C-Kernen haben je nach ihrer Umgebung im Molekül Relaxationszeiten zwischen einigen Millisekunden und mehreren Minuten. Die Relaxationszeiten müssen jedoch beachtet werden, bevor der nächste anregende Impuls gegeben wird: Ist die ursprüngliche BOLTZMANN-Verteilung noch nicht wiederhergestellt, so stehen weniger Kerne zur Verfügung, um den gewünschten Zustandsübergang zu vollziehen, das heißt die Beobachtungsgenauigkeit wird verringert. Der günstige Einfluß der Protonenentkopplung auf die Relaxationszeiten sorgt dabei für eine erhebliche Beschleunigung der Spektrenaufnahme.

Es ist also zusammenzufassen, daß die grundlegende Information, welche zur Aufklärung der Konstitution eines Moleküls benötigt wird, am einfachsten aus protonenbreitbandentkoppelten ^{13}C-NMR-Spektren zugänglich ist. Zum Abschluß formuliert der folgende Abschnitt nun noch einmal das Ziel der vorliegenden Arbeit vor dem Hintergrund der in diesem Kapitel dargestellten Aspekte der organischen Chemie.

2.4 Ziel der Arbeit

Ziel der Strukturaufklärung in der organischen Chemie im allgemeinen ist es, Informationen über die Gestalt eines unbekannten Moleküls zu sammeln. Spektroskopische Methoden spielen dabei eine große Rolle, insbesondere liefert die ^{13}C-NMR-Spektroskopie durch die Untersuchung von Kohlenstoffkernen Informationen über das strukturelle Skelett der fraglichen Substanz.

Protonenbreitbandentkoppelte ^{13}C-NMR-Spektren enthalten die essentielle Information hinsichtlich der Unterscheidung von Kohlenstoffatomen nach ihrer chemischen Umgebung codiert in der Lage der einzelnen Peaks. Entsprechende spektroskopische Untersuchungen sind heutzutage zeit- und kosteneffizient durchzuführen. Darüber hinaus ist es wichtig, alle verfügbaren Informationen systematisch zu nutzen und somit möglichst früh möglichst viele falsche Strukturen auszuschließen.

Hinsichtlich des Detailgrads der Strukturbeschreibung ist auf diese Weise das Ermitteln der Konstitution möglich. Betreffend die Komplexität dieser Aufgabe ist zu bedenken, daß mit wachsender Zahl beteiligter Atome mehr und mehr in Frage kommende isomere Strukturen existieren. Darüber hinaus sollte beachtet werden, daß jeder höhere Detailgrad der Strukturbeschreibung auf der Konstitution aufsetzt und diese somit als Grundlage für weitere Spezialisierungen betrachtet werden kann.

Aus Sicht der organisch-chemischen Strukturaufklärung ist es somit das Ziel der vorliegenden Arbeit, ausgehend von spektroskopischen Daten aus protonenbreitbandentkoppelten ^{13}C-NMR-Spektren Konstitutionsinformation zu gewinnen.

3 Strukturaufklärung in der Computerchemie

Im Zuge der Entwicklung automatischer Methoden wird die chemische Forschung in zunehmendem Maße mit großen Mengen experimenteller Daten konfrontiert. Die noch junge Disziplin der Chemoinformatik oder Computerchemie (*computational chemistry*, [RS98]) gewinnt daher rasant an Bedeutung: Mithilfe von Techniken der künstlichen Intelligenz, der Mustererkennung, des maschinellen Lernens und anderer Methoden der Informatik macht sie die in den Daten enthaltene Information für die verschiedenen chemischen Disziplinen zugänglich.

Auch in der Strukturaufklärung führt der technische Fortschritt und insbesondere die Automatisierung in den Bereichen der chemischen Synthese und der Analytik zu einem immens hohen Datenaufkommen. Dadurch entsteht ein neuer Engpaß im Bereich der Auswertung der nun sehr rasch anfallenden großen Datenmengen. In diesem Bereich befassen sich aktuelle Forschungen der Computerchemie damit, die erhobenen Daten insbesondere spektroskopischer Untersuchungen so weit wie möglich automatisch aufzubereiten und auszuwerten.

Im folgenden sollen zunächst in Abschnitt 3.1 Methoden beschrieben werden, welche in der Computerchemie typischerweise zur Aufklärung einer unbekannten Molekülstruktur zum Einsatz kommen. Abschnitt 3.2 schließt sich mit einem kurzen Überblick über aktuelle Strukturaufklärungssysteme an. Schließlich stellt Abschnitt 3.3 einen alternativen Ansatz vor, welcher Ideen zur Beseitigung bestehender Schwachpunkte aufgreift und dessen Untersuchung Gegenstand dieser Arbeit ist. Abschnitt 3.4 beleuchtet zuletzt deren Ziele aus dem Blickwinkel der Computerchemie.

3.1 Methoden

Ein gängiger Ansatz der Strukturaufklärung, sowohl historisch wie auch in der modernen organischen Chemie, verläuft in zwei Schritten (vgl. [MMWM02, Mun98]): Erst wird eine (große) Menge möglicher Strukturen ermittelt (z.B. basierend auf der Summenformel) und dann jede einzelne dieser Hypothesen validiert, indem eine geeignete Eigenschaft des betreffenden Kandidaten mit einem gegebenen experimentellen Befund der unbekannten Substanz verglichen wird.

Besonders geeignet ist das ^{13}C-NMR-Spektrum der unbekannten Substanz. NMR-Spektren sind heute mit einem vergleichsweise geringem Aufwand an Zeit, Geld und Material zugänglich und besitzen einen direkten Bezug zu strukturellen Eigenschaften des Moleküls. Die Kohlenstoffatome, welche im ^{13}C-NMR-Experiment untersucht werden, sind zudem wenig an intermolekularen Wechselwirkungen beteiligt, so daß die im Spektrum enthaltene strukturelle Information sich vorrangig auf das Molekül selbst bezieht und nicht auf dessen Interaktion mit Nachbarmolekülen der Probe oder des Lösemittels. (Gleichwohl können solche Interaktionen das Spektrum beeinträchtigen, vgl. Abschnitt 2.1.4.) Da der Zusammenhang zwischen Spektrum und Struktur bekannt ist, ist es außerdem prinzipiell möglich, bei gegebener Struktur das zu erwartende Spektrum vorherzusagen (vgl. Abschnitt 3.1.2).

Theoretisch könnte im nächsten Schritt durch Vergleich des Spektrums der unbekannten Substanz mit den berechneten Spektren der Strukturkandidaten die tatsächliche Struktur ein-

deutig bestimmt werden. In der Praxis sind jedoch die dafür nötigen Voraussetzungen nicht erfüllt: Zum einen würden unbegrenzte Rechenressourcen zur Vorhersage des NMR-Spektrums mit beliebiger Genauigkeit benötigt, und zum anderen müßte auch der experimentelle Befund mit derselben Genauigkeit, das heißt frei von jeglichen Unsicherheiten, zur Verfügung stehen.

Systeme, die das oben angedeutete zweischrittige Schema umsetzen, liefern daher nicht eine einzelne Lösung, sondern sie bewerten alle Strukturkadidaten und liefern eine dieser Bewertung entsprechend sortierte Liste der in Frage kommenden Lösungen. Aufgrund der Rechen- und Meßungenauigkeiten kann dabei nicht automatisch angenommen werden, daß die bestbewertete Struktur die tatsächlich korrekte ist, jedoch kann die Verläßlichkeit der Hypothesenbewertung durch einen hohen Informationsgehalt und bestmögliche Qualität der zur Verfügung gestellten Daten erhöht werden. Um darüber hinaus mit den beschränkten Rechenressourcen auszukommen, welche sowohl die Qualität der Ergebnisse als auch die Geschwindigkeit bei deren Ermittlung beeinflussen, bedarf es einer intelligenten Umsetzung geeigneter Methoden vor allem aus den Bereichen der Datenverarbeitung, Musterklassifikation und Musteranalyse.

Heutige Systeme liefern zuverlässige Ergebnisse in akzeptabler Zeit, wobei jedoch teilweise recht unterschiedliche Ansätze verfolgt werden. Im folgenden werden die beiden grundlegenden Aspekte der Hypothesengenerierung und Validierung, eine mögliche Analyse des Ergebnisraumes sowie die Frage einer geeigneten Repräsentation von Molekülstrukturen näher betrachtet, bevor sich in Abschnitt 3.2 ein Überblick über die in aktuellen Systemen eingesetzten Methoden anschließt, bei welchem der Schwerpunkt auf den verwendeten Konzepten und den Ansprüchen an die betreffenden Systeme liegt.

3.1.1 Strukturgeneratoren

Zur Hypothesengenerierung werden in der Strukturaufklärung oft *Strukturgeneratoren* eingesetzt. Ein Strukturgenerator ist ein Programm, welches gegeben eine bestimmte Ausgangsinformation alle passenden Molekülstrukturen liefert (vgl. [Wie97, BKLW96, BKL96]). Diese können als mathematische Modelle des betreffenden Moleküls aufgefaßt werden. Um sie zu berechnen, werden Methoden aus den Bereichen der Gruppentheorie, Kombinatorik und Graphentheorie in einem spezialisierten Computeralgebrasystem vereinigt. Seine Aufgabe ist es, effizient, redundanzfrei und vollständig die mit der gegebenen Information übereinstimmenden Modelle zu liefern.

Dabei gibt es, wie in Abschnitt 2.1.2 (Seite 8) beschrieben, vier prinzipielle Abstraktionsstufen im Detailgrad der Strukturbeschreibung. Am Beispiel von Benzol als einem relativ kleinen Molekül, das nur sechs *schwere Atome* (nicht-Wasserstoff-Atome) besitzt, soll die wachsende Komplexität der Unterscheidungen auf jeder dieser Stufen verdeutlicht werden.

1. Die Summenformel, im Falle von Benzol C_6H_6, dient in der Regel als Eingabe.

2. Hinsichtlich der Konstitution des Moleküls (Lage und Art der Bindungen) lassen sich von C_6H_6 ausgehend 217 Isomere unterscheiden, von welchen Benzol nur eines ist.

3. Unter Einbeziehung räumlicher Aspekte gelangt man zu insgesamt 958 Varianten (Stereo- oder Konfigurationsisomeren) der 217 Bindungsisomere, die sich hinsichtlich der relativen räumlichen Anordnung einzelner Gruppen unterscheiden.

4. Die Menge aller Konformationen (energetisch verschiedener räumlicher Anordnungen der Atome) schließlich ist nicht diskret, da viele der unterschiedlichen Konformationen ein und derselben Konfiguration durch Drehung in einander überführbar sind.

Heutige Strukturgeneratoren sind in der Lage, alle mit der Eingabe übereinstimmenden Konstitutions und Konfigurationsisomere zu berechnen sowie die diskrete Klassifizierung von Konformationen zu leisten (vgl. [BKLW96]). Sie werden in unterschiedlichen Bereichen der Forschung und Lehre eingesetzt. Im folgenden werden die Herausforderungen des Anwendungsfeldes von Strukturgeneratoren im Allgemeinen beschrieben und die zugrundeliegenden mathematischen Prinzipien kurz skizziert.

In eine algebraische Terminologie übertragen sind die Bindungsisomere einer gegebenen Summenformel alle zusammenhängenden *Multigraphen* (Graphen mit potentiell mehreren Kanten zwischen denselben Knoten) mit der *Eckengradfolge*, die sich aus der Summenformel ergibt und deren Knoten mit Atomnamen auf alle wesentlich verschiedenen Weisen *gefärbt* sind. Eine *Färbung* ordnet jedem Knoten des Graphen eine natürliche Zahl zu, welche im Falle von Molekülstrukturen das chemische Element des betreffenden Atoms codiert. Der Eckengrad eines Knotens gibt die Zahl mit ihm verbundener Kanten an, wobei Verbindungen mit sich selbst (*Schleifen*) doppelt gezählt werden. Mit Eckengradfolge wird die (absteigend) geordnete Liste aller Eckengrade eines Graphen bezeichnet; auf das Beispiel von Benzol, C_6H_6, übertragen ergibt sich die Eckengradfolge $(4^6, 1^6)$ für sechs (Kohlenstoff-)Atome der Valenz 4 und sechs (Wasserstoff-)Atome der Valenz 1.

Die Aufgabe eines Strukturgenerators besteht in diesem Sinne darin, zu einer gegebenen Summenformel alle zusammenhängenden (ungerichteten, schleifenfreien Multi-)Graphen zu konstruieren, die eine den Valenzen und Zahlen der Atome entsprechende Eckengradfolge besitzen und die alle etwaigen Zusatzbedingungen, wie etwa das Vorhandensein einer bestimmten funktionellen Gruppe, erfüllen. Die gewünschte Lösungsmenge enthält alle Färbungen dieser Graphen mit den vorgegebenen Atomtypen.

Die Ergebnismenge einer Anfrage an den Strukturgenerator kann jedoch bereits für verhältnismäßig kleine Moleküle sehr groß werden [BKL96]. Um auf effiziente Weise wie gefordert eine vollständige und redundanzfreie Ergebnismenge zu erhalten werden meist auf der *ordnungstreuen Erzeugung* von R. C. READ (READ*'s orderly generation*, [Rea78, CR79]) basierende Verfahren eingesetzt und mit schrittweisen Verfeinerungen kombiniert. So erreicht man eine systematische Erzeugung der gesuchten Graphen, welche frühzeitig abgebrochen werden kann, sobald sich eine Struktur nicht mehr zu einem den Vorgaben entsprechenden Ergebnis erweitern läßt.

Darüber hinaus kann der Generierungsprozeß durch die Vorgabe von *Makroatom*en erleichtert werden. Ein Makroatom ist ein Strukturfragment, das während der Generierung wie ein einzelnes Atom behandelt wird (beispielsweise kann ein Benzolring als ein Makroatom der Valenz 6 angegeben werden).

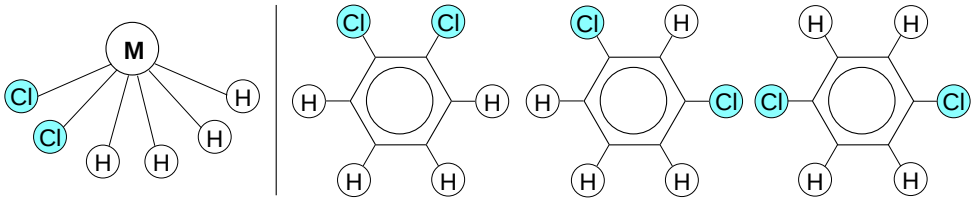


Abb. 3.1: *Gegeben ist die Summenformel $C_6H_4Cl_2$ sowie ein Benzolring als Makroatom. Im Graphen links ist das Makroatom mit M bezeichnet. Rechts sind die drei durch den Graphen zusammengefaßten Molekülstrukturen dargestellt.*

Variationen der an das Makroatom gebundenen Substituenten, die sich nur durch ihre relative Anordnung unterscheiden, fallen bei der Generierung zusammen, da bei der Repräsentation als Makroatom die einzelnen Positionen des tatsächlichen Strukturelements nicht unterscheidbar sind. Die einzelnen Strukturen erhält man durch anschließende Erweiterung des Makroatoms zur vollständigen Substruktur (vgl. [BKLW96, RHR70, RK83]). Zur Verdeutlichung dient Abbildung 3.1.

Auch bei der Speicherung der oft sehr umfangreicher Ergebnisse ist ein intelligentes Vorgehen nötig (vgl. [Jer86]). Eine kanonische Form der Ergebnisstrukturen ermöglicht den Einsatz von *Hashing*, das heißt die eindeutige Abbildung der komplexen Datenstrukturen (Molekülgraphen) auf skalare Werte (*Schlüssel*). Dadurch muß bei der Betrachtung der Einzellösungen keine Überprüfung auf Isomorphie durchgeführt werden, sondern verschiedene Schlüssel sind gleichbedeutend mit der Verschiedenheit der zugehörigen Strukturen.

Im Zusammenhang der Strukturaufklärung spielt zudem die Einbringung von Zusatzinformation eine wichtige Rolle. Ziel der Angabe derartiger Nebenbedingungen ist es, einen möglichst großen Teil der uninteressanten Resultate zu unterdrücken. Typischerweise werden *Positivliste*n oder *Negativliste*n verwendet (vgl. [BKLW96]), auch die Vorgabe von Makroatomen, quasi als zusätzliche Positivliste, kann in diesem Zusammenhang gesehen werden. Positivlisten sind Listen von Strukturelementen, welche in den gewünschten Lösungen vorhanden sein müssen, Negativlisten sind ebensolche Listen verbotener Strukturelemente. Da die innerhalb einer Liste aufgezählten Substrukturen sich überschneiden dürfen, findet in beiden Fällen, anders als im Falle von Makroatomen, eine Überprüfung der generierten Lösungen erst im Nachhinein statt (vgl. auch Abschnitt 3.1.3). Darüber hinaus bieten manche Systeme (z.B. MOLGEN [BKL96, BGH$^+$95]) die Möglichkeit, die Lösungsmenge durch interaktive Beschränkung des Strukturenraums mit einem Strukturediteor einzugrenzen.

Zusammenfassend läßt sich für die Anwendung im Zusammenhang der Strukturaufklärung sagen, daß neben dem allgemeinen Problem der redundanzfreien, jedoch vollständigen Generierung des Strukturenraums dessen enorme Größe ein kritischer Faktor ist. Die Verwendung möglichst aller vorhandenen Zusatzinformationen zum frühestmöglichen Zeitpunkt ist daher anzustreben, mit dem Ziel, den Strukturenraum derart zu begrenzen, daß nur die tatsächlich interessanten Strukturhypothesen generiert werden.

3.1.2 Spektrenvorhersage

NMR-Spektren sind heutzutage experimentell ohne großen Aufwand zu erhalten. Dank der bekannten Zusammenhänge zwischen Spektrum und Struktur können theoretisch berechnete Spektren von Strukturhypothesen im Vergleich mit dem experimentellen Spektrum der unbekannten Substanz zur Hypothesenvalidierung herangezogen werden. Zur Realisierung der Spektrenvorhersage bei gegebener Struktur gibt es verschiedene Ansätze, welche unterschiedliche Gewichtungen von Vorhersagegenauigkeit, Rechengeschwindigkeit und allgemeiner Anwendbarkeit erreichen. Ein Überblick über diese soll im folgenden gegeben werden.

Wie bereits in Abschnitt 2.1.3 erwähnt gibt es für Mehrelektronensysteme keine exakte algebraische Lösung der SCHRÖDINGER-Gleichung, das heißt eine exakte Berechnung der Orbitalenergie und damit der Larmorfrequenz ist nur für ein einzelnes Wasserstoffatom (Ein-Elektronen-System) möglich (vgl. [Atk96], S. 408 ff.). Ausgehend von der Vorstellung, daß die gesuchten Orbitale in Systemen mit mehreren Elektronen aus wasserstoffähnlichen Anteilen aufgebaut sind, erreicht man jedoch gute Näherungen. Dieses Vorgehen wird, im Gegensatz zu empirischen Methoden, als *ab initio* Berechnung bezeichnet. Das quantenphy-

sikalische bzw. quantenchemische Modell erlaubt dabei nicht nur die Bestimmung der chemischen Verschiebung, sondern die Berechnung verschiedener kernmagnetischer Parameter. Systeme, die mit *ab initio* Methoden arbeiten, sind zum Beispiel GAUSSIAN [CF00, gau04] oder COSMOS [Wit03, KLS03].

Eine nicht zu unterschätzende Schwierigkeit hierbei ist jedoch, daß die genaue dreidimensionale Struktur des Moleküls bekannt sein muß. Geht man von der Anwendung aus, zu einem einzelnen gegebenen Molekül das zugehörige NMR-Spektrum berechnen zu wollen, so scheint dies zunächst nur ein geringeres Problem zu sein, auch hier jedoch ist ein vergleichsweise hoher Aufwand nötig, da neben der Konstitution auch die Konfiguration des Moleküls bekannt sein muß und da verschiedene Konformationen eine Rolle spielen können. All diese Faktoren müssen mit berücksichtigt werden.

Im Anwendungsfeld der Strukturaufklärung ist jedoch in der Regel nicht nur das Spektrum einer einzigen Struktur, sondern die Spektren sehr vieler Strukturhypothesen zu bestimmen. Selbst ein kleines Molekül wie Benzol besitzt bereits über 200 Konstitutionsisomere, berücksichtigt man deren unterschiedliche Konfigurationen, so sind es an die 1000 Isomere. Aufgrund der hohen Anforderungen an die Eingangsdaten sind *ab initio* Methoden also für den Einsatz im Rahmen der organisch-chemischen Strukturaufklärung weniger geeignet, da die nötigen Berechnungen für den erforderlichen Detailgrad der Strukturbeschreibung den Gesamtprozeß aufgrund der Menge der Strukturhypothesen sehr rechenzeitintensiv machen. In der Regel kommen daher andere Verfahren zum Einsatz, wobei Schnelligkeit und Genauigkeit stets gegeneinander abgewogen werden.

Inkrementmethoden folgen in Grundzügen einer verwandten Idee: Ausgehend von der Annahme, daß die Einflüsse einzelner Strukturfragmente auf die Verschiebung eines bestimmten Chromophors additiv sind, kommen hier Tabellen zum Einsatz, welche die Inkremente bestimmter Gruppen in einer bestimmten relativen Position auf einen bestimmten Grundtypus von Kern angeben – beispielsweise die Einflüsse von Substituenten in den in Abschnitt 2.2.2 beschriebenen Positionen auf die Kohlenstoffatome eines Benzolringes, wie in Abbildung 3.2 beispielhaft für ein Benzolderivat dargestellt. Eine zu berechnende chemische Verschiebung s setzt sich aus der Grundverschiebung σ_0 des betreffenden Kohlenstofftypus und der Summe aller für die vorhandenen Substituenten i erforderlichen Inkrementterme σ_i sowie einem Korrekturterm K für sterische Effekte zusammen:

$$s = \sigma_0 + \sum_i \sigma_i + K \tag{3.1}$$

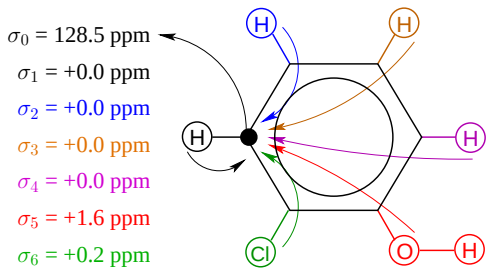


Abb. 3.2: *Die Einflüsse benachbarter Strukturfragmente auf die Verschiebung eines Chromophors sind additiv und ergeben zusammen mit einer vom Typ des Chromophors abhängigen Grundverschiebung die letztendliche Lage des zugehörigen Peaks.*

Der Vorteil dieser Methode ist ihre Schnelligkeit, welche das Verfahren auch für große Mengen von Strukturhypothesen einsetzbar macht. Die Systeme CHEMDRAW und SPECTOOL beinhalten das Paket CHEMNMR [CF00], welches den Inkrementansatz verwendet, jedoch gegenüber Gleichung (3.1) mit einer wichtigen Verfeinerung: Besonders an mehrfachsubstituierten bzw. verzweigten Systemen ist es wichtig, die Interaktion der Substituenten miteinander mit einzubeziehen. Dies erreicht man durch die Verwendung von Kreuztermen σ_{ij} neben den reinen Inkrementtermen:

$$s = \sigma_0 + \sum_i \sigma_i + \sum_{(i,j)} \sigma_{ij} + K \tag{3.2}$$

Bei allen Inkrementtermen handelt es sich jedoch um empirisch ermittelte Werte. Daher wird es um so schwieriger, fundierte und verläßliche Werte für die Kreuzterme zu allen möglichen Kombinationen der betrachteten Gruppen zu ermitteln, je mehr Substituenten an solchen Interaktionen beteiligt sind. Darüber hinaus ist die Genauigkeit dieses Verfahrens stets von der Qualität der tabellarisch gespeicherten Daten abhängig, insbesondere liegt es in der Natur der Sache, daß keine Inkremente für unbekannte Gruppen zur Verfügung stehen können. Auch der Detailgrad in der Unterscheidung von Substituenten spielt natürlich eine Rolle.

Die bevorzugte Methode der Spektrenvorhersage sind heutzutage *neuronale Netze*, da sie große Schnelligkeit und Präzision auf sich vereinen. Ein neuronales Netz wird trainiert, indem ihm eine große Menge von Eingabedaten zusammen mit den gewünschten Ausgabewerten präsentiert wird. Die internen Netzparameter werden mit jedem präsentierten Beispiel angepaßt. Nach dem Training ist das System unabhängig von experimentellen Daten, so daß es prinzipiell für Verbindungen beliebiger Stoffklassen genutzt werden kann. Gleichwohl kann durch die Wahl einer entsprechenden Trainingsstichprobe eine Spezialisierung auf einen bestimmten Einsatzbereich erfolgen.

Die Systeme C_SHIFT [MMW00] und ANALYZE [MM02] arbeiten mit einer auf neuronalen Netzen basierenden Spektrenvorhersage. In der Regel werden in solchen Systemen mehrere neuronale Netze eingesetzt, die auf einzelne Typen von Kohlenstoffatomen spezialisiert sind. Die Ausgabe eines Netzes ist stets nur eine einzelne chemische Verschiebung; das Gesamtspektrum wird im Anschluß aus den Einzelpeaks zusammengesetzt. Wie bei den zuvor beschriebenen Inkrementmethoden spielt auch hier der Detailgrad der Betrachtung eine Rolle, und zwar hinsichtlich der Strukturbeschreibung der Eingabedaten: Je detaillierter die Umgebung des Kohlenstoffatoms beschrieben wird, dessen chemische Verschiebung zu berechnen ist, desto genauer werden auch die Einflüsse aus dieser Umgebung berücksichtigt.

Insgesamt ist die Vorhersage von NMR-Spektren ein wichtiger Punkt in der automatischen Strukturaufklärung. Wird sie zur Hypothesenvalidierung eingesetzt, so ist aufgrund des zu erwartenden großen Umfangs der Hypothesenmenge die Schnelligkeit der eingesetzten Methode neben ihrer Präzision ein entscheidender Faktor.

3.1.3 Substrukturanalyse

Neben der Validierung bzw. Bewertung von Strukturhypothesen ist die *Substrukturanalyse* eine interessante Möglichkeit, sich die Gesamtmenge der Strukturkandidaten nutzbar zu machen. Ihr Ziel ist es, Aufschluß darüber zu erhalten, worin sich Untermengen der Gesamthypothesenmenge unterscheiden (vgl. [Mea02]). Sie kann in Kombination mit der Validierung oder unabhängig von dieser eingesetzt werden.

Die Substrukturanalyse einer gegebenen Menge von Strukturen gibt Aufschluß über die Unterschiede der einzelnen Elemente dieser Menge (*Clustering*). Sie ist ein hartes Problem,

da sie sich mit dem paarweisen Vergleich aller Moleküle befaßt und dabei das größte gemeinsame Strukturfragment sucht. Bei jedem paarweisen Vergleich werden zuerst zwei übereinstimmende Atome in den beiden verglichenen Strukturen mit einander assoziiert, und diese (initial einatomige) Substruktur wird dann inkrementell um weitere übereinstimmende Atome erweitert. Da das Ergebnis initialisierungsabhängig ist, muß über alle möglichen initialen Atompaare iteriert werden; ebenso sind Iterationen über alle möglichen Folgeassoziationen bei der Erweiterung der Teilstruktur nötig.

Das Ergebnis der Unterteilung der Gesamtstrukturenmenge kann durch Angaben wie etwa eine Mindestzahl von Atomen in jeder Substruktur oder eine Mindestzahl von enthaltenden Molekülen, um eine bestimmte Substruktur zu berücksichtigen, parametrisiert werden. Günstig ist nach der Analyse eine Organisation der gefundenen Substrukturen als Baum, welcher das Enthaltensein kleinerer Fragmente in größeren wiedergibt. Man erhält so eine Unterteilung der gegebenen Strukturen in Strukturfamilien, wobei die Familien um so allgemeiner gefaßt ist, je kleiner das gemeinsame Strukturfragment ist.

Im Kontext der Strukturaufklärung ermöglicht es eine solche Untergliederung der Hypothesenmenge, gezielt ganze Gruppen von Kandidaten aufgrund gemeinsamer Eigenschaften in den weiteren Prozeß eingehen zu lassen oder davon auszuschließen. Auf einer sehr allgemeinen Definitionsebene von Strukturfamilien ist z.B. ein Ausschluß möglich, wenn ein Peak in einem für eine bestimmte funktionelle Gruppe typischen Bereich des Spektrums fehlt. Auch kann etwaiges Vorwissen genutzt werden, um bestimmte Substrukturen als erforderlich oder verboten zu kennzeichnen (Positiv- oder Negativlisten, vgl. Abschnitt 3.1.1). Die betreffenden Strukturfamilien werden dann aus der weiteren Betrachtung entfernt (Negativlisten) bzw. ausschließlich betrachtet (Positivlisten). Durch Kombination mit der Spektrenvorhersage kann dann innerhalb der verbleibenden Strukturfamilien eine Bewertung der Hypothesen vorgenommen werden, wobei das Verfahren durch den aufgrund der Vorauswahl verkleinerten Strukturenraum beschleunigt wird.

Ebenso ist es umgekehrt auch denkbar, eine Vorauswahl der Kandidaten für die Substrukturanalyse anhand ihrer Abweichungen vom experimentellen Spektrum zu treffen. Eine weitere Möglichkeit besteht darin, die Spektrenvorhersage zu nutzen, um anstelle der genauen Struktur die wahrscheinlichste Strukturfamilie zu ermitteln. In diesem Fall wird die mittlere Abweichung der berechneten Spektren aller Strukturen einer Familie vom experimentellen Spektrum der unbekannten Substanz berechnet, ein Vorgehen, das aus statistischen Gründen besonders robust ist, da die Schätzfehler bei der Vorhersage der einzelnen Spektren sich durch die Mittelung eliminieren.

3.1.4 Integration von Hypothesengenerierung und Validierung

Der Schwachpunkt des klassischen zweischrittigen Vorgehens ist die Generierung eines sehr großen und kaum überschaubaren Raumes von Molekülstrukturen, welcher in einem Folgeschritt vollständig nach der korrekten Struktur der unbekannten Verbindung durchsucht werden muß. Eine möglichst frühzeitige Suchraumbeschränkung ist somit der Hauptansatzpunkt für Verbesserungen. Hierzu ist es naheliegend, die beiden Schritte der Hypothesengenerierung und der Validierung nicht einzeln zu betrachten, sondern mit einander zu verflechten. Im folgenden sollen hier zwei Methoden vorgestellt werden, welche auf Datenbanken sowie auf genetischen Algorithmen basieren und diese Idee umsetzen.

In *Spektrendatenbanken* (vgl. z.B. [Spe98]) wird der Strukturenraum organisch-chemischer Verbindungen durch die gespeicherten Beispiele repräsentiert. Ist eine Zuordnung der Peaks in jedem Spektrum zu den sie verursachenden Atomen des korrespondierenden Mole-

küls gegeben, so kann die Datenbank genutzt werden, um eine SSC-Bibliothek (*subspectra-substructure-correlation* Bibliothek) [WFR96, WR97] aufzubauen. In einer solchen Bibliothek enthält jeder Datensatz ein Subspektrum und die Konnektivitätsinformation der dazugehörigen Substruktur. Jedes Kohlenstoffatom wird dabei hinsichtlich seiner molekularen Umgebung genau beschrieben, und im Subspektrum ist für seine chemische Verschiebung der quadratische Mittelwert angegeben, der sich aus all denjenigen Spektren in der Datenbank ergibt, deren verursachende Moleküle die betreffende Substruktur enthalten.

Die Grundidee dieses Ansatzes zur Repräsentation des Strukturenraumes ähnelt in gewisser Weise denen der Methoden zur Spektrenvorhersage. Da hier nicht einzelne Kohlenstoffatome und deren zugehörige Peaks, sondern Substrukturen und ihre zugehörigen Subspektren betrachtet werden, findet zudem implizit eine Berücksichtigung der wechselweisen Einflüsse innerhalb der jeweiligen chemischen Umgebung statt. Einflüsse, die von außerhalb dieses Strukturfragments herrühren, werden dabei durch die Mittelung über alle Spektren, deren zugehörige Strukturen eine bestimmte Substruktur enthalten, eliminiert. Sofern sie aus bekannten Substrukturen aufgebaut sind, ist auch die Betrachtung von Verbindungen, die selbst nicht in der Datenbank enthalten sind, möglich.

Im Anwendungsfeld der Strukturaufklärung werden SSC-Bibliotheken wie folgt genutzt: Zuerst wird nach Ähnlichkeiten zwischen dem experimentellen Spektrum und den Subspektren gesucht. Aus den auf diese Weise gefundenen Substrukturen wird nach und nach die Gesamtmolekülstruktur aufgebaut, wobei die Information überlappender Teilfragmente genutzt wird. Eine Spektren- bzw. Subspektrenvorhersage findet bei jedem Aufbauschritt eines größeren Strukturfragments zur Validierung statt (vgl. [WFR96, WR97]). Somit sind Hypothesengenerierung und Validierung in diesem Ansatz stark vernetzt. In der Regel liefert diese Methode nur sehr wenige in Frage kommende Strukturhypothesen.

Wie bei der Spektrenvorhersage als unabhängigem Validierungsschritt ist auch hier die Detailtiefe der Strukturbeschreibung von Bedeutung für die Genauigkeit des Verfahrens. Kritisch ist jedoch darüber hinaus die Qualität der zugrundegelegten Daten: Da es sich um reale, experimentelle Daten handelt, können systematische Fehler im NMR-Experiment, Verunreinigungen, fehlerhafte Peakzuordnungen und dergleichen die Genauigkeit erheblich beeinträchtigen. Daher ist ein relativ hoher Wartungsaufwand zur Qualitätssicherung notwendig. Unter dieser Voraussetzung jedoch ist die Präzision des Verfahrens sehr hoch, da prinzipiell mit jedem NMR-Experiment, das durchgeführt wird, neue Daten eingetragen oder bestehende untermauert werden können.

Als zweite Methode, die eine Verflechtung des Hypothesengenerierungs- und des Validierungsschrittes realisiert, setzt ein auf genetischen Algorithmen basierendes Verfahren [MW01, MW02] auf eine dynamische Generierung des Strukturenraums. *Genetische Algorithmen* orientieren sich an der biologischen Evolution; so wird hier jede Molekülstruktur wird als ein Individuum einer Population angesehen und als „Gen“ codiert. Von Generation zu Generation soll nun diejenige „Erbinformation“ erhalten bleiben, die an die Umgebungsbedingungen am besten angepaßt ist. Dies wird mit Hilfe einer sogenannten *Fitness-Funktion* bewertet. Eine neue Generation entsteht, indem die Individuen der vorhergehenden entweder geklont (direkt übernommen), mutiert (leicht variiert) oder rekombiniert werden. Je besser der Fitness-Wert eines Individuums, desto eher kommt er für Klonierung und Rekombination in Frage, um dadurch „gute“ Erbinformation für die nächste Generation zu konservieren. Weniger optimale Individuen haben eine steigende Mutationswahrscheinlichkeit; durch diese Variation werden neue Hypothesen in die Population eingebracht.

Klonierungs-, Mutations- und Rekombinationsrate sind Parameter des Systems, welche zu optimieren sind, ebenso wie die Größe der Population. Darüber hinaus stellt sich die Frage, wann die Iterierung weiterer Generationen beendet wird. Wie bereits zu Beginn von Abschnitt 3.1 erwähnt sind im Zusammenhang mit der Bewertung von Strukturhypothesen Unsicherheiten durch fehlerbehaftete Spektrenaufnahme und Spektrenvorhersage von Bedeutung. Dies spielt auch hier eine Rolle, da die Abweichung des vorhergesagten Spektrums eines Individuums von dem experimentellen Spektrum der unbekannten Substanz für die Fitness-Funktion verwendet wird. Daher kann nicht die exakte Übereinstimmung von berechnetem und experimentellem Spektrum als Abbruchkriterium herangezogen werden. Stattdessen kann beispielsweise eine feste Zahl von Generationen, ein Schwellwert für die Fitnessfunktion oder ein absolutes Zeitlimit vorgegeben werden.

Allerdings liegt es in der Natur genetischer Algorithmen als einer nichtdeterministischen Methodik, daß das Finden des globalen Optimums (in diesem Falle der korrekten Struktur) nicht garantiert werden kann. Darüber hinaus ist der Ansatz initialisierungsabhängig, und zahlreiche empirisch zu optimierende Parameter haben Einfluß auf seine Leistungsfähigkeit. Hier wäre eine Vorgabe für eine gute Parametrisierung wünschenswert, die auf theoretisch-wohlfundierter Ebene nachvollziehbar ist.

3.1.5 Strukturbeschreibung mit Hilfe des HOSE-Codes

Neben der konkreten Strategie ist die eindeutige, systematische Beschreibung von Strukturen ein wichtiger Aspekt der Strukturaufklärung in der Chemoinformatik. Aus diesem Bedarf heraus wurde der der *HOSE-Code* (engl. *hierarchically ordered system of spherical environments*: „hierarchisch geordnetes System sphärischer Umgebungen“ [Bre78]) entwickelt, um Strukturteile, die zu einem bestimmten spektralen Befund korrespondieren, eindeutig zu beschreiben und kompakt zu repräsentieren. Er kam zuerst in Spektrendatenbanken zum Einsatz, wo eine HOSE-Code-basierte Ähnlichkeitssuche durchgeführt werden kann, dient aber ebenso zur Codierung der Eingabedaten (Strukturen) von neuronalen Netzen.

Molekülstrukturen oder Teilstrukturen werden im HOSE-Code in sogenannte *Sphären* untergliedert beschrieben. Das Atom im Fokus der Betrachtung besetzt die nullte Sphäre. Die erste Sphäre enthält seine direkten Nachbarn, die zweite Sphäre all diejenigen Atome, die zwei Bindungen entfernt liegen und so weiter. Abbildung 3.3 verdeutlicht dies an einem beispielhaften Benzolderivat.

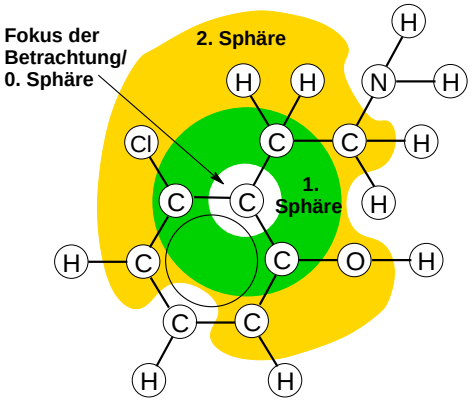


Abb. 3.3: *Der HOSE-Code beschreibt Molekülstrukturen untergliedert nach Sphären. Mit dem gekennzeichneten Atom als Fokuspunkt (nullte Sphäre) gehören die Atome im grünen Bereich zur ersten, die im gelben Bereich zur zweiten Sphäre.*

Durch klare Prioritätsregeln erlaubt es der HOSE-Code, Strukturen oder Substrukturen eindeutig wiederzugeben. Bestimmte Zeichen (Buchstaben, Zahlen, Klammern usw.) repräsentieren dabei bestimmte Atom- und Bindungstypen, Ringschlüsse und die aktuell beschriebene Sphäre. Die Genauigkeit dieser Repräsentation ist von der maximalen Länge des Codes abhängig. Ist sie beliebig (was jedoch in den genannten Anwendungen nicht der Fall ist), so kann auch jedes beliebige Molekül exakt repräsentiert werden. Bei beschränkter Länge werden alle Sphären so weit wie möglich exakt wiedergegeben; für die übrigen wird als letzte eine „Summensphäre“ eingeführt.

3.2 Stand der Forschung

Wie bereits beschrieben verfolgt die Computerchemie bei der Strukturaufklärung meist den Ansatz, zunächst eine Menge von Strukturhypothesen zu generieren und diese dann durch den Vergleich des theoretischen NMR-Spektrums jeder Hypothese mit dem experimentellen Spektrum der unbekannten Substanz zu bewerten. Spektrenaufnahme und Spektrenberechnung sind jedoch mit Unsicherheiten behaftet, so daß das bestbewertete Ergebnis nicht automatisch als die korrekte Lösung angenommen werden kann.

Ein Problem ist dabei die Größe des Strukturenraums. Hier gibt es zum einen Ansätze, den Umfang der Hypothesenmenge vor dem Validierungsschritt z.B. mittels Positiv- oder Negativlisten oder auch interaktiv zu verringern, und zum anderen die Idee, sich den Gesamtraum möglicher Strukturen durch eine Substrukturanalyse (Clustering) nutzbar zu machen. Vielversprechender erscheint jedoch der Gedanke, die beiden Schritte der Hypothesengenerierung und Validierung nicht unabhängig, sondern integriert zu betrachten.

Maßgeblich für die Wahl der in Abschnitt 3.1 beschriebenen Methoden und deren Kombination in einem Strukturaufklärungssystem ist jeweils der Anspruch, welchem das resultierende System genügen soll. Die Ressourcen an Rechenzeit und Speicherplatz, Detailgrad der Strukturbetrachtung und die Art der verfügbaren Ausgangsdaten spielen dabei ebenso eine Rolle wie die erforderliche Verläßlichkeit des Systems.

Im folgenden werden nun exemplarisch einige Systeme vorgestellt, die in der Computerchemie zur Strukturaufklärung eingesetzt werden. Sie werden bezüglich der eingesetzten Methoden gegeneinander abgegrenzt und ihre Leistungsfähigkeit sowie ihre Anforderungen an die Eingabedaten dargelegt. Hierdurch entsteht ein Überblick über die Möglichkeiten heutiger Systeme und somit auch der Erwartungen, welchen ein neues System zur Strukturaufklärung heutzutage gerecht werden muß.

3.2.1 MOLGEN und ANALYZE: Strukturvalidierung mittels neuronaler Netze

Gegeben sind zu Beginn Summenformel und experimentelles ^{13}C-NMR-Spektrum einer unbekannten Verbindung. Der Strukturgenerator MOLGEN [BGH^{+}95] bestimmt bis auf Konfigurationsebene alle Isomere der gegebenen Summenformel. Er liefert dabei effizient eine vollständige, redundanzfreie Auflistung der zugehörigen Strukturen. In dem Programm ANALYZE [MM02] dienen neun neuronale Netze für unterschiedliche Typen von Kohlenstoffatomen dazu, zu den von MOLGEN gelieferten Strukturen ^{13}C-NMR-Spektren zu berechnen. Mit Hilfe der mittleren quadratischen Abweichung zwischen dem experimentellen Spektrum und den vorhergesagten Spektren werden dann die Hypothesen bewertet. Das Ergebnis ist eine nach Wahrscheinlichkeit geordnete Liste aller in Frage kommenden Konstitutions- und Konfigurationsisomere der gegebenen Summenformel.

Die Autoren Meiler und Mehringer beschreiben die Anwendbarkeit ihres Verfahrens für Moleküle mit bis zu zwölf schweren (das heißt Nicht-Wasserstoff-)Atomen [MM02]. Die Rechenzeit wird mit 6 Sekunden bis über 2 Stunden beziffert. Die Zuverlässigkeit des Vorgehens ist um so höher, je unähnlicher sich die Spektren innerhalb einzelner Bereiche des Strukturenraumes sind: Wird der Strukturenraum in Unterräume unterteilt, welche jeweils diejenigen Strukturen enthalten, deren Spektren einander ähnlich sind, so ist die Größe dieser Unterräume ein Maß dafür, wie hoch die Wahrscheinlichkeit von *False Positives* ist, also die Wahrscheinlichkeit, eine andere Strukturhypothese als die tatsächliche Molekülstruktur besser zu bewerten als diese.

Der Schwachpunkt des Verfahrens ist die Generierung eines sehr großen und kaum überschaubaren Suchraums von Molekülstrukturen. Die möglichst frühzeitige Beschränkung desselben ist der Hauptansatzpunkt für Verbesserungen. Meiler und Mehringer schlagen zu diesem Zweck eine Verflechtung des Hypothesengenerierungs- und des Validierungsschrittes vor [Mea02, FS96, MW01].

3.2.2 COCON: Zusätzliche Information durch zweidimensionale NMR-Experimente

COCON [FS96] ist ebenso wie MOLGEN ein Strukturgenerator. Es werden jedoch nicht alle mit der Summenformel übereinstimmenden Isomere generiert, sondern nur diejenigen, die außerdem mit der Konnektivitätsinformation übereinstimmen, welche aus den Befunden spezieller zweidimensionaler NMR-Experimente gewonnen wird. Köck *et al.* kombinieren COCON mit einer Substrukturanalyse (vgl. 3.1.3) und der Hypothesenvalidierung mittels neuronaler Netze [Mea02]. Sie erreichen eine Reduktion der Hypothesenmenge auf etwa 1% der ursprünglichen Größe (basierend auf der Summenformel ohne zusätzliche Konnektivitätsinformation), ohne die Zuverlässigkeit des Strukturaufklärungsverfahrens zu beeinträchtigen.

Befunde aus zweidimensionalen NMR-Experimenten können jedoch nicht grundsätzlich als gegeben vorausgesetzt werden, da solche Untersuchungen einen hohen Aufwand an Kosten, Zeit und Probenmaterial mit sich bringen. In ungünstigen Fällen sind zudem die erhaltenen Daten aufgrund von Mehrdeutigkeiten unterbestimmt, das heißt man erhält nur ein geringes Maß der gewünschten Konnektivitätsinformation. Unabhängig vom gewählten Strukturgenerator ist jedoch die Kombination des klassischen zweischrittigen Vorgehens bei der Strukturaufklärung mit einer Substrukturanalyse, das heißt einer Analyse der inneren Struktur des Hypothesenraumes, vorteilhaft.

3.2.3 SPECSOLV: Subspektrum-Substruktur-Korrelationen

In dem System SPECSOLV [WFR96, WR97] legen die Autoren Richert, Fachinger und Will Wert darauf, daß als Eingabe lediglich das experimentelle ^{13}C-NMR-Spektrum der unbekannten Substanz benötigt wird und es somit unabhängig von der Angabe einer Summenformel ist. Ausgehend von Lage-, Intensitäts- und Multiplettinformation im experimentellen NMR-Spektrum wird eine Subspektrensuche in einer Bibliothek von Subspektrum-Substruktur-Korrelationen (SSC) durchgeführt, und anschließend werden die gefundenen Substrukturen zu einer Gesamtstruktur zusammengesetzt. Als Grundlage der SSC-Bibliothek dient den Entwicklern eine hausinterne SPECINFO-Datenbank.

Die Substruktursuche liefert typischerweise etwa 500 Resultate für Strukturen um 25 schwere Atome. Nicht alle davon sind jedoch notwendigerweise tatsächlich Substrukturen der gesuchten Verbindung. Im nachfolgenden Strukturgenerierungsprozeß finden daher ver-

schiedene Überprüfungen (einschließlich einer SSC-basierten Spektrenvorhersage) statt, bevor Substrukturen zu einem neuen, größeren Strukturfragment verschmolzen werden. Im Prinzip wird also basierend auf dem ^{13}C-NMR-Spektrum eine Positivliste von Strukturbausteinen erstellt und an einen spezialisierten Strukturgenerator übergeben, welcher zudem bereits während des Aufbaus der Gesamtstruktur Zwischenergebnisse validiert, um nur einen möglichst kleinen Teil des Gesamtraumes infragekommender organisch-chemischer Strukturen absuchen zu müssen.

Richert, Fachinger und Will geben an, etwa 80% aller in ihrem Labor untersuchten Reinstoffproben voll automatisch aufklären zu können. Wird jedoch eine Substanz untersucht, die eine unbekannte Substruktur enthält, so muß das System zwangsläufig scheitern. Außerdem ist eine strikte Qualitätsüberwachung der Datenbank- und Bibliothekseinträge durch einen Experten erforderlich, da neben menschlichen Fehlern (z.B. Verwechslungen bei der Zuordnung von Atomen und Absorptionen) auch verschiedene chemische Faktoren wie Lösemittel- und stereochemische Einflüsse die Spektraldaten und damit insbesondere die Zwischenvalidierung durch Subspektrenvorhersage erheblich beeinträchtigen. Auch ungenaue oder unvollständige Strukturbeschreibungen in der Datenbank führen zu Problemen.

3.2.4 GENIUS: ein genetischer Algorithmus zur Hypothesengenerierung

Eine ähnlich starke Verflechtung von Hypothesengenerierung und Validierung realisiert das System GENIUS [MW02, MW01], welches einen genetischen Algorithmus zur Hypothesengenerierung benutzt, dessen Fitnessfunktion die Abweichung des berechneten Spektrums einer Strukturhypothese vom experimentellen Spektrum ist. Das System erhält als Eingabe die Summenformel und das experimentelle ^{13}C-NMR-Spektrum der unbekannten Verbindung. Basierend auf der Summenformel wird eine (nicht sehr große) Startpopulation von Strukturen generiert. Diese werden anhand ihrer mit Hilfe neuronaler Netze (vgl. [Mea02, MMW00, MM02]) vorhergesagten Spektren bzw. deren Abweichung vom experimentellen Spektrum bewertet. Die bestbewerteten gehen in die nächste Generation von Strukturen ein. Es ist parametrisierbar, wie viele Strukturen jede Generation umfaßt, wie viele Individuen direkt übernommen (geklont), wie viele in Variation übernommen (mutiert) und wie viele neue Strukturen durch Rekombination generiert werden. Außerdem ist es möglich, mehrere Populationen sich parallel entwickeln zu lassen.

Der limitierende Faktor für die Verläßlichkeit ist die Anzahl schwerer Atome im Molekül, da sie die Gesamtgröße des Strukturenraums bestimmt. GENIUS ist in der Lage für Moleküle mit bis zu 14 schweren Atomen zuverlässig die korrekte Struktur zu bestimmen, diese Grenze kann jedoch durch Randbedingungen, insbesondere Positiv- oder Negativlisten, weiter nach oben geschoben werden, da solche Bedingungen den Strukturenraum eingrenzen. Negativlisten verbotener Strukturfragmente ermöglichen die Behandlung von Molekülen mit bis zu 20 schweren Atomen, Positivlisten haben einen noch günstigeren Einfluß.

Genetische Algorithmen gehören jedoch zur Gruppe nichtdeterministischer Methoden. Nichtdeterministisch bedeutet unter anderem, daß für das Finden des globalen Optimums hinsichtlich der gegebenen Bewertungsfunktion keine Garantie gegeben werden kann. Darüber hinaus ist die Methode initialisierungsabhängig, und es spielen viele empirisch zu optimierende Parameter für die Leistungsfähigkeit des Systems eine Rolle. Schließlich stellt sich noch die Frage, wann die Iteration immer neuer Generationen von Strukturen zu beenden ist, ob bei Erreichen eines bestimmten Schwellwerts für die Fitnessfunktion, einer feste Zahl von Generationen oder einem absoluten Zeitlimit.

3.3 Grundidee eines neuen Ansatzes

Trotz der Vielfalt heutzutage eingesetzter Methoden bleibt im Bereich der automatischen Strukturaufklärung noch genug Raum für Verbesserungen und neue Ansätze. Insbesondere im Abweichen vom klassischen zweischrittigen Vorgehen hin zu einem starken Ineinandergreifen der Generierung und Validierung von Strukturhypothesen liegt großes Potential. Der Grundgedanke hierbei ist, zum frühestmöglichen Zeitpunkt jede verfügbare Information zu nutzen und so falsche Strukturhypothesen auszuschließen.

Im einfachsten Fall des Vorgehens dient lediglich die Summenformel als Ausgangsinformation für die Generierung von Hypothesen, welche in einem anschließenden Schritt validiert werden. Die Vernetzung beider Schritte mit einander hat es zum Ziel, den bei alleiniger Definition über die Summenformel sehr großen Strukturenraum nicht vollständig generieren und nach der passenden Molekülstruktur durchsuchen zu müssen.

Besonders naheliegend ist es, die in Gestalt des experimentellen ^{13}C-NMR-Spektrums der unbekannten Substanz gegebene Information nicht nur für einen nachgelagerten Abgleich mit den berechneten Spektren zahlloser Strukturhypothesen zu nutzen. Außerdem wäre es wünschenswert, etwaige Zusatzinformationen einbringen zu können, ohne daß die Methode durch die unbedingte Einforderung bestimmter Daten einen hohen Anspruch stellt, welcher in zahlreichen Einzelfällen nicht oder nur unter großem Aufwand erfüllt werden kann.

Um ein alternatives Verfahren zu den bestehenden Methoden zu entwickeln, soll zunächst ein beispielhaftes Anwendungsszenario vorgestellt werden, welches in überschaubarem Rahmen die Möglichkeit bietet, den neuen Ansatz exemplarisch zu realisieren und seine Stärken zu untersuchen. Anschließend werden Bayes-Netze als Methodik vorgestellt, welche eine intelligente Realisierung der oben skizzierten Ideen erlaubt.

3.3.1 Szenario

Für einen neuen Ansatz innerhalb des Feldes der automatischen Strukturaufklärung in der Computerchemie soll zunächst ein begrenztes Szenario betrachtet werden. Es muß überschaubar, darf aber nicht trivial sein, so daß auf der einen Seite die Möglichkeit besteht, einzelne Entscheidungen im Verlauf der Systementwicklung zu beurteilen, ohne auf der anderen Seite ihre Übertragbarkeit auf allgemeinere, möglicherweise um ein Vielfaches komplexere Anwendungen aus dem Blick zu verlieren.

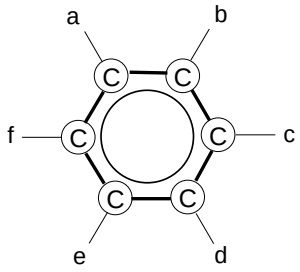


Abb. 3.4: *Schematische Darstellung von Benzolderivaten: Als Substitutionsmuster einer solchen Verbindung wird die Art und relative Anordnung der Substituenten in den Positionen a) – f) bezeichnet.*

Für diesen Zweck wurde die Erkennung von Substitutionsmustern an Benzolringen (vgl. Abbildung 3.4) gewählt. Benzolderivate sind eine interessante Stoffgruppe innerhalb der organischen Chemie, da sie, wie in Abschnitt 2.2 angedeutet, die klassischen exemplarischen

Vertreter der Stoffgruppe der Aromaten sind, so daß diese Aufgabe auch eine gewisse praktische Bedeutung besitzt. Gegenüber der Bestimmung der gesamten Molekülstruktur, selbst auf der Abstraktionsebene der Konstitution, ist sie jedoch weniger komplex: Zum einen ist das Strukturelement eines Benzolringes im Molekül vorgegeben, zum anderen ist lediglich eine Kategorisierung (und nicht die auf Konstitutionsebene genaue Struktur) der einzelnen Substituenten gesucht, welche zu einer Beschreibung von Strukturcharakteristika weiterverarbeitet wird. Die Erweiterbarkeit über Benzolderivate hinaus und hin zu einer vollständigen Strukturbeschreibung sollte bei der Entwicklung jedoch stets mit berücksichtigt werden.

Trotz der durch den Benzolring gegebenen strukturellen Einschränkung ist die Zahl unterschiedlicher Verbindungen noch immer sehr groß. Nicht nur deshalb ist die angestrebte Erkennung von Substitutionsmustern keinesfalls als trivial zu bezeichnen. Zum einen haben Benzol und seine Derivate als klassische Vertreter der Aromaten eine besondere Bedeutung in der organischen Chemie. Zum anderen propagieren aromatische Systeme wie der Benzolring Substituenteneinflüsse auf die Elektronenstruktur des Moleküls besonders gut (vgl. Abschnitt 2.2) – mit der Fokussierung auf aromatische Kohlenstoffatome in dem beschriebenen Szenario ist bei der Entwicklung eines entsprechenden Systems damit eine besonders gewissenhafte Einbeziehung der zugrundeliegenden physikalisch-chemischen Gegebenheiten unerläßlich. Mit Hilfe der am Beispiel aromatischer Verbindungen gemachten Erfahrungen sollte es ohne Probleme möglich sein, entsprechende Systeme zu entwickeln, die auf die drei anderen Typen von Kohlenstoffatomen (vgl. Tabelle 2.1) zugeschnitten sind.

Eine solche Spezialisierung auf die einzelnen Typen von Kohlenstoffatomen ist sinnvoll, da die Charakteristik der Elektronenumgebung maßgeblich von der Hybridisierung der Kohlenstoffatome geprägt wird. Über den Hybridisierungsgrad hinaus werden hinsichtlich sp^2-hybridisierter Kohlenstoffatome wegen der herausragenden Elektroneneigenschaften aromatischer Systeme noch Alkene und Aromaten unterschieden. Ähnlich, mit einer noch tiefergehenden Unterscheidung, verfahren Meiler *et al.* (vgl. [MW01, MM02, MMW00, Mea02, MMWM02]) in ihrem auf neuronalen Netzen basierenden Ansatz: Zur Vorhersage chemischer Verschiebungen werden neun verschiedene neuronale Netze für neun Arten von Kohlenstoffatomen verwendet. Man kann in einer solchen Spezialisierung auch eine konzeptionelle Parallele zu Positiv- und Negativlisten von Strukturelementen oder, bei entsprechendem Spezialisierungsgrad, der Behandlung von Makroatomen sehen.

3.3.2 Das Potential von Bayes-Netzen

In jüngster Zeit erhält das Feld der Graphenmodelle (*graphical models*) mehr und mehr Aufmerksamkeit in verschiedensten Anwendungsbereichen (Überblick: [Smy97], Sprachverarbeitung: [Bil03], Proteinstrukturen: [KB94], Datamining: [BK02]). Zu dieser Methodengattung zählen unter anderem auch Bayes-Netze. Sie besitzen einige vorteilhafte Eigenschaften, welche den Einsatz im Bereich der Strukturaufklärung begünstigen. Zum ersten erlauben sie die Handhabung von Unsicherheiten sowie von Informationen, die vage oder unvollständig sind. Unsicherheiten sind in der Strukturaufklärung in zweierlei Hinsicht von Belang: Zum einen sind die gemessenen Daten, also das experimentelle Spektrum, mit Unsicherheiten durch Meßungenauigkeiten behaftet, zum anderen zeigen die Erfahrungen aus der Spektrenvorhersage, daß bei der Wiedergabe des Zusammenhangs zwischen Spektrum und Struktur Näherungen verwendet werden müssen (vgl. Abschnitt 3.1.2).

Der Aspekt potentieller Unvollständigkeit der gegebenen Daten spielt im Zusammenhang mit Zusatzinformationen eine Rolle: Existierende Systeme, die dem klassischen zweischrittigen Vorgehen von Hypothesengenerierung und anschließender Validierung folgen oder sich

daran orientieren, setzen zumeist die Summenformel der unbekannten Verbindung neben dem ^{13}C-NMR-Spektrum als gegeben voraus. Andererseits benötigt der datenbankgetriebene Ansatz (SPECSOLV [WR97]) diese Information explizit *nicht*, so daß selbst die Summenformel schon als Zusatzinformation gelten könnte. Welche Informationen darüber hinaus vorausgesetzt werden oder optional verarbeitet werden können, variiert von System zu System. Manchmal existiert Vorwissen, oder es sind Befunde aus zusätzlichen Experimenten (etwa 2D-NMR-Spektren wie in COCON [FS96]) gegeben, dies kann jedoch nicht grundsätzlich als gegeben vorausgesetzt werden. Es ist also kein einheitlicher Umfang der Eingangsinformation vorgegeben, welcher als Maßstab herangezogen werden könnte.

Ein Bayes-Netz kann jedoch so gestaltet werden, daß es vielfältige Informationen nutzt, ohne daß dadurch festgelegt wäre, daß die entsprechenden Daten in jedem Einzelfall vorhanden sein müssen. Gleichwohl muß man im Blick behalten, daß nicht jedes nur irgendwie denkbare Eingabedatum mit aufgenommen werden sollte, sondern daß man sich aus Gründen der Übersicht wie auch der Komplexität auf diejenigen Eingaben beschränken sollte, deren Vorhandensein in einem großen Teil der Fälle zu erwarten ist – in diesem Fall bieten sich etwa neben der spektralen Information Angaben über die beteiligten chemischen Elemente an, welche sich entweder auf einer gegebenen Summenformel oder anderem Vorwissen, wie etwa dem Ursprung der Substanz, begründen können.

Der zweite Vorteil von Bayes-Netzen ist die Möglichkeit, Expertenwissen explizit wiederzugeben, was zu einer hohen Nachvollziehbarkeit der Ergebnisse führt, wodurch deren weitere Nutzung durch menschliche Experten erleichtert wird. Einem Bayes-Netz liegt ein Kausalmodell der ursächlichen Zusammenhänge innerhalb der betrachteten Domäne zugrunde. Es repräsentiert das Wissen, welche der im Rahmen der gegebenen Fragestellung relevanten Ereignisse in welchem Maße von einander abhängen. Innerhalb dieses Netzes kausaler Einflüsse ist es möglich, mit oder entgegen der Kausalrichtung Schlußfolgerungen anzustellen, das heißt z.B. von spektralen Merkmalen auf die verursachenden strukturellen Eigenschaften eines Moleküls zu schließen oder umgekehrt. Dies ist lediglich eine Frage der Formulierung der Anfrage an das System, nicht aber des Modells bzw. der Wissensrepräsentation.

Dadurch ist es möglich, in den Strukturaufklärungsprozeß gleichzeitig strukturelle und spektrale Information einfließen zu lassen und deren Auswertung integriert zu nutzen. Ein geeignetes Modell, das die Kausalzusammenhänge sachlich richtig wiedergibt, verbindet Aspekte der Hypothesenbewertung (diagnostischer Rückschluß vom Spektrum auf strukturelle Eigenschaften) mit solchen der Hypothesengenerierung (kausaler Vorwärtsschluß von Informationen aus der Summenformel auf die Molekülstruktur). Die Schlußfolgerungen sind nicht unabhängig von einander, da sie über die Betrachtung der Molekülstruktur und deren Eigenschaften mit einander verbunden sind. Auf diese Weise werden die beiden Schritte Hypothesengenerierung und Strukturvalidierung nicht nur eng mit einander verflochten, sondern quasi in ein und demselbem Schritt integriert.

Gleichwohl stellt die Entwicklung eines entsprechenden Modells, das beiden Sichtweisen gerecht wird, eine erhebliche Herausforderung dar, da es zugleich mit Rücksicht auf die nicht unendlichen Ressourcen an Zeit und Rechnenleistung keine beliebig genaue Wiedergabe der Realität sein kann. Die erforderliche Präzision muß also bereits bei der Wissensrepräsentation berücksichtigt werden. Der Aspekt der Geschwindigkeit wird ebenfalls durch das Kausalmodell beeinflußt; je umfangreicher und detaillierter die Modellierung, desto mehr Rechenoperationen sind zur Beantwortung einer Anfrage nötig.

Zum dritten sind Bayes-Netze prinzipiell unabhängig von experimentellen Daten. Die ursächlichen Zusammenhänge der betrachteten Domäne werden mit bedingten Wahrscheinlichkeiten gewichtet, welche basierend auf wohlfundierten Theorien bzw. Fachwissen ermittelt

werden können. Dadurch ist ein auf Bayes-Netzen basierendes System für beliebige Verbindungsklassen unabhängig von deren Vorkommen in einer Stichprobe anwendbar.

Durch eine entsprechende Anpassung der Parametrisierung, etwa die Verwendung empirisch über relative Häufigkeiten innerhalb einer repräsentativen Stichprobe ermittelter bedingter Wahrscheinlichkeiten, ist es zugleich aber dennoch möglich, eine Spezialisierung auf eine bestimmte Aufgabenstellung, besondere Laborbedingungen oder auch bestimmte Substanzklassen vorzunehmen. Voraussetzung ist lediglich, daß die verwendete Stichprobe in ihrer Charakteristik tatsächlich repräsentativ für das betreffende Spezialgebiet ist. Damit sind Bayes-Netze in der Strukturaufklärung sowohl universell wie auch spezialisiert einsetzbar. Der folgende Abschnitt ordnet vor diesem Hintergrund das Ziel der gegenwärtigen Arbeit in den Zusammenhang der Computerchemie ein.

3.4 Ziel der Arbeit

In der Computerchemie kommen Methoden aus unterschiedlichen Bereichen der Informatik zum Einsatz, um Informationen aus chemischen Daten zu gewinnen. Strukturaufklärungssysteme orientieren sich dabei stark an dem klassischen zweischrittigen Vorgehen, zunächst eine Menge von Hypothesen zu generieren und diese anschließend zu validieren. Grundlage der Hypothesengenerierung ist zumeist die Summenformel, was zu einer sehr umfangreichen Menge zu betrachtender Strukturkandidaten führt. Durch Vergleich ihrer theoretisch berechneten Spektren mit dem experimentellen Spektrum der unbekannten Substanz werden sie bewertet.

Obwohl insbesondere durch den Einsatz neuronaler Netze heutzutage ^{13}C-NMR-Spektren schnell und zuverlässig vorhergesagt werden können, ist eine Reduktion der Hypothesenmenge Hauptansatzpunkt für Verbesserungen. Am vielversprechendsten erscheint es dabei, die beiden Schritte der Generierung und der Validierung der Strukturhypothesen nicht unabhängig von einander zu betrachten, sondern eng mit einander zu verknüpfen.

Für die Untersuchung eines dementsprechenden neuen Ansatzes wird zunächst ein überschaubares, jedoch nicht triviales Szenario gewählt, um seine Leistungsfähigkeit zu testen und Erfahrungen bei seiner Realisierung zu sammeln. Die gewählte Methode muß in ihren Stärken den Anforderungen der Domäne Strukturaufklärung entsprechen.

Bayes-Netze erscheinen vor diesem Hintergrund vielversprechend. Als Versuchsszenario wird die Erkennung von Substitutionsmustern an Benzolderivaten ausgewählt. Die Aufgabe, ausgehend von spektroskopischen Daten aus protonenbreitbandentkoppelten ^{13}C-NMR-Spektren Konstitutionsinformation zu gewinnen, wird dahingehend konkretisiert, die an den Benzolring gebundenen Substituenten nach strukturellen Merkmalen kategorisiert werden sollen und darüber hinaus ihre relative Anordnung erkannt werden soll.

4 Grundlagen

In den vorangegangenen Kapiteln wurde ein Einblick in die Domäne der Strukturaufklärung innerhalb der organischen Chemie sowie ein Überblick über diesbezügliche Ansätze in der Computerchemie gegeben. Es geht nun um die Entwicklung und Realisierung eines neuen Ansatzes in dem in Abschnitt 3.3.1 beschriebenen experimentellen Szenario. Hierfür ist es essentiell, eine fundierte Ausgangsbasis an Grundlagen aus der Informatik zu besitzen. Zugleich ist es wichtig, sich über die Einordnung der gegebenen Aufgabe innerhalb der Disziplin der Informatik und damit verbunden mit den in Frage kommenden Methoden Gedanken zu machen.

Für die Einordnung innerhalb der Informatik als der instrumentellen Mutterdisziplin ist neben den grundlegenden Eigenheiten der Daten des Anwendungsbereichs vor allem das Ziel ihrer Verarbeitung von Bedeutung. Im vorliegenden Fall geht es darum, aus den Eingabedaten Information über die sich in ihnen widerspiegelnden Eigenschaften des durch sie repräsentierten Objekts in der realen Welt zu gewinnen, wobei ein schematischer Zusammenhang zwischen unterschiedlichen Objekten zugrundegelegt wird. Damit läßt sich die Aufgabe dem Feld der Mustererkennung zuordnen.

Die Grundlagen der Mustererkennung sind Gegenstand von Abschnitt 4.1. Im Besonderen ist hierbei die Betrachtung von ^{13}C-NMR-Spektren als Mustern von Interesse. Abschnitt 4.2 gibt eine Einführung in Bayes-Netze und legt im besonderen deren Vorteile und Stärken hinsichtlich der Auswertung von ^{13}C-NMR-Spektren dar. Sie bieten vor allem die Möglichkeit, explizit Expertenwissen aus dem Bereich der Spektrenauswertung in das zu entwickelnde System einzubringen, indem eine geeignete kausale Modellierung zugrundegelegt wird. Abschließend werden daher in Abschnitt 4.3 die Grundzüge der Entwicklung von Kausalnetzen vorgestellt. Abschnitt 4.4 faßt schließlich die essentiellen Punkte noch einmal zusammen und formuliert in diesem Kontext das Ziel der gegenwärtigen Arbeit.

4.1 Mustererkennung

Das Aufgabenfeld der *Mustererkennung* befaßt sich damit, Wahrnehmungsleistungen, welche vom Menschen oder allgemein von Lebenwesen bekannt sind, zu automatisieren, indem das „Muster“ hinter einem bestimmten Ablauf, einer bestimmten Szene, einem bestimmten Objekt usw. erkannt und es auf diese Weise als Instanz eines Oberbegriffs identifiziert oder schematisch beschrieben wird. In diesem Zusammenhang ist es üblich (vgl. z.B. [Nie83]), die Repräsentation des betrachteten Objekts (z.B. Bild- oder Meßdaten) als *Muster* zu bezeichnen. Dies steht im Einklang mit dem Verständnis des allgemeinen Sprachgebrauchs, wo der Begriff Muster [1] eine exemplarische Probe oder ein Schema bezeichnet, welches prototypisch Abläufe oder Objekte charakterisiert.

Diesem Muster soll, abhängig von der gegenwärtigen Domäne, ein bestimmter Sinn abgewonnen werden, welcher sich aus der Relation seiner einzelnen Bestandteile ergibt. Es lassen sich dabei zwei Ansätze unterscheiden: Zum einen kann das Muster als atomar betrachtet

[1] von lat. *monstrare*: „zeigen“

werden; die Zuordnung eines Musters als Gesamtheit zu einer von endlich vielen Kategorien (Klassen) ist das Ziel der *Musterklassifikation*. Demgegenüber verarbeitet die *Musteranalyse* Muster im Kontext der betrachteten Domäne zu einer symbolischen Beschreibung, welche auf der Beziehung der einzelnen Bestandteile des komplexen Musters zueinander basiert. Gleichwohl ist der Schritt der Klassifikation als Brücke zwischen Beobachtungen auf subsymbolischer Ebene und den symbolischen Konzepten der einzelnen Klassen elementare Voraussetzung für derartige weitere Prozesse, die über die bloße Erkennung einer bestimmten Kategorie oder Klasse hinaus in den Bereich des Verstehens hineinreichen.

Die Interpretation von NMR-Spektren als Mustern findet im Kontext der Strukturaufklärung und, allgemeiner, der organischen Chemie statt. Die einzelnen Peaks des Spektrums werden dabei in Bezug zu einander gesetzt, um Hinweise auf das Zusammenspiel möglicher Ursachen der beobachteten Verschiebungen zu erhalten. So wird Information über die im Molekül präsenten Strukturbestandteile gewonnen, die zu einer symbolischen Beschreibung des Molekülaufbaus führt. Das NMR-Spektrum kann somit als komplexes Muster verstanden werden, welches einen Musteranalyseprozeß durchläuft. Ebenso kann es jedoch Gegenstand eines Klassifikationsprozesses sein, wenn beispielsweise die Zuordnung zu einer bestimmten Substanzklasse gesucht ist und nicht eine vollständige symbolische Beschreibung der Molekülstruktur. Im folgenden wird zunächst ein kurzer Überblick über die Felder der Musterklassifikation und der Musteranalyse gegeben. Anschließend sollen Spektren als Muster sowohl als Gegenstand von Klassifikations- wie auch Analyseaufgaben betrachtet werden, um zu einem angemessenen Vorgehen zu gelangen, welches der gegebenen Problemstellung gerecht wird.

4.1.1 Musterklassifikation

Musterklassifikation ist der Schluß von einer Beobachtung auf ein allgemeines, symbolisches Konzept. Dabei ist die Variabilität der Beobachtungen zu berücksichtigen, das heißt es stellt sich die Frage, wie stark diese variieren dürfen, um dennoch als Instanz desselben Konzepts aufgefaßt zu werden. Im folgenden soll ein kurzer Einblick in das Feld der Musterklassifikation gegeben werden, eine detaillierte Einführung sowie Erörterungen betreffend bestimmte Klassifikatoren ist der Literatur (z.B. [Sch96, Nie83, Nie90, Dud73, Fin03]) zu entnehmen.

Im mathematischen Sinne ist die Klassifikation von Mustern eine Abbildung g von einem *Merkmalsvektor* $\vec{c}$ auf eine von endlich vielen disjunkten Klassen ω_i:

$$g : C \mapsto \Omega$$
$$g(\vec{c}) = \omega_i, \quad \omega_i \in \Omega = \{\omega_1, \ldots, \omega_K\}$$

Ein Merkmalsvektor beschreibt dabei das zu klassifizierende Objekt, welches als Muster, das heißt durch seine Meßdaten repräsentiert, vorliegt. Die Abbildung derselben auf einen Merkmalsvektor nennt man *Merkmalsextraktion*; die einzelnen Komponenten des Merkmalsvektors heißen *Merkmale*. Ihre Zahl soll einerseits möglichst gering sein, da der mathematische Aufwand bei der Klassifikation von der Dimensionalität des Merkmalsvektors abhängt, andererseits müssen jedoch ausreichend viele und vor allem geeignete Merkmale gewählt werden, so daß anhand derselben die gewünschte Kategorisierung vorgenommen werden kann.

In einem Klassifikationssystem verläuft die Verarbeitung im allgemeinen folgendermaßen (vgl. Abbildung 4.1): Zunächst werden die Daten aufgenommen, welche das zu klassifizierende Objekt repräsentieren. Diese durchlaufen dann einen Vorverarbeitungsschritt, welcher das Muster im Hinblick auf die durchzuführende Klassifikation verbessert (z.B. Rauschunterdrückung bei der Verarbeitung von Sprachdaten). Im darauffolgenden Merkmalsextraktionsschritt wird es dann in einen Merkmalsvektor überführt. Dieser ist die Eingabe für den

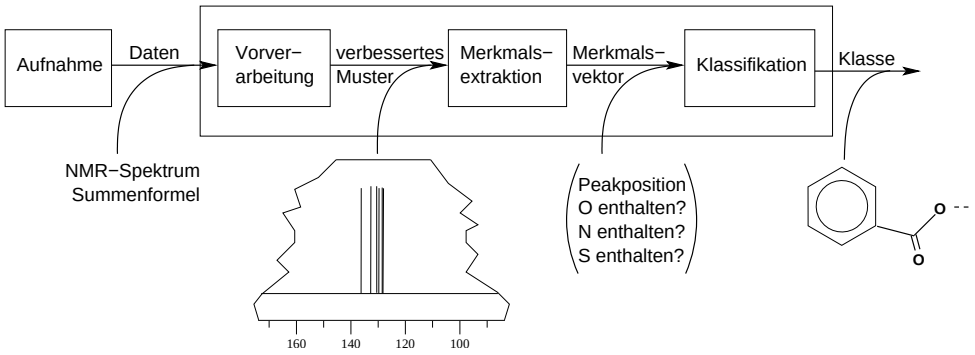


Abb. 4.1: *Klassifikation des Substituenten an einem Benzolring anhand des ^{13}C-NMR-Spektrums und der Summenformel der Verbindung. Nach der Aufbereitung der Daten gehen als Merkmale die chemische Verschiebung des ipso-Kohlenstoffs und die enthaltenen chemischen Elemente in die Klassifikation ein. Die schematische Darstellung eines Klassifikationssystems orientiert sich an [Nie90], Kap. 1.*

eigentlichen Klassifikationsschritt, welcher als Ausgabe die Klasse des betreffenden Objekts liefert. Um Merkmalsvektoren, welche nicht mit ausreichender Sicherheit zugeordnet werden können, zurückzuweisen, kann zusätzlich eine sogenannte *Rückweisungsklasse* ω_0 eingeführt werden. Im folgenden wird dieser Fall jedoch nicht weiter betrachtet.

Betrachtet man die Klassifikation als statistischen Entscheidungsprozeß, so lassen sich der grundsätzliche Ansatz der *Risikominimierung* und seine Spezialfälle, der *Bayes-Klassifikator* und der *Maximum-Likelihood-Klassifikator*, unterscheiden. Alle drei arbeiten mit einer Kostenmatrix, welche Fehlklassifikationen bestraft, die Gestalt der Matrix ist jedoch jeweils unterschiedlich. Beim Risikominimierungsansatz werden die Zuordnungen anhand der *Rückschlußwahrscheinlichkeit* der einzelnen Klassen getroffen, das ist die Wahrscheinlichkeit der jeweiligen Klasse gegeben den Merkmalsvektor. Der Bayes-Klassifikator unterscheidet sich davon nur in der Bewertung von Fehlklassifikationen, die in beiden Fällen ebenfalls in die Entscheidung für eine Klasse eingeht. Maximum-Likelihood-Klassifikatoren dagegen entscheiden anhand der *klassenspezifische Dichte*, wählen also die Klasse, für welche die Wahrscheinlichkeit des gegebenen Merkmalsvektors maximal ist.

Die benötigten Wahrscheinlichkeiten und Wahrscheinlichkeitsdichten erhält man durch Modellierung des mustererzeugenden Prozesses als stochastischem Prozeß. Ein solcher liefert zufällig (aber nicht regellos) Paare aus einem Merkmalsvektor und seiner zugehörigen Klasse. Oft geht man dabei wie in Gleichung (4.1) von einer Normalverteilung aus:

$$p(\vec{c}|\omega_i) = \mathcal{N}_{\vec{c}}(\vec{\mu}_i, \underline{K_i}) = \frac{1}{\sqrt{(2\pi)^N \det \underline{K_i}}} e^{-\frac{1}{2}(\vec{c}-\vec{\mu}_i)^T K_i^{-1}(\vec{c}-\vec{\mu}_i)} \tag{4.1}$$

Dabei ist $\vec{c}$ ein Merkmalsvektor, *Mittelwerte* $\vec{\mu}_i$ und *Kovarianzmatrizen* $\underline{K_i}$ der einzelnen Klassen ω_i sind zu optimierende Parameter. Ziel der Optimierung ist eine möglichst gute Trennbarkeit der zu unterscheidenden Klassen. Ellipsoide, paraboloide oder hyperboloide Klassengrenzen lassen sich gut mit Hilfe von Normalverteilungen approximieren, sind die betrachteten Klassen linear trennbar, so kann der Ansatz durch die Verwendung einer gemeinsamen Kovarianzmatrix vereinfacht werden. In komplizierteren Fällen, für welche ein *Normalverteilungsklassifikator* nicht ausreicht, kann ein *Gemischtverteilungsklassifikator* eingesetzt werden, welcher die tatsächliche Wahrscheinlichkeitsdichte durch eine gewichtete Summe von Normalverteilungen approximiert.

Als Universalapproximatoren sind *Polynomklassifikator*en anzusehen, da nach dem Approximationssatz von WEYERSTRASS (vgl. [BS91], Abschnitt 7.1.2) jede Funktion durch ein Polynom approximiert werden kann, wenn der Grad des Polynoms ausreichend hoch gewählt wird. Auch *neuronale Netze* werden zur Klassifikation eingesetzt. Werden nicht einzelne Muster, sondern Folgen von Mustern betrachtet, so tritt das Problem der Segmentierung der Folge hinzu, etwa in der Sprach- oder Handschrifterkennung – hier sind *Hidden Markov Modelle* [Fin03] das Mittel der Wahl, da sie Segmentierung und Klassifikation in einem Schritt integrieren. Es hängt also stark von der Charakteristik des gegebenen Problems ab, welcher Typ von Klassifikator am besten zu wählen ist. Näheres zum Einsatz von Bayes-Netzen zur Klassifikation ist Abschnitt 4.2 zu entnehmen.

Um die internen Parameter eines Klassifikators (z.B. Mittelwerte, Kovarianzen, Polynomkoeffizienten usw.) zu optimieren, wird eine *repräsentative, klassifizierte Stichprobe* benötigt. Dies ist eine Menge von Beispielmustern, für welche zum einen jeweils die zugehörige Klasse bekannt ist, und welche zum anderen in ihrer Zusammensetzung die Charakteristik des betrachteten Anwendungsfeldes gut wiedergibt. Jedes dieser Beispiele geht als klassifizierter Merkmalsvektor in das *Training* des Klassifikators ein (*überwachtes Lernen*; vgl. auch [Dud73] Kap. 6 zum nicht überwachten Lernen sowie [Fle87] allgemein zu Optimierungsverfahren). Es ist dabei jedoch eine mögliche anschließende Evaluierung des Klassifikators zu beachten, das heißt die Gesamtstichprobe muß in diesem Fall in Trainings- und Testdaten aufgeteilt werden. Werden ein und dieselben Daten für Training und Test benutzt, so hat die Evaluierung auf diesen dem System bereits „bekannten“ Daten keine Aussagekraft.

4.1.2 Musteranalyse

Musteranalyse geht, wie eingangs dargestellt, insofern über die bloße (Wieder-)Erkennung von Mustern im Sinne einer Klassifikation hinaus, als daß sie sich weitgehend auf einer symbolischen Ebene bewegt und somit dem Bereich der künstlichen Intelligenz zugeordnet werden kann. Die meisten für Musteranalysesysteme relevanten Aspekte werden in der Literatur dieses Feldes, wie etwa [Jac99, Win92, Nil98], behandelt. Darüber hinaus geht [Nie90] auf die Verwandtschaft und das Zusammenspiel mit der Musterklassifikation ein und gibt anschaulich am Anwendungsfeld der Szenenerkennung einen Überblick über diese grundlegenden Aspekte und den Aufbau eines Musteranalysesystems.

Anders als in Klassifikationssystemen läßt sich für Musteranalysesysteme kein von der gewählten Domäne unabhängiger schematischer Verarbeitungsverlauf angeben. Auch bei der Musterklassifikation spielt die jeweilige Domäne natürlich eine Rolle, sie bestimmt jedoch hauptsächlich über die innerhalb der einzelnen Verarbeitungsschritte verwendeten Methoden. Da die Musteranalyse sich dagegen, wie bereits erwähnt, weitgehend auf einer symbolischen Ebene der Betrachtung bewegt, hängen die dabei relevanten Prinzipien und Strategien, die den Ablauf bestimmen, stark von der jeweiligen Domäne ab. Darüber hinaus hat auch das Ziel der Verarbeitung einen Einfluß auf deren konkreten Verlauf. Dennoch lassen sich auch für Musteranalysesysteme einige immer wiederkehrende Aspekte identifizieren, welche die Grundbausteine des Systems bilden. Sie sind in einen abstrakten Schema in Abbildung 4.2 dargestellt. Der genaue Verarbeitungsverlauf und die Kopplung der einzelnen Schritte ist jedoch systemspezifisch, und die Einzelmodule werden problemabhängig aktiviert.

Über einen steuernden Kontrollmechanismus sind Komponenten aus den Bereichen Methodik, Wissensrepräsentation, Erklärung sowie Lernen mit einander verbunden. Die Kontrolle muß dabei nicht zwangsläufig in einem eigenen Modul realisiert sein, sondern kann auch im Festlegen eines Verarbeitungsablaufs bestehen. Es ist zu regeln, wann auf welche Sensordaten und Zwischenergebnisse welche Methoden angewendet werden und wann

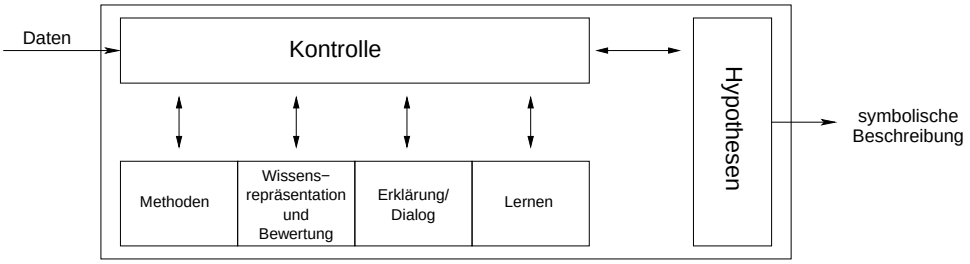


Abb. 4.2: *In einem Musteranalysesystem verbindet ein steuerndes Kontrollelement unterschiedliche Komponenten miteinander und mit der Generierung und Verwaltung von Hypothesen. Der Verlauf der Verarbeitung ist nicht linear und hängt stark vom Kontext der Domäne ab. Die schematische Darstellung orientiert sich an [Nie90].*

welches Wissen zum Einsatz kommt. Darüber hinaus werden bedarfsweise Erklärungs- und Lernmechanismen aktiviert.

Im Bereich der Methoden stehen dabei in der Regel verschiedene allgemeinere Verfahren aus der jeweiligen Domäne zur Verfügung. Sie sind dem Ansatz nach zwar auf das Einsatzfeld (z.B. Bildverarbeitung) abgestimmt, müssen jedoch keinen besonderen Bezug zu dem speziellen Szenario besitzen, in welchem das Musteranalysesystem eingesetzt wird. In diesen Bereich fallen etwa Vorverarbeitung der eingehenden Daten und Merkmalsberechnung. Während dieser Schritt auf die subsymbolische Ebene beschränkt bleibt, ist der Wissensbereich eng mit dem konkreten Szenario verbunden. Die qualitative Repräsentation des Wissens und seine quantitative Bewertung schaffen ein Modell der betrachteten Domäne und bilden so die Grundlage des Musteranalyseprozesses.

Die Lernkomponente soll in diesem Zusammenhang die Möglichkeit bieten, das vorhandene Domänenwissen zu erweitern oder sogar automatisch zu gewinnen, da es in vielen Fällen sehr aufwendig sein kann, das benötigte Wissen von Hand akkurat und fehlerfrei wiederzugeben. Sofern dies möglich ist und geeignete Methoden zur Verfügung stehen, können Repräsentation oder Parametrisierung des Wissens (etwa die Gewichtung von Abhängigkeiten oder Kausalzusammenhängen) automatisch gewonnen werden.

Bedeutsam für die Interaktion mit dem Benutzer, aber auch für Wartung und Weiterentwicklung des Systems, ist die Erklärungskomponente. Sie dient dazu, die Systemantwort transparent zu machen: Ein Experte, der mit dem System arbeitet, kann diese so besser einordnen, ihre Verläßlichkeit einschätzen und z.B. darauf basierend entscheiden, die Anfrage an das System anders zu stellen, andere oder zusätzliche Daten verfügbar zu machen oder auch die Notwendigkeit feststellen, das repräsentierte Wissen zu korrigieren oder zu ergänzen.

Die Bewertung, Verwaltung und Weiterverarbeitung von Hypothesen ist schließlich derjenige Bestandteil des Musteranalysesystems, welcher zur letztendlichen Gewinnung der gewünschten Information führt. Die unterschiedlichen Systemkomponenten liefern verschiedene Zwischenergebnisse auf dem Weg zu der gewünschten symbolischen Beschreibung. Die Ansprüche an diese sind je nach Fortschritt der Verarbeitung unterschiedlich; genügen die vorhandenen Hypothesen in ihrer Qualität, Verläßlichkeit, oder dem Detailgrad ihrer Beschreibung den Ansprüchen nicht, müssen durch Rückmeldung an die Kontrolle entsprechende Maßnahmen eingeleitet werden (z.B. Erklärung an den Benutzer, weshalb keine adäquate Systemantwort erreicht wurde oder Durchführung zusätzlicher Aufbereitungsschritte).

Bei der Akzentuierung eines konkreten Systems kommt jedoch nicht allen der aufgezeigten Bestandteile stets die gleiche Bedeutung zu. In der vorliegenden Arbeit liegt der Schwerpunkt auf der expliziten Repräsentation von Expertenwissen. Damit hängt die Entscheidung für

Bayes-Netze zusammen, welche Gegenstand von Abschnitt 4.2 sind. In den Bereich der Methodik fällt es, zuvor chemische Daten, im einzelnen Strukturdaten und digitale ^{13}C-NMR-Spektren, einzulesen und daraus die nötige Ausgangsinformation zu gewinnen. Nach dem Klassifikationsschritt werden die Ergebnisse an ein Modul zur Hypothesengenerierung weitergeleitet, welches daraus das gesuchte Substitutionsmuster generiert.

Der Aspekt des Lernens soll zugunsten einer fundierten und nachhaltigen Bearbeitung der übrigen Komponenten zurückgestellt werden, wenngleich Bayes-Netze die grundsätzliche Möglichkeit bieten, auch dies zu realisieren. Zumindest kann die Anpassung der bedingten Wahrscheinlichkeiten, welche in Bayes-Netzen zur Quantifizierung der bekannten Kausalzusammenhänge (das heißt des Wissens) dienen, als Lernen der Charakteristika der Einsatzumgebung aufgefaßt werden.

Auch die Entwicklung einer Erklärungskomponente wird zurückgestellt: In diesem Bereich eröffnet sich ein weites Feld möglicher Arbeiten zu Dialog und Interaktion zwischen Mensch und Computer, welches zusätzlich zum bereits genannten Schwerpunkt der Wissensrepräsentation zu behandeln den Rahmen einer einzelnen Arbeit sprengen würde. Lediglich in rudimentärer Weise wird es durch die Forderung einer adäquaten Ausgabe des Analyseergebnisses berührt. Langfristig wäre es jedoch wünschenswert, wenn das System sich darüber hinaus quasi in einen sachlichen Dialog mit dem Benutzer begeben könnte, um auf diese Weise mit menschlichen Experten als „virtueller Laborassistent“ zusammenzuarbeiten.

Die Kontrolle der Verarbeitung ist immer ist dann von Bedeutung, wenn Fehler oder Unzulänglichkeiten auftreten. In diesem Fall müssen geeignete Maßnahmen eingeleitet werden, um den Fehlerfall zu bereinigen. Weiterhin kann es jedoch günstig sein, unter bestimmten Umständen bereits vorab Maßnahmen zu ergreifen, um einen Konflikt in der Verarbeitung von vornherein zu vermeiden, anstatt ihn im Nachhinein auszuräumen. Offensichtlich sind hierbei umfangreiche Überlegungen und sorgfältiges Abwägen erforderlich, so daß eine intensive Bearbeitung neben dem Schwerpunkt der Wissensrepräsentation nicht realistisch erscheint. Daher soll zunächst von einem Optimalfall ausgegangen werden, das heißt das System geht intern von einer fehlerfreien Verarbeitung aus. Beim Eintreten eines Konflikts soll dieser zwar an den Benutzer gemeldet werden, die Erarbeitung einer Strategie, die solche Fehlerfälle bereinigt, liegt jedoch jenseits des Rahmens dieser Arbeit.

Für einige Anwendungsgebiete der Musteranalyse (z.B. Sprachverarbeitung und Bildverstehen) hat sich ein mehr oder minder festes Verarbeitungsschema etabliert. In ähnlich schematischer Weise kann auch das klassische zweischrittige Prinzip der Strukturaufklärung beschrieben werden, wie Abbildung 4.3 zeigt: Ausgehend von der Summenformel werden zahlreiche Strukturhypothesen generiert, deren Spektren berechnet und diese schließlich mit dem Spektrum der unbekannten Substanz verglichen. Dazwischen können Auswahlschritte wie die Anwendung von Positiv- oder Negativlisten oder die Berechnung der *Fitness* stattfinden.

Bestehende Systeme variieren oder ergänzen bereits dieses Vorgehen. Im Falle des auf genetischen Algorithmen basierenden Systems GENIUS wäre etwa eine Rückkopplung von der Hypothesenliste über den genetischen Selektionsmechanismus zurück zum Strukturgenerator einzufügen. Und der auf SSC-Bibliotheken basierende Ansatz von SPECSOLV benötigt keine Summenformel als Eingabe, sondern dem Strukturgenerator wird eine Subspektrensuche mit dem experimentellen Spektrum als Eingabe vorgelagert; außerdem finden die nachfolgenden Schritte auf der Ebene von Subspektren statt, und ihre Ergebnisse gehen wiederum in die Strukturgenerierung ein.

Damit ist jedoch die Vielfalt vorstellbarer Vorgehensweisen längst nicht erschöpft. Insbesondere wird, wie in Kapitel 3 dargestellt, eine stärkere Verbindung oder Integration der Schritte der Hypothesengenerierung und Validierung angestrebt. Im folgenden sollen daher

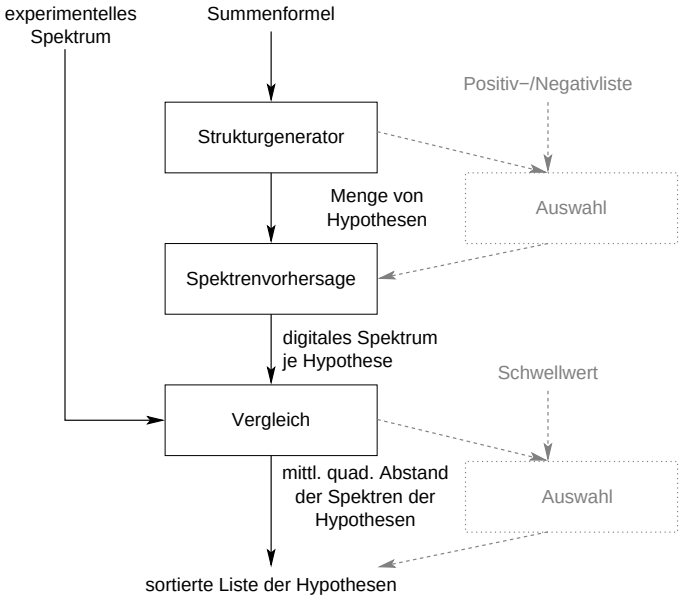


Abb. 4.3: *Aufbau eines Strukturaufklärungssystems nach dem klassischen Schema von Hypothesengenerierung und anschließender Validierung. Der Aufbau ist weitgehend linear.*

Spektren als Muster in Klassifikations- oder Analyseprozessen betrachtet werden, so daß schließlich ein alternatives Verarbeitungsschema entwickelt werden kann.

4.1.3 Spektren als Muster

Die Meßdaten eines digitalen Spektrums können im Kontext der Strukturaufklärung als Muster aufgefaßt werden. Es ist nun die Frage, welche Information bezüglich des Moleküls gewünscht ist: Die Zuordnung zu einer bestimmten Klasse von Molekülen (Klassifikation), oder eine symbolische Beschreibung der Struktur (Musteranalyse). Unterschiedliche Aufgabenstellungen sind denkbar; manche sind sogar auf den ersten Blick nicht ausschließlich dem einen oder dem anderen Bereich zuzuordnen:

- **Zuordnung der Testsubstanz zu einer Substanzklasse**: Dies ist eindeutig eine Klassifikationsaufgabe. Gegeben eine endliche Menge von Substanzklassen werden für jedes Spektrum Merkmale berechnet, anhand derer sich die Zugehörigkeit zu einer bestimmten Substanzklasse entscheiden läßt.

- **Zuordnung von Peaks zu Chromophoren**: Diese Aufgabe ähnelt der vorgenannten, sofern eine endliche Menge von Chromophoren gegeben ist. Je detaillierter aber die Unterscheidung der Typen von Kohlenstoffatomen ist (wie in Abschnitt 2.3.2 dargestellt haben auch mehrere Bindungen entfernte Atome noch einen Einfluß auf die Elektronenumgebung eines Kohlenstoffkerns), desto weniger ist aufgrund der Vielzahl zu unterscheidender Klassen eine Realisierung als Klassifikationsaufgabe praktikabel.

- **Zuordnung von Gruppen von Peaks zu strukturellen Untereinheiten**: Auch hier gilt ähnliches wie im Fall der Zuordnung von Peaks zu Chromophoren. Zum einen kann eine Subspektrenklassifikation durchgeführt werden, zum anderen kann eine symbolische Beschreibung des zugrundeliegenden Strukturfragments gewünscht sein.

- **Identifikation einer bestimmten Art von Peakgruppe**: Neben dem Klassifikationsaspekt, das Vorliegen der betreffenden Peakgruppe festzustellen, kann es bei dieser Aufgabe sinnvoll sein ebenfalls anzugeben, wo im Spektrum die betreffende Formation vorliegt, unter welchen Voraussetzungen die vorgefundene, möglicherweise verzerrte Ausprägung zustandekommen kann und wie plausibel unter diesen Umständen die Identifikation ist. Erklärungen, Einbringen von Wissen und Hypothesenbewertungen dieser Art entsprechen der Bearbeitung eines Musteranalyseproblems.
- **Strukturbeschreibung der Testsubstanz aufgrund spektraler Befunde**: In diesem Fall kann man mit Sicherheit von einer Musteranalyseaufgabe sprechen. Gegeben das Spektrum ist eine symbolische Beschreibung der zugrundeliegenden Molekülstruktur gesucht. Innerhalb der Analyse können gleichwohl Klassifikationsaspekte wie die Zuordnung von Peaks zu bestimmten Klassen von Kohlenstoffatomen eine Rolle spielen.

Das in Abschnitt 3.3.1 beschriebene Szenario, die Erkennung von Substitutionsmustern an Benzolderivaten, ist innerhalb des letzten Punktes einzuordnen. Es sind ausdrücklich keine bestimmten Klassen von Substitutionsmustern vorgegeben, sondern mit Blick auf eine zukünftige Erweiterbarkeit des Ansatzes auf allgemeinere Strukturaufklärungsaufgaben ist eine Beschreibung des Substitutionsmusters gesucht. Mit Bezug auf die Bestandteile eines Musteranalysesystems sind als eingehende Daten Spektrum und Summenformel zu nennen, die Ausgabe ist eine schematische Beschreibung des Substitutionsmusters. In einer solchen Beschreibung sind üblicherweise Substitutionsgrad sowie Positionen und Arten der Substituenten zu nennen.

Hinsichtlich der Arten von Substituenten spielt einmal mehr die Sichtweise des chemischen Anwendungsfeldes eine Rolle. Hier soll daher die Unterscheidung rein nach strukturellen Merkmalen vorgenommen werden. Der systematischen Beschreibung chemischer Strukturen dient der HOSE-Code (vgl. Abschnitt 3.1.5). Die ihm zugrundeliegenden Ideen können zur systematischen Unterscheidung der Substituenten verwendet werden, welche an einen Benzolring gebunden sind. Es stellt sich jedoch die Frage der Granularität dieser Unterscheidung. Vorrangige Kriterien sollen hier, wie in Abbildung 4.4 verdeutlicht, die höchste innerhalb des Substituenten noch besetzte Sphäre (also die Länge des Substituenten) sowie chemisches Element und Hybridisierung der Atome der ersten und zweiten Sphäre sein.

Der dargestellte Baum unterscheidet zunächst nach ein- und mehratomigen Substituenten und dann nach chemischem Element und Hybridisierung des Atoms der ersten Sphäre. Für die mehratomigen Substituenten wird für jede dieser Ausprägungen ein „Standardsubstituent“ definiert, in welchem alle freien Bindungen entweder mit Wasserstoffatomen oder im Fall von Mehrfachbindungen mit demselben Atomtypus wie in der ersten Sphäre besetzt sind. Im letzteren Fall darf die dritte Sphäre ausschließlich Wasserstoffatome enthalten. Eine Ausnahme bilden die NO_2- und die SO_3H-Gruppe, welche alle freien Valenzen (Bindungen) der ersten Sphäre mit Sauerstoffatomen in der zweiten Sphäre besetzt haben, um die Charakteristik ihres typischen Vorkommens zu erhalten. Anschließend wird weiter unterschieden, ob es sich um den „Standardsubstituenten“ oder ein Derivat desselben handelt. Nach demselben Schema ist auch eine genauere Beschreibung unter Berücksichtigung von mehr als zwei Sphären möglich, sofern gewünscht.

Damit ist ein einzelner Substituent eindeutig einer strukturellen Klasse zuzuordnen, für die Betrachtung der Verbindung insgesamt sind jedoch alle sechs Positionen des Benzolringes zu betrachten. Nach ihrer Klassifikation kann mit Hilfe der Position der betreffenden Substituenten relativ zu einander das Substitutionsmuster der vorliegenden Verbindung beschrieben werden.

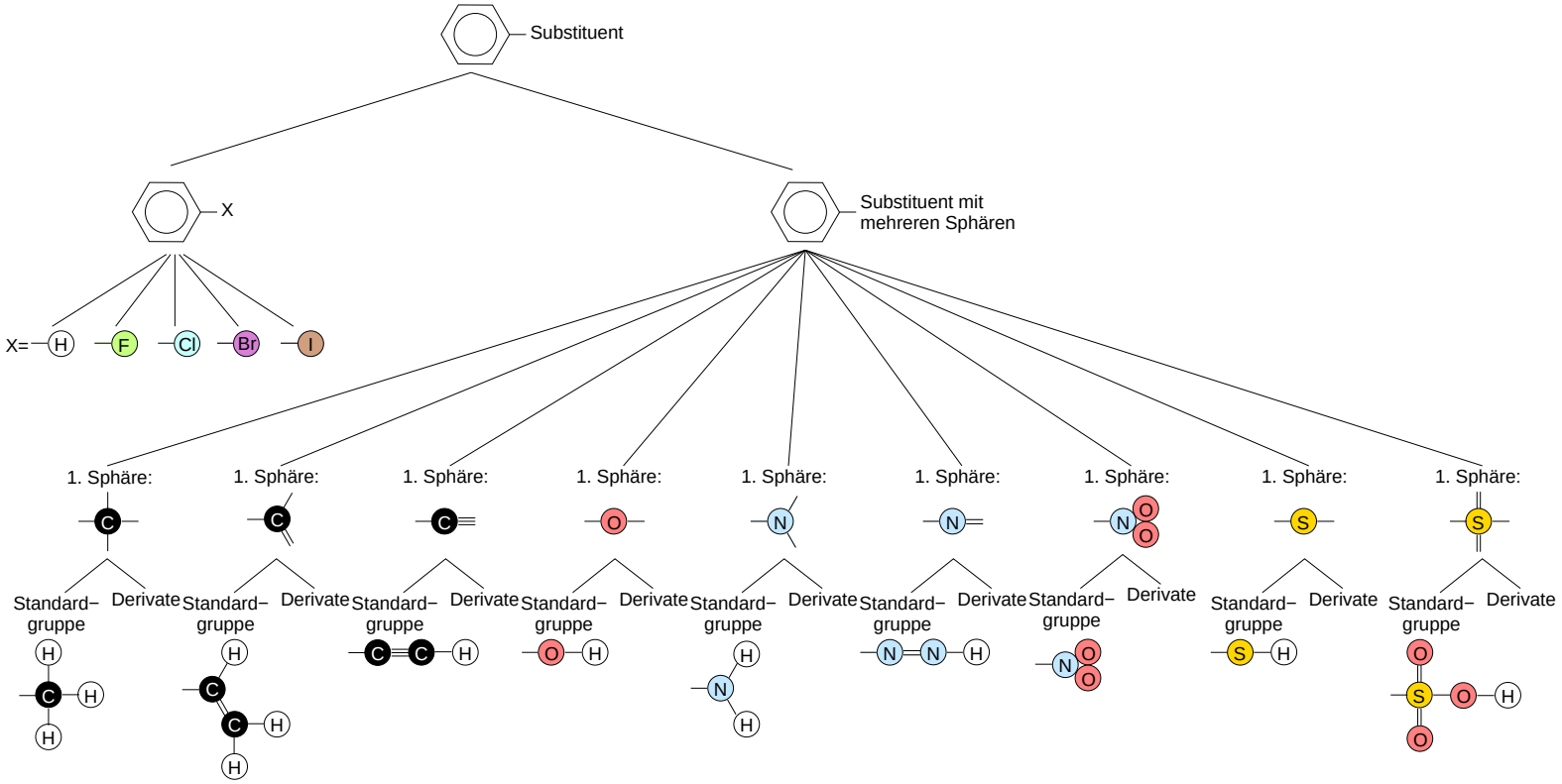


Abb. 4.4: *Klassifikationsbaum zur Unterscheidung von Substituenten an Benzolringen nach strukturellen Gesichtspunkten. Um ein Benzolderivat einem bestimmten Substitutionsmuster zuzuordnen, müssen jedoch alle sechs Positionen des Ringes und ihre relative Stellung zueinander betrachtet werden.*

Der wechselseitige Einfluß der einzelnen Ringatome und ihrer Substituenten auf ihre chemische Verschiebung ist dabei jedoch zu berücksichtigen, sowohl bei der Klassifikation der Einzelpositionen, als auch hinsichtlich der relativen Anordnung der Substituenten. Dieser Aspekt kann zum einen für die Hypothesengenerierung ausgenutzt werden; es ist ferner denkbar, daß bei einem Fehlschlag oder einer schlechten Hypothesenbewertung auf dieser Basis eine Revision des Klassifikationsprozesses veranlaßt werden kann.

Die gegenwärtige Problemstellung beinhaltet also einen Klassifikationsaspekt: Die Substituenten des Benzolringes sollen jeweils einer strukturellen Klasse zugeordnet werden. Darüber hinaus reicht sie jedoch in den Bereich der Musteranalyse hinein, da nicht die einzelnen Substituentenklassen, sondern das Substitutionsmuster gesucht ist. Der folgende Abschnitt widmet sich nun der gewählten Methodik von Bayes-Netzen, die, wie in Abschnitt 3.3.2 beschrieben, im Zusammenhang der Problemstellung besonders geeignet erscheinen.

4.2 Bayes-Netze

Ein *Bayes-Netz* [Pea88, Spi84, Pea86] ist ein Graphenmodell, dessen Schwerpunkt auf der Betrachtung kausaler Beziehungen zwischen den innerhalb einer gegebenen Domäne interessanten *Ereignissen* liegt. Dies wird durch die Verbindung eines *Kausalnetz*es mit Prinzipien der Wahrscheinlichkeitstheorie erreicht. In einem solchen Kausalnetz sind die Knoten Ereignisse und die gerichteten Kanten zwischen ihnen ursächliche Verknüpfungen. Während auf die Entwicklung von Kausalmodellen in Abschnitt 4.3 eingegangen wird, soll im folgenden der methodische Ansatz von Bayes-Netzen motiviert werden. Es schließen sich grundlegende mathematische Definitionen und Begriffe an. Ein kurzer Überblick über den Einsatz von Bayes-Netzen in der Mustererkennung sowie eine Darstellung der Rolle des Bayes-Netzes innerhalb des geplanten Mustererkennungssystems schließen den Abschnitt ab.

4.2.1 Motivation

Bei der Entwicklung der ersten *Expertensysteme* oder *wissensbasierten Systeme* in den 1960er Jahren hatte man das hoch gesteckte Ziel, menschliche Experten durch präzisere Maschinen ersetzen zu können [Jac99, Jen96]. Diese Systeme besaßen eine *Wissensbasis* von *Produktionsregeln* und ein *Inferenzsystem* zum Anstellen von Schlußfolgerungen. Sie erzielten anfänglich große Erfolge, *unsicheres Schließen* [Pea88, Som92] jedoch führte sie bald an ihre Grenzen.

Unsicherheiten können zum einen in einer Unvollständigkeit des Wissens begründet sein, wenn etwa die zugrundeliegenden Naturgesetze nicht bekannt oder zu komplex sind, um sie erschöpfend wiederzugeben. Die Erstellung einer Wissensbasis von Hand oder durch maschinelles Lernen sowie das korrekte Folgern aus Beobachtungen wird dadurch erschwert oder unmöglich gemacht. Zum anderen kann das System auch *unsichere Information* zur Verarbeitung erhalten, die es unter Umständen unmöglich macht, zu einem eindeutigen Schluß zu kommen. Diese Unsicherheit kann darin bestehen, daß ein Teil der erwarteten Angaben nicht verfügbar ist, oder kann dadurch gegeben sein, daß Beobachtungen zwar vorhanden, aber vage oder uneindeutig sind.

Die Fähigkeiten des Menschen, einzelne Fakten unterschiedlich zu gewichten, um so das Fehlen weniger wichtiger Beobachtungen kompensieren zu können, sowie die Verwendung vager Begriffe (*„gutes Wetter“*) versuchte man wissensbasierten Systemen durch die Erweiterung um Wahrscheinlichkeiten zu verleihen. Da Inferenzsysteme jedoch kontextfrei sind,

während die Wahrscheinlichkeiten aber durchaus vom Kontext ihrer Betrachtung abhängen, konnte auf diese Weise keine befriedigende Lösung erreicht werden.

Anstelle der Verwendung um Wahrscheinlichkeiten erweiterter, aber kontextfreier Inferenzsysteme erwuchs die Idee, gänzlich auf die Wahrscheinlichkeitstheorie zurückzugreifen, welche in der *Entscheidungstheorie* zu einem präzisen mathematischen Rahmen des rationalen Treffens von Entscheidungen aufgearbeitet worden war. Außerdem sei die Domäne zu modellieren, nicht der Experte, um diesen bei seiner Arbeit innerhalb derselben zu unterstützen. Wenngleich diese Gedanken bereits seit den 1960er Jahren existierten, dauerte es bis in die Neunziger, bis entsprechend leistungsfähige Maschinen zu seiner Realisierung zur Verfügung standen.

Die herausragende Eigenschaft von Bayes-Netzen ist in diesem Zusammenhang, daß sie die Verwendung von vager, unsicherer oder unvollständiger Information ermöglichen. Fehlen einige der Befunde, deren Einbringung das Bayes-Netz prinzipiell erlaubt, so können die vorhandenen Informationen dennoch ausgewertet werden. Dies kann in der gegenwärtigen Aufgabe ausgenutzt werden, indem bei der Klassifikation eines Substituenten die Nachbarpositionen im Benzolring berücksichtigt werden. Ist entsprechende Information nicht vorhanden, so können Peakposition und Summenformelinformation dennoch ausgewertet werden; wurden jedoch bereits einige Positionen klassifiziert, so können deren bereits bekannte Substituenten wertvolle Zusatzinformation liefern.

Im Kontext der Strukturaufklärung ist es außerdem sehr vorteilhaft, daß sowohl *diagnostische Evidenzen* als auch *kausale Evidenzen* zum Tragen kommen. Diagnostische Evidenzen sind solche Beobachtungen, bei deren Auswertung entgegen der Kausalrichtung (von der beobachtbaren Wirkung auf die mögliche Ursache) geschlossen wird, kausale Evidenzen werden dagegen mit der Ursache-Wirkungs-Richtung propagiert. Spektroskopische Befunde sind diagnostische Evidenzen: Betrachtet man eine beobachtete chemische Verschiebung, so hängt diese von der Gestalt des zugehörigen Chromophors ab. Demgegenüber ist die Information aus der Summenformel eine kausale Evidenz, da sie die Gestalt, die der Chromophor haben kann, ursächlich beeinflußt. Ist etwa bekannt, daß kein Stickstoff im Molekül vorkommt, so ist eine dem Kohlenstoffatom benachbarte Aminogruppe ausgeschlossen, oder enthält das Molekül genau ein Sauerstoffatom, so kann zwar eine benachbarte Alkohol-, aber keine benachbarte Carboxylgruppe (Carbonsäuregruppe) vorliegen. Auch Informationen betreffend die übrigen Ringpositionen sind kausale Evidenzen.

Auf diese Weise kann verglichen mit dem klassischen zweischrittigen Vorgehen bei der Strukturaufklärung eine starke Verflechtung des Hypothesengenerierungs- und des Validierungsschrittes erreicht werden. Der Strukturenraum wird dadurch beschränkt, daß diagnostische und kausale Evidenzen in einem Schritt zusammenwirken. Voraussetzung hierfür ist jedoch ein geeignetes Modell, welches die kausalen Gegebenheiten der Domäne korrekt und mit ausreichender Genauigkeit widerspiegelt. Daher soll in Abschnitt 4.3 im besonderen auf Aspekte der Modellentwicklung eingegangen werden, nachdem die grundlegenden mathematischen Prinzipien dargelegt sowie der Einsatz von Bayes-Netzen in der Mustererkennung allgemein wie auch im Rahmen der gegebenen Aufgabenstellung erörtert wurde.

4.2.2 Begriffe und mathematische Grundlagen

Im mathematischen Sinne ist ein Bayes-Netz ein gerichteter, azyklischer Graph, dessen Knoten Zufallsvariablen sind (vgl. [Jen96], Abschnitt 2.3.6). Jede Variable V des Netzes repräsentiert ein bestimmtes Ereignis in der betrachteten Domäne und besitzt eine endliche Menge diskreter Zustände $v_0 \dots v_n$, welche unterschiedliche *Ausprägungen* des modellierten Ereig-

nisses wiedergeben. (Kontinuierliche Variablen sollen an dieser Stelle nicht Gegenstand der Betrachtung sein.) Die gerichteten Kanten des Graphen sind als Kausalverbindungen zu interpretieren, die von einer Ursache A zum durch sie beeinflußten Ereignis B verlaufen.

Jede Variable B ist mit einer Tabelle $P(B|A_1 \ldots A_n)$ bedingter Wahrscheinlichkeiten versehen, welche die Stärke des Einflusses der einzelnen möglichen Ursachen $A_1 \ldots A_n$ auf das durch B modellierte Ereignis quantifiziert. Für Variablen, die innerhalb des Graphen keine Eltern besitzen, sind (nicht-bedingte) *a-priori-Wahrscheinlichkeiten* $P(A)$ gegeben. Abbildung 4.5 verdeutlicht diese Zusammenhänge. Die benötigten Wahrscheinlichkeitsverteilungen können wohlfundierten Theorien entstammen, sie können jedoch ebenso empirisch ermittelt oder (z.B. basierend auf der Einschätzung eines Experten) frei festgelegt werden.

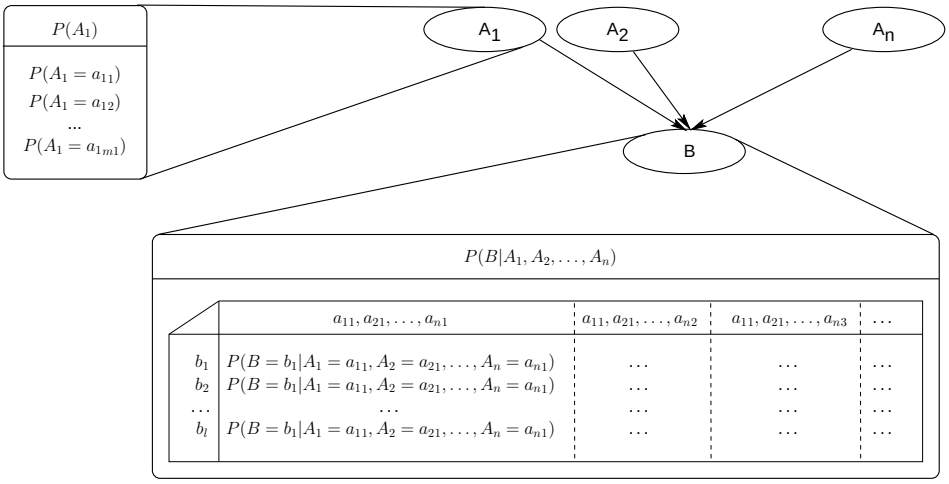


Abb. 4.5: *Ein sehr rudimentäres Bayes-Netz. Die Knoten des Netzes sind Zufallsvariablen. A_1 bis A_n sind Ursachen von B; der Einfluß, den sie auf B haben, wird durch die mit B assoziierte Wahrscheinlichkeit $P(B|A_1 \ldots A_n)$ quantifiziert. Mit A_1 bis A_n sind jeweils* a-priori-*Wahrscheinlichkeiten $P(A_1)$ bis $P(A_n)$ assoziiert.*

Während die Auswahl und Repräsentation von Ereignissen und ihren Kausalzusammenhängen Gegenstand von Abschnitt 4.3 ist, sollen an dieser Stelle die mathematischen Gesetzmäßigkeiten der Wahrscheinlichkeitstheorie betrachtet werden (siehe auch [Bos95]), welche der Informationsgewinnung mit Hilfe von Bayes-Netzen zugrundeliegen. Die Wahrscheinlichkeit eines beliebigen Ereignisses E wird mit $P(E)$ bezeichnet. E ist eine Kombination von *Elementarereignis*sen ω_i, welche selbst nicht als Kombination anderer Ereignisse betrachtet werden können, und ist in diesem Sinne als Menge zu verstehen. Die Menge aller möglichen Elementarereignisse heißt *Wahrscheinlichkeitsraum* und wird mit Ω bezeichnet. Für jedes *Wahrscheinlichkeitsmaß P* gelten die folgenden Grundaxiome:

$$P(\Omega) = \sum_i P(\omega_i) = 1 \tag{4.2}$$

$$0 \leq P(A) \leq 1 \qquad \forall A : A \subseteq \Omega \tag{4.3}$$

$$P(A \vee B) = P(A) + P(B) \qquad \forall A, B : A \cap B = \oslash \tag{4.4}$$

Für die quantitative Betrachtung kausaler Zusammenhänge werden *bedingte Wahrscheinlichkeiten* benötigt: $P(A|B)$ ist die „Wahrscheinlichkeit von A gegeben B“, die Wahrscheinlichkeit, daß A eintritt, wenn B bereits eingetreten ist. Sie stellt einen Übergang in einen anderen Wahrscheinlichkeitsraum dar (vgl. [Bos95], S. 31/32). Es sei außerdem $P(A,B)$ die Wahr-

scheinlichkeit des Eintretens beider Ereignisse A und B, also $P(A,B) = P(E)$ mit $E = A \cap B$. Da offensichtlich $P(A,B) = P(B,A)$ können die folgenden Gleichungen (4.5) und (4.6), welche für bedingte Wahrscheinlichkeiten stets gelten, gleichgesetzt werden, was zum *Satz von* BAYES (4.7) führt:

$$P(A|B)P(B) = P(A,B) \tag{4.5}$$

$$P(B|A)P(A) = P(B,A) \tag{4.6}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{4.7}$$

Sind A und B *disjunkte Ereignisse*, das heißt $A \cap B = \emptyset$, so ist $P(A,B) = P(A)P(B)$. Der *Satz von der Totalen Wahrscheinlichkeit* (4.8) verallgemeinert dies für eine Folge $(b_n)_{n\in\mathbb{N}}$ disjunkter Ereignisse; mit seiner Hilfe läßt sich der Satz von BAYES (4.7) wie in Gleichung (4.9) verallgemeinern. Diese Verallgemeinerungen sind nötig, um Ereignisse betrachten zu können, die im Rahmen einer kausalen Modellierung auftreten und mehrere mögliche Ausprägungen besitzen.

$$P(B) = \sum_{i=1}^{n} P(a_i)P(B|a_i) \tag{4.8}$$

$$P(b_j|A) = \frac{P(A|b_j)P(b_j)}{\sum_i P(b_i)P(A|b_i)} \tag{4.9}$$

In diesem Zusammenhang kann zur Verdeutlichung für das Folgende wiederum Abbildung 4.5 dienen. Sei A eine Zufallsvariable in einem Kausalmodell, dann ist $P(A)$ ein Vektor reeller Zahlen zwischen 0 und 1. Seine i-te Komponente ist die Wahrscheinlichkeit $P(A = a_i)$, daß sich A im Zustand a_i befindet. Kurz schreibt man auch $P(a_i)$. A habe n Zustände und sei innerhalb des Kausalmodells Ursache der m-wertigen Zufallsvariable B. Die bedingte Wahrscheinlichkeit $P(B|A)$, daß bei Eintreten von A als Folge auch B eintritt, ist dann eine $m \times n$ Einträge große Tabelle mit den Einträgen $P(b_k|a_i)$ mit $0 \leq a_i \leq n$ und $0 \leq b_k \leq m$. Hat B mehrere Ursachen, so wird deren Zusammenwirken durch $P(B|A_1 \ldots A_l)$ erfaßt; die Einzelwahrscheinlichkeiten $P(B|A_1) \ldots P(B|A_l)$ sind dagegen nicht von Bedeutung.

Da die Zustände a_i einer Variablen A disjunkt und erschöpfend sein müssen, gelten in Analogie zu Gleichung (4.2) auch die Gleichungen (4.10) und (4.11), das heißt die Summe der einzelnen Komponenten jeder *a-priori*-Wahrscheinlichkeit $P(A)$ und die Spaltensumme jeder bedingten Wahrscheinlichkeit $P(B|A_1 \ldots A_m)$ beträgt jeweils 1.

$$\sum_{k=1}^{n} P(a_k) = 1 \tag{4.10}$$

$$\sum_{k=1}^{n} P(b_k|a_1^{(i)} \ldots a_m^{(i)}) = 1 \tag{4.11}$$

mit i : Index der betrachteten Spalte

Interessant im Kontext kausaler Zusammenhänge ist auch die Frage der Abhängigkeit oder Unabhängigkeit von Ereignissen. Zwei Ereignisse sind unabhängig von einander, wenn das Eintreten des einen die Wahrscheinlichkeit des anderen nicht beeinflußt (Gleichung (4.12)). *Bedingt unabhängig* sind zwei Ereignisse A und C, wenn gegeben ein drittes Ereignis B keinerlei Kenntnisse über C die Wahrscheinlichkeit von A beeinflussen und umgekehrt. Die

in Gleichung (4.13) dargestellte Äquivalenz, welche dies wiedergibt, läßt sich mit Hilfe des Satzes von BAYES zeigen.

$$P(A|C) = P(A) \tag{4.12}$$

$$P(A|B,C) = P(A|B) \Leftrightarrow P(C|B,A) = P(C|B) \tag{4.13}$$

Sind zwei Ereignisse A und C in einem Bayes-Netz unabhängig gegeben die eingetragenen Evidenzen, so heißen sie *d-separiert*. Dies ist genau dann der Fall, wenn für alle Pfade zwischen A und C entweder eine *konvergierende Verbindung* zu einem Ereignis B existiert und weder B noch seine Nachkommen Evidenzen erhalten haben, oder eine *serielle* oder *divergierende Verbindung* zu einem Ereignis B besteht und dessen Eintreten bekannt ist (das heißt der Zustand der betreffenden Zufallsvariable steht fest). Abbildung 4.6 illustriert die unterschiedlichen Arten von Verbindungen.

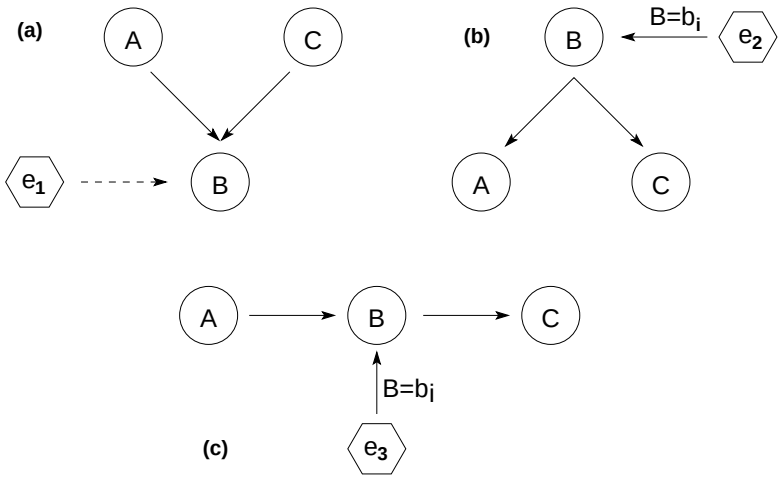


Abb. 4.6: *Konvergenz (a), Divergenz (b) und serielle Verbindung (c) zwischen Ereignissen innerhalb eines Bayes-Netzes. Die eingehende Evidenz e_1 stellt eine Abhängigkeit zwischen den Variablen A und C her, e_2 und e_3 führen zu (bedingter) Unabhängigkeit.*

Um die Wahrscheinlichkeit jedes beliebigen Ereignisses innerhalb eines Bayes-Netzes bei gegebenen Beobachtungen berechnen zu können, wird formal die Gesamtwahrscheinlichkeit (*joint probability*) aller Variablen des Bayes-Netzes benötigt. Sie enthält einen Wahrscheinlichkeitswert für jede Zustandskombination aller in dem betreffenden Netz enthaltenen Variablen; ist also sehr umfangreich. Durch die Ausnutzung bedingter Unabhängigkeiten ist es jedoch möglich, alle benötigten Einträge aus den gegebenen bedingten und *a-priori*-Wahrscheinlichkeiten zu berechnen (*Kettenregel*, vgl. [Jen96] S. 19–20). Sei U die Menge aller Variablen A_i des Bayes-Netzes und seien $B_{i_1}, \ldots, B_{i_k}$ jeweils die Eltern von A_i, dann kann die Gesamtwahrscheinlichkeit $P(U)$ berechnet werden als

$$P(U) = \prod_{i=1}^{m} P(A_i|B_{i_1}, \ldots, B_{i_k}) \tag{4.14}$$

Heutzutage gibt es intelligente Algorithmen (vgl. z.B. [Dec96]), welche die Informationspropagierung in Bayes-Netzen effizient realisieren. Demgegenüber liegt in dieser Arbeit jedoch der Schwerpunkt auf der Entwicklung eines geeigneten Modells für die spezielle Anwendung zur Interpretation von ^{13}C-NMR-Spektren. Auf diesen Aspekt wird daher im folgenden genauer eingegangen, während die Algorithmik nicht Gegenstand der Betrachtungen

ist. Zuvor sollen jedoch einige allgemeine Bemerkungen zum Einsatz von Bayes-Netzen im Bereich der Mustererkennung sowie zu ihrer Rolle im Gesamtkontext der gegenwärtigen Arbeit diesen Abschnitt abschließen.

4.2.3 Bayes-Netze und Mustererkennung

In welcher Weise Bayes-Netze für Aufgaben der Musterklassifikation eingesetzt werden können, ist nicht schwer herzuleiten: Die Klassen ω_i eines Klassifikationsproblems können unmittelbar mit den Zuständen a_i einer Hypothesenvariablen A assoziiert werden. Außerdem sind diejenigen Ereignisse zu repräsentieren, welche den betrachteten Merkmalen entsprechen, und alle Variablen sind in einer kausalen Struktur mit einander zu verbinden, wie es Abschnitt 4.3 zu entnehmen ist. Die Klassifikation kann dann durchgeführt werden, indem man bestimmt, für welchen Zustand die Wahrscheinlichkeit bei den aus den gegebenen Beobachtungen berechneten Merkmalen maximal ist.

In ähnlicher Weise können aber auch Musteranalyseaufgaben realisiert werden. Denkbar wäre z.B. ein Modell mit mehreren Hypothesenvariablen, deren Zustandskombination eine Szenenbeschreibung ergibt. Die Anwendung ist jedoch nicht darauf limitiert, den wahrscheinlichsten Zustand oder die wahrscheinlichste Zustandskombination zu bestimmen: Werden die betreffenden Variablen den vorliegenden Beobachtungen entsprechend instantiiert, so kann die Gesamtwahrscheinlichkeit (*joint probability*) neu berechnet werden. Diese wiederum erlaubt es, jede beliebige Wahrscheinlichkeit innerhalb des Bayes-Netzes im Zusammenhang der gegebenen Beobachtungen zu bestimmen. Durch die so berechneten Wahrscheinlichkeitswerte wird neu gewichtet, wie plausibel das Eintreten jedes durch das Modell erfaßten Ereignisses unter den gegebenen Beobachtungen ist.

Klassische Anwendungen für Bayes-Netze kommen aus dem Bereich der Medizin: In [BO03] werden das System MUNIN, das der Diagnose von Muskel- und Nervenkrankheiten dient, sowie die Anwendung von Bayes-Netzen zur Planung von Insulin-Gaben bei Diabetes diskutiert. Bei beiden Anwendungen liegt die letztendliche Entscheidung in der Hand des Mediziners, den das Computersystem unterstützen soll (*decision support system*). In [Jen96] wird außerdem das System CHILD vorgestellt, das zur Diagnose angeborener Herzkrankheiten dient: Es unterstützt den Facharzt bei der Entscheidung, ein auffälliges Neugeborenes in eine Spezialklinik überweisen zu lassen. Aspekte der Entwicklung des Systems sind in [LTS94] diskutiert.

Die beschriebenen Systeme, die zur Entscheidungsfindung dienen, kann man grundsätzlich in den Bereich der Musterklassifikation einordnen: So kann etwa die Entscheidung, ob ein Neugeborenes in die Spezialklinik überwiesen werden soll oder nicht, als Zwei-Klassen-Problem betrachtet werden. Auf den zweiten Blick wird jedoch deutlich, daß derartige Expertensysteme deutlich über die reine Klassifikationsaufgabe hinausgehen, indem sie nicht einfach die Entscheidung als Resultat der Klassifikation liefern. Vielmehr wird durch die von Bayes-Netzen gebotene Möglichkeit, das maschinelle Schlußfolgern für den Menschen transparent zu machen, dem Experten eine Hilfe für seine eigene Entscheidung gegeben. Diese umfaßt ebenfalls nicht nur die Feststellung einer bestimmten Diagnose, sondern wichtiger sind gewiß die sich daraus ergebenden Maßnahmen – auch Vorsichtsmaßnahmen, wenn die Diagnose nicht mit letzter Gewißheit gestellt werden kann.

Einerseits gehen also die genannten Systeme über reine Klassifikationssysteme hinaus, andererseits sind außerdem mit der Anwendung im Bereich des Bildverstehens, z.B. [BLM89], bzw. des integrierten Bild- und Sprachverstehens wie in [Wac01] Anwendungen in typischen Bereichen der Musteranalyse beschrieben. Die Stärken von Bayes-Netzen qualifizieren sie

also nicht nur für den Einsatz in der Mustererkennung. Es ist darüber hinaus festzustellen, daß sie einen starken Bezug zur symbolischen Ebene erkennen lassen, so daß ihr Potential beim Einsatz für reine Klassifikationsaufgaben kaum ausgereizt wird. Sie eignen sich also besonders für Musteranalysesysteme, in deren Kontext sie Aufgaben sowohl auf symbolischer wie auf subsymbolischer Ebene und im besonderen den Übergang vom einen zum anderen abdecken können. Eine diesen Stärken entsprechende Rolle soll das Bayes-Netz auch im Rahmen des zu entwickelnden Systems erhalten.

4.2.4 Das Bayes-Netz im Kontext der Aufgabenstellung

Das zu entwickelnde Gesamtsystem kann in Anlehnung an das auf Seite 53 gegebene Schema abstrakt wie in Abbildung 4.7 dargestellt werden. Wie bereits erwähnt sind Datenfluß und Aktivierung der einzelnen Komponenten stark system- und anwendungsabhängig. Daher kann ein konkreteres Schema erst nach entsprechender Konzipierung der Einzelmodule (wie es in Kapitel 5 geschehen soll) angegeben werden.

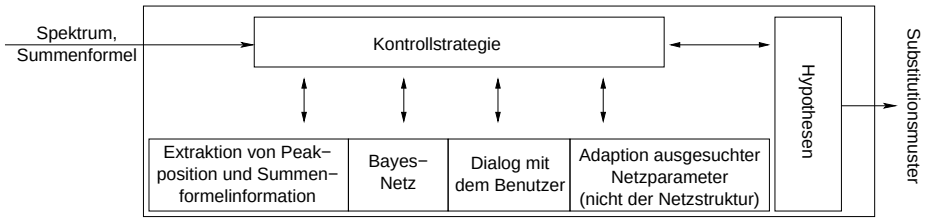


Abb. 4.7: *Schematischer Aufbau des zu entwickelnden Musteranalysesystems zur Erkennung von Substitutionsmustern an Benzolderivaten.*

Das Bayes-Netz repräsentiert dabei in Gestalt des ihm zugrundegelegten Kausalmodells das dem System zur Verfügung stehende Wissen. Auf der funktionellen Seite soll es außerdem als Klassifikator eingesetzt werden. Es liefert für jede Position des Benzolringes die Klasse des dort befindlichen Substituenten basierend auf der chemischen Verschiebung des betreffenden Ringatoms. Aufgrund der besonderen Eignung von Bayes-Netzen für unvollständige Information ist es dabei möglich, potentielle Rückmeldungen aus der Hypothesengenerierung oder bereits vorhandene Klassifikationsergebnisse anderer Positionen im Ring zu berücksichtigen, um so die Verläßlichkeit des Ergebnisses zu verbessern oder eine vorhergehende Einordnung zu revidieren. Essentiell ist in jedem Fall die Entwicklung eines geeigneten Modells, welches die Zusammenhänge der betrachteten Domäne korrekt wiedergibt.

Darüber hinaus werden zur Gewichtung der Kausalzusammenhänge die den modellierten Ereignissen und ihren Verknüpfungen entsprechenden bedingten Wahrscheinlichkeiten benötigt, sowie *a-priori*-Wahrscheinlichkeiten derjenigen Ereignisse, welche nicht von anderen beeinflußt werden. Diese Parameter des Netzes können frei festgelegt, aufgrund theoretischer Überlegungen ermittelt oder empirisch erhoben werden. Bei einem empirischen Vorgehen kann die Gewinnung der Parametrisierung auch als Lernen der Charakteristika der Einsatzumgebung aufgefaßt werden. Ob dies für einige oder alle Wahrscheinlichkeiten sinnvoll und für welche es im besonderen erfolgversprechend ist, muß eine genauere Betrachtung des Modells unter diesem Aspekt zeigen.

Eine Merkmalsextraktion findet hinsichtlich der chemischen Verschiebung des jeweils betrachteten Ringatoms statt. Darüber hinaus kann die Summenformel der Verbindung als gegeben erwartet werden, welche ebenfalls auszuwerten wäre. Im Zusammenhang mit dem Ler-

nen der Systemparametrisierung ist es darüber hinaus von Bedeutung, daß die Eingangsdaten auch Strukturinformation (über die Summenformel hinaus) enthalten können, welche die zu den einzelnen Absorptionen korrespondierenden Atome beschreibt. Bei der Betrachtung der Eingabedaten, insbesondere des Datenformats, ist hier Aufmerksamkeit geboten.

Ein wichtiger Punkt ist auch die Kommunikation mit dem Benutzer, insbesondere die Frage, in welchem Umfang diese stattfindet und wie sie dem angemessen zu gestalten ist. Eine Minimalanforderung ist sicherlich eine verständliche Ausgabe des Analyseergebnisses, darüber hinaus wären jedoch weitere Erklärungen wünschenswert, deren Umfang von einer knappen Benennung der durchgeführten Folgerungs- bzw. Verarbeitungsschritte bis hin zu einer Art sachlichem Dialog mit dem Benutzer reichen könnte. In diesem Zusammenhang wäre es ebenfalls wünschenswert, wenn über die Eingabe der Ausgangsinformation hinaus eine Möglichkeit der Einflußnahme bestünde und Revisionen der Klassifikationsergebnisse oder der Hypothesen nicht nur systemintern, sondern auch durch den Benutzer veranlaßt werden könnten. Gleichwohl ist die Realisierbarkeit einer so umfangreichen Interaktion zusätzlich zu dem gegebenen Schwerpunkt im Bereich der Wisensrepräsentation innerhalb des gegebenen Zeitrahmens fraglich. In Kapitel 5 werden die unbedingt erforderlichen Aspekte der Erklärungskomponente festgelegt, wobei Raum zur Vertiefung dieses Aspekts in weiterführenden Arbeiten bleibt.

Mit dieser Beschreibung der zu bewältigenden Teilgebiete der gegebenen Aufgabenstellung ist nun ein Überblick über die einzelnen Aspekte erreicht, die später im Detail ausgearbeitet werden sollen. Der Schwerpunkt liegt im Bereich der Wissensrepräsentation, das heißt bei der Entwicklung eines Kausalmodells, welches die Zusammenhänge der betrachteten Domäne in geeigneter Weise wiedergibt. Daher befaßt sich der folgende Abschnitt mit Aspekten der Modellentwicklung für Bayes-Netze.

4.3 Kausale Modellierung in Bayes-Netzen

Wie bereits erwähnt setzt sich ein Bayes-Netz aus einem Kausalmodell der Domäne und einem Propagierungsalgorithmus basierend auf Gesetzen der Wahrscheinlichkeitstheorie zusammen. Die Implementierung eines entsprechenden Algorithmus wie in [Dec96] beschrieben steht bereits zur Verfügung. Es bleibt also ein entsprechendes Kausalmodell zu entwickeln, um Bayes-Netze für die gegebene Strukturaufklärungsaufgabe nutzen zu können. Hierfür gibt es kein allgemeines Schema, nach welchem vorzugehen ist, jedoch sind einige Leitfragen bei der Modellentwicklung hilfreich:

- Welche Information ist gesucht?
- Welche Information ist zugänglich?
- Wie sind gesuchte und verfügbare Information in diskrete Zufallsvariablen zu fassen?
- In welchem ursächlichen Zusammenhang stehen gesuchte und beobachtbare Fakten?
- Da in der Regel nicht ausschließlich direkte Abhängigkeiten zwischen den genannten Ereignissen bestehen, welche sonstigen Ereignisse sind innerhalb dieser Zusammenhänge wichtig, und wie lassen sie sich in diskrete Zufallsvariablen fassen?

Während die Entwicklung eines geeigneten Kausalmodells Gegenstand von Kapitel 6 ist, sollen im folgenden einige grundsätzliche Überlegungen angestellt werden, welche im Kontext von Bayes-Netzen hilfreich sind, um auf der Basis eines guten Problemverständnisses der

Domäne zu einem befriedigenden Gesamtmodell zu gelangen. Es schließen sich Strategien an, um in bestimmten Situationen günstige Modellierungsentscheidungen zu treffen.

4.3.1 Grundsätzliche Überlegungen

Im Rahmen des gewählten Versuchsszenarios sollen Substitutionsmuster an Benzolderivaten erkannt werden, das heißt gesucht sind die Typen von Substituenten an den sechs Kohlenstoffatomen eines Benzolringes sowie deren relative Anordnung. Dabei soll das Bayes-Netz im Rahmen eines Musteranalysesystems eine Klassifikation der einzelnen Substituenten leisten. Als Eingabe steht dabei zum einen das breitbandentkoppelte ^{13}C-NMR-Spektrum der Verbindung zur Verfügung, und ferner kann die Summenformel als gegeben angenommen werden. Darüber hinaus kann aber auch Vorwissen über die übrigen Ringpositionen genutzt werden, sofern es vorhanden ist, da über das aromatische System eine wechselseitige Beeinflussung der Elektronendichte an den sechs Positionen stattfindet.

Es kann gleichwohl sinnvoll sein, nicht alle potentiell verfügbaren Informationen, etwa alle fünf übrigen Ringatome, in das Modell zu integrieren: Zwar kann ein Bayes-Netz auch mit unvollständigen Eingabedaten arbeiten, jedoch sollte nicht jede nur irgendwie denkbare Informationsquelle in die Modellierung aufgenommen werden, da das Modell durch die Vielzahl zusätzlicher Variablen zunehmend komplizierter wird, während unter Umständen kaum etwas zur Leistung des Systems beigetragen wird. Insofern sollte nur solche Information berücksichtigt werden, deren Vorhandensein als gegeben vorausgesetzt oder zumindest als erwartbar eingestuft werden kann, oder die die Präzision des Modells nennenswert verbessert. Betreffend die übrigen Ringatome ist z.B. zu bedenken, daß in den seltensten Fällen alle fünf übrigen Substituenten bekannt sein dürften.

Ähnliche Überlegungen spielen auch bei der Einbringung von Zwischenvariablen eine Rolle. Der Versuch, ein möglichst präzises Modell zu entwickeln, kann leicht zu einer kombinatorischen Explosion der benötigten Wahrscheinlichkeiten führen, denn neben der Modellierung der Ereignisse und ihrer Kausalzusammenhänge müssen letztere am Ende immer über bedingte Wahrscheinlichkeiten quantifiziert werden. Im Umkehrschluß bedeutet dies jedoch keineswegs, daß so weit wie möglich auf die Einbringung von Zwischenvariablen verzichtet werden sollte. Vielmehr ist es erforderlich, stets sowohl Aspekte der Modellierungsgenauigkeit, der Kombinatorik hinsichtlich zu betrachtender Zustandskombinationen wie auch der Zugänglichkeit entsprechenden Wissens zur Gestaltung der Zustände und der Gewinnung der Wahrscheinlichkeitswerte im Blick zu behalten.

4.3.2 Strategien

Durch geschicktes Einbringen von Zwischenvariablen kann in verschiedenen Situationen sogar eine Vereinfachung der Netzstruktur hinsichtlich der oben genannten Punkte erreicht werden. Dabei spielen die bereits entwickelten Kausalstrukturen eine Rolle. Im Idealfall sollten sie sich mit der Repräsentation der entsprechenden Ereignisse entwickeln, das heißt bei jedem modellierten Ereignis sollte der Entwickler eine Vorstellung davon besitzen, welche weiteren Ereignisse mit diesem zusammenhängen. Einige nützliche Strategien (vgl. z.B. [Jen96] S. 47 ff.) zur geschickten Einbringung von Zwischenvariablen sollen im folgenden vorgestellt werden.

Es kann z.B. vorkommen, daß es Abhängigkeiten zwischen Ereignissen gibt, ohne daß die Richtung dieser Abhängigkeit feststellbar wäre. Dabei kann es sich etwa um mögliche Konfigurationen von Eigenschaften innerhalb der gegebenen Domäne handeln, von welchen

einige gültig und andere ungültig sind. In einem solchen Fall kann man bedingte Abhängigkeiten ausnutzen, indem man eine zusätzliche, binäre Variable V einführt, deren Ursachen die bewußten Ereignisse sind, zwischen welchen die ungerichtete Abhängigkeit besteht.

Seien diese Ereignisse A, B und C und sei $R(A,B,C)$ eine Relation, welche ihre Abhängigkeit in Zahlen aus $[0,1]$ beschreibt, so ist die bedingte Wahrscheinlichkeitstabelle für V gegeben durch

$$\begin{aligned} P(V = v_0|A,B,C) &= R(A,B,C) \\ P(V = v_1|A,B,C) &= 1 - R(A,B,C) \end{aligned} \tag{4.15}$$

Wird nun die Evidenz $V = v_0$ eingetragen, so sind A, B und C nicht mehr d-separiert, das heißt der Informationsfluß zwischen ihnen wird hergestellt. Dieses Vorgehen setzt jedoch voraus, daß A, B und C keine Eltern haben, oder, falls doch, daß $R(A,B,C)$ nicht probabilistischer Natur ist. Das Vorgehen in anderen Fällen ist der Literatur über *Chain Graphs*, z.B. [Lau96], zu entnehmen.

Ein weiterer Fall, der problematisch sein kann, ist, daß eine Variable V sehr viele Eltern $A_1 \ldots A_n$ hat: In diesem Fall ist für jede Zustandskombination der Eltern eine Wahrscheinlichkeitsverteilung über die Zustände von V erforderlich. Wenn diese empirisch aus einem gegebenen Trainingsdatensatz zu ermitteln ist, kann es leicht aufgrund einer unzureichenden Datenlage zu Schwierigkeiten kommen. Auch kann es vorkommen, daß die einzelnen Wahrscheinlichkeiten $P(V|A_1) \ldots P(V|A_n)$ zugänglich sind, während aber $P(V|A_1 \ldots A_n)$ benötigt wird und über das Zusammenwirken der einzelnen Ursachen nichts bekannt ist.

In beiden Fällen kann zur Vereinfachung *Noisy Or* verwendet werden. Seien $A_1 \ldots A_n$ Ursachen von B und seien alle Variablen binär mit den Zuständen X und O. $A_i = \mathrm{X}$ führt zu $B = \mathrm{X}$, es sei denn ein Inhibitor wirkt, was mit Wahrscheinlichkeit q_i der Fall ist. Unter der Annahme, daß alle Inhibitoren unabhängig seien, kann nun $P(B = \mathrm{O}|A_1 \ldots A_n)$ wie in Gleichung (4.16) angegeben werden. Y sei dabei die Indexmenge derjenigen Variablen, die sich im Zustand X befinden.

$$P(B = \mathrm{X}|A_1 \ldots A_n) = \prod_{j \in Y} q_j \tag{4.16}$$

Voraussetzung für dieses Vorgehen ist, daß B nie im Zustand X sein darf, wenn keine der Ursachen im Zustand X ist, das heißt es muß immer einen Grund für das Eintreten von $B = \mathrm{X}$ geben. Ist dies nicht gegeben, muß zusätzlich ein Hintergrundereignis eingeführt werden, welches die Rolle dieses Grundes übernimmt. Realisiert wird *Noisy Or*, indem die abhängige Variable B n-mal dupliziert wird. Die Kopien $B_1 \ldots B_n$ hängen von den unveränderten Ursachen $A_1 \ldots A_n$ ab. Ein gemeinsames Kind von $B_1 \ldots B_n$ realisiert ein logisches Oder.

Noisy Or kann als Spezialfall eines allgemeineren Vorgehens angesehen werden: *Divorcing* gruppiert einzelne Ursachen eines Ereignisses, um dadurch die Zahl der benötigten Wahrscheinlichkeitsverteilungen zu verringern. Dies geschieht, indem eine Zwischenvariable D zwischen den zu gruppierenden Variablen und ihrem gemeinsamen Kind eingeführt wird. Alle übrigen Eltern bleiben direkt mit dem Kind verbunden. Der Gedanke hinter diesem Vorgehen ist, daß es Zustandskombinationen der zu gruppierenden Variablen gibt, welche sich hinsichtlich ihres Einflusses auf den Zustand des gemeinsamen Kindes äquivalent verhalten. Diese Kombinationen werden auf denselben Zustand der Zwischenvariablen abgebildet.

Die Ersparnis bei den zu ermittelnden Wahrscheinlichkeiten beruht darauf, daß die Zahl der Zustände der Eltern als Faktor in die Zahl der benötigten Wahrscheinlichkeitsverteilungen des Kindes eingeht. Seien zum Beispiel $A_1 \ldots A_4$ Eltern von B und alle diese Variablen dreiwertig, so werden für B 81 Verteilungen benötigt (eine Verteilung je Zustandskombinationen der Eltern). Wird eine Zwischenvariable D mit m Zuständen als Kind von A_1 und A_2

und als Ursache von B eingeführt, so werden nun für B $9 \cdot m$ Verteilungen (Kombinationen von A_3, A_4 und D) anstelle von 81 und zusätzliche 9 Verteilungen für D (Zustandskombinationen von A_1 und A_2) benötigt. Ist $m = 3$, so fällt die Zahl zu ermittelnder Verteilungen von 81 auf 36, auch für $m = 5$ ist mit 54 Verteilungen die Ersparnis immer noch beachtlich.

Damit soll nun der Überblick über die Grundlagen der Mustererkennung im allgemeinen sowie im besonderen aus dem Bereich der Bayes-Netze und der kausalen Modellierung abgeschlossen werden. Vor diesem Hintergrund kann nun klar definiert werden, welche Teilaufgaben zur Bewältigung der gegebenen Aufgabenstellung vonnöten sind.

4.4 Definition des zu entwickelnden Systems

Gegeben ist die Aufgabe, basierend auf protonen-breitbandentkoppelten ^{13}C-NMR-Spektren Substitutionsmuster an Benzolderivaten zu erkennen. Es ist eine symbolische Beschreibung der Substitutionsmuster gesucht, so daß diese Aufgabenstellung in den Bereich der Musteranalyse fällt. Es sind explizit keine Klassen von Substitutionsmustern vorgegeben, wohl aber findet eine Klassifikation der einzelnen Substituenten statt, wobei die Klassen über strukturelle Eigenschaften definiert sind. In einem Folgeschritt wird dann eine symbolische Beschreibung der relativen Anordnung der Substituenten ermittelt.

Der Schwerpunkt innerhalb des zu entwickelnden Musteranalysesystems liegt im Bereich der Wissensrepräsentation, wo ein Bayes-Netz zum Einsatz kommt. Es gewährt eine gewisse Flexibilität hinsichtlich der erforderlichen Eingangsdaten und ermöglicht es gleichzeitig, das streng zweischrittige Schema von Hypothesengenerierung und anschließender Validierung zu verlassen. Durch die Auswertung sowohl kausaler wie auch diagnostischer Evidenzen werden die beiden Schritte integriert und somit eine möglichst frühzeitige Reduktion des Strukturenraums erreicht. Das Bayes-Netz wird als Klassifikator für die einzelnen Substituenten eingesetzt, aus welchen anschließend das Substitutionsmuster aufgebaut werden soll. Die Ergebnisse früherer Klassifikationsschritte sollten als zusätzliche kausale Evidenzen verwendet werden können, um Klassifikationsergebnisse zu untermauern oder, im Falle eines Widerspruchs, deren Revision zu initiieren.

Darüber hinaus sind Module aus den Bereichen Lernen, Dialog und Vorverarbeitung vonnöten, um das System zu vervollständigen. Da vorerst nur der Regelfall eines reibungslosen Verarbeitungsverlaufs betrachtet werden soll, kann dieser Verlauf einfach vorgegeben werden. Zukünftige Arbeiten können sich mit der weiteren Ausarbeitung des Kontrollmoduls befassen. Lern- und Dialogmodul werden vorerst ebenfalls nur in rudimentärer Weise realisiert, da die angemessene Bearbeitung weiterer Schwerpunkte neben der Wissensrepräsentation den Rahmen einer einzelnen Arbeit sprengen würde. Die Vorverarbeitung der chemischen Eingangsdaten (Spektren, Strukturen) ist dagegen ein wichtiger Schritt als solide Basis der weiteren Verarbeitung, sowohl hinsichtlich der Merkmalsextraktion für den Klassifikationsprozeß als auch zur Analyse einer gegebenen Stichprobe mit dem Ziel der Adaption der Systemparameter (Training bzw. Lernen der Charakteristik der Einsatzumgebung).

Ziel der Entwicklung ist ein grundsätzlich einsatzfähiges Gesamtsystem, welches gegeben spektrale und Summenformelinformation eine Hypothese für das Substitutionsmuster generiert. Noch vor der Frage, inwiefern diese Hypothese mit der korrekten Molekülstruktur übereinstimmt, ist die Leistung des Klassifikationsschrittes interessant, da einerseits hier die Ausgangsinformation für die Hypothesengenerierung gewonnen wird und sich andererseits hier der Einfluß der Wissensrepräsentation unmittelbar niederschlägt, welche den Schwerpunkt der Arbeit darstellt.

5 Systemaufbau

Ziel der gegenwärtigen Arbeit ist die Entwicklung eines Systems für die Auswertung von ^{13}C-NMR-Spektren zur Erkennung von Substitutionsmustern an Benzolderivaten. Diese Aufgabe ist dem Feld der Mustererkennung, oder genauer der Musteranalyse, zuzuordnen. Ausgehend von dem in Abschnitt 4.1.2 vorgestellten Schema wird deutlich, daß hierzu mehrere Teilaufgaben zu lösen sind. Diese können jedoch nicht unabhängig von einander betrachtet werden, da sie mit Blick auf das Gesamtziel wie Zahnräder ineinandergreifen.

Es ist daher von zentraler Bedeutung für den erfolgreichen Abschluß der Entwicklungen, ein prinzipielles Schema des Gesamtsystems und seiner einzelnen Bestandteile (Module) zu erarbeiten. Überlegungen hinsichtlich interner Zusammenhänge, das heißt des Zusammenspiels einzelner Module, wie auch externer Faktoren wie etwa den Bedingungen, unter welchen das System entwickelt wird, spielen dabei gleichermaßen eine Rolle.

In Abschnitt 5.1 werden die einzelnen Teilbereiche eines schematischen Musteranalysesystems betrachtet und innerhalb der gegenwärtigen Arbeit gewichtet. Es schließt sich in Abschnitt 5.2 ein Überblick über die in diesem Zusammenhang geltenden Vorgaben in Gestalt relevanter Datenformate und einzusetzender Algorithmen an. Abschnitt 5.3 faßt die Ergebnisse dieser Betrachtungen mit Blick auf die durchzuführenden praktischen Arbeiten zusammen.

5.1 SASCHA

Bei der Entwicklung des gewünschten Systems ist es zunächst von großer Bedeutung, sich einen Überblick über die in unterschiedlichen Stadien der Verarbeitung vorliegenden Daten zu machen. Unmittelbar damit verbunden ist natürlich die Frage, wie diese aus einander hervorgehen. Dabei sind zwei grundsätzliche Fälle zu unterscheiden: der ideale Verarbeitungsverlauf und der Fall des Auftretens von Fehlern oder Unzulänglichkeiten. Nur ersterer soll in den folgenden Abschnitten betrachtet werden.

Ist dieser Optimalfall nachhaltig realisiert, fällt es in den Bereich der Entwicklung der einzelnen Module, sich mit irregulären Vorkommnissen zu beschäftigen. Im besonderen ist es Aufgabe des Kontrollmoduls des Systems, im Falle hinsichtlich ihrer Qualität oder Vollständigkeit unzureichender Zwischenergebnisse entsprechende Maßnahmen zu veranlassen. Deren Umsetzung fällt wiederum in den Bereich der betreffenden Module. Jedes davon wird einzeln für sich realisiert, wobei jedoch der Kontext des Gesamtsystems hinsichtlich der zu liefernden Daten und Inhalte zu beachten ist.

Im folgenden wird zunächst ein Überblick über den Aufbau des Gesamtsystems und seine Anwendungsfälle gegeben. Das System erhält den Namen SASCHA – ein **s**achlich **a**rgumentierendes **S**ystem für die **ch**emische **A**nalyse. Gerade in diesem sehr daten- und funktionsbezogenen Stadium soll dadurch präsent gehalten werden, daß sich das System am menschlichen Vorgehen bei derartigen Strukturaufklärungsaufgaben orientieren soll (wenngleich die Entwicklung eines Systems, das sich tatsächlich in einen sachlich argumentativen Dialog mit dem Benutzer begeben kann, vorerst Utopie bleibt).

5.1.1 Aufbau des Gesamtsystems

In Abbildung 4.7 wurde der Aufbau eines Musteranalysesystems, wie es hier zu entwerfen ist, abstrakt dargestellt. Die einzelnen Bereiche sollen nun in konkrete Module mit entsprechenden Aufgaben gefaßt werden. Darüber hinaus ist zu berücksichtigen, daß neben der Auswertung eines NMR-Spektrums mit dem Ziel der Substitutionsmustererkennung zwei weitere Anwendungsfälle existieren: die Evaluierung des Systems bzw. einzelner Komponenten sowie die Adaption interner Systemparameter. Auch deren Ansprüche sind bei der Ausarbeitung zu berücksichtigen.

Die Extraktion von Peakposition und Summenformelinformation sind dem Vorverarbeitungsmodul zuzuordnen. Es hat die Aufgabe, die relevante Information aus den eingehenden Rohdaten zu selektieren und zur Verfügung zu stellen. Dies schließt für die Anwendungsfälle der Parameteradaption und der Evaluierung auch die Verarbeitung strukturbezogener Daten ein (vgl. Abschnitt 5.1.2).

Um Spektren auswerten zu können umfaßt das Wissensmodul Algorithmen zur Wissenspropagierung sowie das Bayes-Netz selbst. In der Netzstruktur ist dabei das dem System zugrundeliegende Wissen repräsentiert. Seine Parametrisierung dagegen ist nicht Gegenstand des Wissens-, sondern des Lern- bzw. Statistikmoduls. Die Adaption ausgesuchter Netzparameter, die dieses Modul leisten soll, kann als Lernen von Umgebungscharakteristika aus einer Stichprobe aufgefaßt werden (vgl. Abschnitt 5.1.6). Da jedoch die statistische Untersuchung einer Stichprobe, durch welche dies geschieht, eine eher rudimentäre Form des Lernens ist, erscheint die Bezeichnung Statistikmodul treffender.

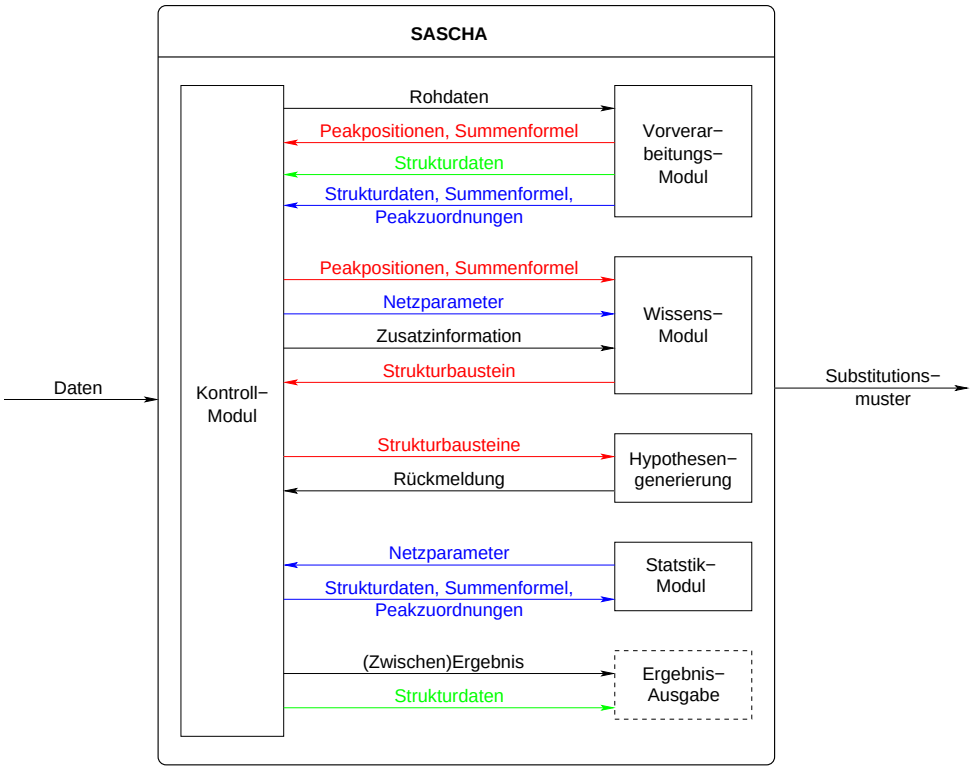


Abb. 5.1: *Aufbau des zu entwickelnden Musteranalysesystems* SASCHA. *Der Datenaustausch zwischen den einzelnen Modulen ist abhängig vom Anwendungsfall; die betreffenden Informationen sind dementsprechend farbig hervorgehoben (Spektrenauswertung=rot, Evaluierung=rot+grün, Parameteradaption=blau).*

Alle Module sollen miteinander über ein Kontrollmodul verbunden werden, das die Bereitstellung der benötigten Daten veranlaßt und diese an die entsprechenden Module weiterleitet. Abbildung 5.1 veranschaulicht den modulbezogenen Systemaufbau sowie den Fluß der benötigten Informationen für die drei Anwendungsfälle Spektrenauswertung, Evaluierung und Parameteradaption. Der Kontakt mit dem Benutzer könnte dabei optimalerweise über ein eigenes Dialogmodul erfolgen; wie jedoch dargestellt werden wird, ist derzeit eine möglichst informative Ausgabe von Ergebnissen und Zwischenergebnissen ausreichend.

In den folgenden Abschnitten wird der Weg von den Eingangsdaten zum Analyseergebnis Schritt für Schritt theoretisch vollzogen. Dies stellt im weiteren Verlauf die Basis für die Definition der benötigten Datenstrukturen und Funktionalitäten dar.

5.1.2 Eingangsdaten

Als Ausgangsinformation steht zunächst das ^{13}C-NMR-Spektrum der Substanz und die Summenformel zur Verfügung. Die Summenformel wird in Gestalt von Paaren aus chemischem Element und Zahl seiner Vorkommen in der unbekannten Substanz zugänglich. Dabei genügt es, diejenigen Elemente zu betrachten, die in organischen Substanzen am häufigsten vorkommen. Die meisten Strukturaufklärungssysteme berücksichtigen Sauerstoff (O), Stickstoff (N), Schwefel (S) sowie die Halogene Fluor (F), Chlor (Cl), Brom (Br) und Iod (I). Diese sieben Elemente soll auch SASCHA betrachten.

Außerdem wird hinsichtlich der spektrale Information die Lage der sechs zu den Benzolringatomen korrespondierenden Peaks benötigt. Diese automatisch zu selektieren würde eine eigene Anwendung für sich darstellen, da der in Frage kommende Bereich sehr groß ist und in der Regel mehr als sechs Peaks innerhalb dieses Bereichs zu finden sind. Für die gegenwärtige Arbeit soll vorerst vorausgesetzt werden, daß bekannt ist, welche die relevanten Peaks sind. Die automatische Auswahl kann Gegenstand zukünftiger Arbeiten sein.

Für die Anwendungsfälle der Parameteradaption und die Evaluierung sind hingegen interpretierte Spektraldaten erforderlich, das heißt mit der spektroskopischen Information muß Strukturinformation einhergehen. Mit Blick auf die Klassifikation der einzelnen Positionen des Benzolringes ist es sogar erforderlich, daß im Detail Verknüpfungen zwischen den einzelnen Peaks im Spektrum und den sie verursachenden Kohlenstoffatomen in den Strukturdaten gegeben sind. Diesbezügliche Überlegungen hinsichtlich des Formats der Rohdaten sind Gegenstand von Abschnitt 5.2.

Es wird also eine geeignete Datenstruktur für Strukturdaten benötigt, welche ihre Zuordnung zur spektralen Information bewahrt. Dabei muß jedes Atom eindeutig identifiziert werden können und mit Informationen über sein chemisches Element sowie seine chemische Verschiebung (sofern es sich um Kohlenstoff handelt) versehen sein. Zur Klassifikation hinsichtlich struktureller Merkmale (vgl. Abschnitt 3.1.5 zur dem zugrundegelegten Idee des HOSE-Codes) müssen auch die Bindungen repräsentiert werden: Ein Molekül ist in diesem Sinne aus Strukturelementen aufgebaut, die neben den beschriebenen Informationen bezüglich einzelner Atome auch deren Hybridisierung erfassen und die benachbarten Strukturelemente (Bindungspartner) aufführen. Um die Genauigkeit der Strukturrepräsentation nicht im Vorhinein zu limitieren, ist an dieser Stelle eine Repräsentation des gesamten Moleküls und nicht nur des Benzolringes und seiner auf eine bestimmte Anzahl von Sphären begrenzten Umgebung gefordert. Die Realisierung eines Vorverarbeitungsmoduls, welches die genannten Aufgaben erfüllt, wird in Kapitel 7 beschrieben.

5.1.3 Klassifikation der Ringatome

Die vom Vorverarbeitungsmodul gelieferte Summenformel- und Peakinformation dient als Eingabe für die Klassifikation der einzelnen Positionen des Benzolrings. Nacheinander werden die beteiligten chemischen Elemente und ihre Multiplizitäten mit jeweils einer Peakposition als Evidenzen in das Bayes-Netz eingetragen. Dieses liefert die Wahrscheinlichkeitsverteilung der Zustände der Hypothesenvariable. Da der Algorithmus zur Durchführung der notwendigen Berechnungen in der im Rahmen von [Wac01] durchgeführten Implementierung von Sven Wachsmuth zur Verfügung steht, sind in diesem Bereich keine Vorarbeiten erforderlich. Die entsprechenden Fakten werden in Abschnitt 5.2 dargestellt.

Der Schwerpunkt der Entwicklung liegt vielmehr im Bereich der Modellierung, das heißt in der Wiedergabe der Gesetzmäßigkeiten und Zusammenhänge zwischen Spektrum und Struktur. Dabei ist insbesondere zu beachten, daß nicht nur der *ipso*-Substituent, sondern auch die Substituenten in den übrigen Ringpositionen die chemische Verschiebung des *ipso*-Kohlenstoffs beeinflussen. Dies soll ausgenutzt werden, um das Klassifikationsergebnis einer Position zu bestätigen oder zu revidieren, wenn Ergebnisse anderer Positionen vorliegen. Diese können als Vorwissen genutzt werden, um den Raum infragekommender Strukturen möglichst frühzeitig zu beschränken und den Aufbau des Substitutionsmusters aus den einzelnen Klassifikationsergebnissen zu erleichtern. Die Modellentwicklung ist Gegenstand von Kapitel 6.

Der Anwendungsfall der Parameteradaption stellt hinsichtlich der Datenstrukturen des Wissensmoduls keine zusätzlichen Anforderungen. Es ist zwar erforderlich, die Repräsentation des Bayes-Netzes einlesen und hinsichtlich der bedingten Wahrscheinlichkeiten bearbeiten zu können, die Überlegungen zu den in diesem Zusammenhang benötigten Funktionalitäten und Datenstrukturen sind jedoch Gegenstand des Lernmoduls (vgl. Abschnitt 5.1.6).

Hinsichtlich der Evaluation des Klassifikationsschrittes ist zu bemerken, daß das Klassifikationsergebnis mit den zur Eingabe gehörigen Strukturdaten verglichen werden muß, um seine Korrektheit zu beurteilen. Außerdem ist zu beachten, daß die Ausgabe darüber hinaus als Eingabe für den Aufbau des Substitutionsmusters genutzt wird. Die Ausgabe des Bayes-Netzes sollte also grundsätzlich eine Form haben, welche der Sichtweise chemischer Strukturen gerecht wird.

5.1.4 Aufbau des Benzolringes

Nach der Klassifikation der einzelnen Peakpositionen wird aus diesen das gesamte Substitutionsmuster zusammengesetzt. In diesem ausschließlich auf Molekülstrukturen bezogenen Schritt ist, wie bereits hinsichtlich der Ausgabe des Bayes-Netzes erwähnt, eine entsprechende Repräsentation zu wählen. Zudem sind mit Blick auf den Anwendungsfall der Evaluation auch die gegebenen Strukturdaten in der gleichen Weise darzustellen: So kann leicht festgestellt werden, ob die gefundene Struktur der tatsächlich im Molekül vorkommenden entspricht. Die Parameteradaption hat demgegenüber keine Berührungspunkte mit der Hypothesengenerierung.

Wird in der Modellierung des Bayes-Netzes der wechselseitige Einfluß der einzelnen Ringpositionen auf einander berücksichtigt, so kann nach der Gewinnung eines Klassifikationsresultats dasselbe als zusätzliche Evidenz erneut eingegeben werden, um eine Klassifikation hinsichtlich der Nachbarpositionen innerhalb des Rings zu erhalten. Auf diese Weise lassen sich dreigliedrige Ringsegmente aufbauen, deren Überschneidungen bei der Hypothesengenerierung Informationen für den Aufbau des Gesamtringes liefern. Außerdem kann diese

Information im Falle eines Widerspruchs zur Revision früherer Klassifikationsergebnisse herangezogen werden. In beiden Fällen ist es hilfreich, den einzelnen Strukturfragmenten und den Hypothesen für das Substitutionsmuster die jeweils zugrundegelegten Annahmen und Wahrscheinlichkeitsbewertungen im Rahmen der Klassifikation zuzuordnen.

Durch eine geeignete Aufbereitung der Systemantwort wird deren Informationsgehalt für den menschlichen Benutzer zugänglich gemacht, und auch die Leistungsfähigkeit des Systems kann bei entsprechender Ausgabe, etwa der Erfolgsbewertung zusammen mit den Vergleichsdaten des Analyseergebnisses und der tatsächlichen Molekülstruktur, besser beurteilt und im Detail analysiert werden.

5.1.5 Dialog mit dem Benutzer

Hinsichtlich der Ausarbeitung eines Dialog- oder Erklärungsmoduls ist es einerseits wünschenswert, möglichst ausführliche Erklärungen zu den Ergebnissen des Systems zu liefern, um die Systemantwort für den Benutzer transparent und nachvollziehbar zu machen. Andererseits bedeutet dies aber einen sehr hohen Aufwand bei der Entwicklung, da die Erläuterungen verständlich sein müssen und erst dann ihren vollen Sinn entfalten, wenn dem Benutzer auch die Möglichkeit gegeben wird, darauf zu reagieren und entsprechenden Einfluß auf den Analyseprozeß zu nehmen. Nicht ohne Grund erhält das Feld der Mensch-Maschine-Kommunikation in Informatik, Sprach- und Sozialwissenschaft gleichermaßen viel Aufmerksamkeit. Vor diesem Hintergrund soll zunächst der Umfang der Erklärungskomponente in Entwicklung und Realisierung auf das notwendige Minimum beschränkt werden, da die Bearbeitung eines zweiten, derart vielschichtigen Schwerpunkts neben dem Aspekt der Wissensrepräsentation den Rahmen einer einzelnen Arbeit sprengen würde. Erweiterungen können Gegenstand zukünftiger Arbeiten sein.

Verläuft die Datenverarbeitung regulär, so ist keine Einflußnahme durch den Benutzer nötig, und eine reine Erklärung der Systemantwort ist ausreichend. Das gleiche gilt für die Evaluation des Systems, wobei hier jedoch eine andere, möglicherweise technischere oder umfangreichere Ausgabe der Erklärungen erwünscht ist. Der Anwendungsfall der Parameteradaption kommt wiederum mit wenigen Erklärungen aus, welche sich nun aber auf die angepaßten Parameter und nicht die berechneten Hypothesen beziehen. Die zu realisierenden Funktionalitäten beschränken sich also vorerst darauf, die von SASCHA berechneten Ergebnisse verständlich und dem Anwendungsfall angemessen auszugeben.

5.1.6 Lernen

Der Aspekt des Lernens soll für SASCHA nur untergeordnete Bedeutung erhalten. Im Falle der regulären Spektrenauswertung und der Evaluation soll er nicht zum Tragen kommen. Lediglich die Adaption der Systemparameter kann als Lernen angesehen werden.

Die Struktur des Bayes-Netzes, welches das dem System zugrundegelegte Wissen repräsentiert, wird explizit vorgegeben; seine Parameter in Gestalt der bedingten Wahrscheinlichkeiten sollen jedoch basierend auf einer repräsentativen klassifizierten Stichprobe angepaßt werden. Auf diese Weise wird die Charakteristik der durch die Stichprobe repräsentierten Einsatzumgebung gelernt.

Das Lern- oder Statistikmodul greift dabei auf die durch das Vorverarbeitungsmodul gelieferten Spektral- und Strukturdaten zurück. Sie werden hinsichtlich der für das Kausalmodell der Domäne relevanten Eigenschaften untersucht, und die darauf basierenden relativen Häufigkeiten werden zur Approximation der entsprechenden bedingten Wahrscheinlichkeiten des

Bayes-Netzes verwendet. Dabei sollen die zu adaptierenden Wahrscheinlichkeiten einzeln auswählbar sein. Wird eine leere Stichprobe verwendet, so resultiert dies in einer Gleichverteilung der betreffenden Wahrscheinlichkeiten, das heißt, das Bayes-Netz kann auf diese Weise gezielt veranlaßt werden, die vorherige Quantifizierung der Kausalzusammenhänge wieder zu „vergessen“. Kapitel 8 beschreibt, wie die Adaption der Netzparameter durchgeführt wird.

5.2 Datenformate und Algorithmen

Neben den durch die Definition der Aufgabenstellung gegebenen Anforderungen, die im vorangegangenen Abschnitt beschrieben wurden, spielen auch die vorhandenen technischen Gegebenheiten sowie bestehende Standards der Datenverarbeitung innerhalb des gegenwärtigen Arbeitsfeldes eine Rolle für die Realisierung. In diesem Zusammenhang sind die zu verarbeitenden chemischen Daten, das in Gestalt eines Bayes-Netzes zu speichernde Wissen und die Informationspropagierung in Bayes-Netzen zu betrachten. Im folgenden sollen daher das *JCAMP-Format* für chemische Daten sowie betreffend Bayes-Netze das BNIF-Format zur Repräsentation und der *Bucket-Elimination*-Algorithmus zur Informationspropagierung vorgestellt werden.

5.2.1 Das JCAMP-Format für chemische Daten

Das JCAMP[1] ist eine nicht gewinnorientierte Organisation, die sich mit der Generierung, Sammlung, Evaluierung und Bearbeitung atomarer und molekularer physikalischer Daten befaßt. (Seit 1995 hat jedoch die IUPAC[2] den Verantwortungsbereich des JCAMP übernommen.) 1983 wurde eine Arbeitsgruppe zur Portabilität spektraler Daten unter der Leitung von Paul A. Wilks, Jr. ins Leben gerufen, um es den Anwendern spektroskopischer Systeme zu ermöglichen, ihre Daten zwischen unterschiedlichen Systemen zu transferieren, was bis dahin aufgrund der unterschiedlichen proprietären Datenformate nahezu unmöglich war. In dieser Gruppe wurden das JCAMP-DX-Format für spektroskopische Daten [MWJ88, DL93] und das JCAMP-CS-Format für Strukturdaten [GHH+91] entwickelt.

Von besonderem Vorteil für die gegenwärtige Arbeit ist die Möglichkeit einer Verbindung von Spektral- und zugehöriger Strukturinformation, wie sie für die Anwendungsfälle der Evaluation und der Adaption der Systemparameter benötigt wird: Das zeitlich später entwickelte JCAMP-CS-Format wurde auf der technischen Seite so weit wie möglich an die bereits bestehenden JCAMP-DX-Konventionen angelehnt und und erlaubt die Verbindung von spektroskopischen und Strukturdaten bis auf die Ebene der Zuordnung von Chromophoren zu Absorptionen. Maßgeblich ist seitens der spektroskopischen Daten die Spezifikation des Formats JCAMP-DX 5.00 [DL93], welche ihre Vorläufer um NMR-bezogene Datenfelder ergänzt. Strukturdaten werden im Format JCAMP-CS 4.24 [GHH+91] angegeben.

Obwohl die JCAMP-Datenformate weit verbreitet sind, können jedoch Schwierigkeiten auftreten. Da sich die JCAMP Arbeitsgruppe bei ihrer Arbeit stark an den Bedürfnissen menschlicher Benutzer orientierte, ist die Struktur von JCAMP-Dateien einfach; im Grunde handelt es sich um eine schlichte Folge von Bezeichner-Wert-Paaren (Datenfeldern). Dadurch sind die Daten zwar für den Menschen leicht verständlich zu lesen, man muß sich jedoch bewußt sein, daß das Format primär dem Daten*austausch* dienen sollte, so daß Aspekte der automatisierten Daten*verarbeitung* in den Hintergrund rückten. Als Folge davon sind we-

[1] *Joint Committee on Atomic and Molecular Physical Data*
[2] *International Union of Pure and Applied Chemistry*

der alle Informationen einer JCAMP-Datei für die maschinelle Verarbeitung geeignet, noch dafür bestimmt. Auch dehnen viele Anwender die Definitionen des Formats sehr stark, so daß es trotz der Bezugnahme auf den JCAMP-Standard zu Problemen bei der Verarbeitung der betreffenden Daten kommen kann. Im Einzelfall ist daher zu prüfen, ob die verwendeten Daten dem Standard entsprechen oder wo und wie sie davon abweichen.

Daher ist im Zusammenhang der Entwicklung des Vorverarbeitungsmoduls, welches JCAMP-Daten verarbeiten soll, eine eingehendere Beschäftigung mit dem Format der vorliegenden Daten nötig. An dieser Stelle soll dementsprechend auch eine genauere Beschreibung des Aufbaus von JCAMP-Dateien erfolgen (vgl. Kapitel 7).

5.2.2 Der Bucket-Elimination-Algorithmus

Für SASCHA steht ein von Sven Wachsmuth im Rahmen von [Wac01] implementierter Algorithmus für die notwendigen Berechnungen innerhalb des Bayes-Netzes zur Verfügung, welcher auf Rina Dechters *Bucket-Elimination*-Konzept [Dec96] basiert. Dieses stellt einen algorithmischen Rahmen dar, welcher die Prinzipien des *Dynamic Programming* dergestalt verallgemeinert, daß sie Algorithmen unterschiedlichster, komplexer Anwendungen, im besonderen des probabilistischen Schließens, gerecht werden. Der folgende Überblick über *Bucket-Elimination*-Verfahren soll sich auf die prinzipielle Funktionsweise solcher Algorithmen beschränken. Zusätzliche Überlegungen zur Verbesserung der Verfahren, etwa durch *Conditioning*, sind der Originalarbeit [Dec96] zu entnehmen, ebenso wie Komplexitätsbetrachtungen und die Beziehung zu anderen Methodengattungen.

Die Vorteile des *Bucket-Elimination*-Prinzips sind seine Einfachheit und Universalität: Da das Verfahren auf der einen Seite ohne umfangreiche Hilfsdefinitionen und aufwendige Konstrukte auskommt, ist es auch ohne intensive Einarbeitung rasch zu verstehen und anzuwenden. Auf der anderen Seite lassen sich viele existierende Algorithmen als *Bucket-Elimination*-Algorithmen oder Spezialisierungen derselben auffassen. So können diese Konzepte nicht nur ohne viel Umstand in neue Anwendungsfelder eingebracht werden, welche sich über diese gemeinsame Sichtweise gegenseitig bereichern können: Durch die Übertragung bestimmter Techniken in *Bucket-Elimination*-Verfahren können diese über die Grenzen ihrer klassischen Anwendung hinweg eingesetzt und so Wissen und Erfahrungen ausgetauscht werden.

An dieser Stelle soll in erster Linie die Verwendung für Bayes-Netze als spezielle Methodik im Bereich des probabilistischen Schließens betrachtet werden. Hier sind drei Hauptaufgaben zu unterscheiden: *Belief Updating* ist die Berechnung der *a-posteriori*-Wahrscheinlichkeiten aller Aussagen innerhalb des Netzes bei gegebenen Beobachtungen. Unter der Bestimmung der *mpe* (*most probable explanation*) versteht man, bei gegebenen Beobachtungen bezüglich einiger Variablen die wahrscheinlichste Belegung der übrigen Variablen zu bestimmen. Der dritte Fall ist die Bestimmung der *map* (*maximum a-posteriori hypothesis*): Dies ist die Bestimmung der wahrscheinlichsten Zustandskombination der Hypothesenvariablen bei gegebenen Beobachtungen.

Dechter beschreibt in [Dec96] einen allgemeinen Variablen-Eliminierungs-Algorithmus für *Belief Updating*, die wohl häufigste Berechnungsaufgabe im Zusammenhang mit Bayes-Netzen: Mit dem Eintreffen neuer Evidenzen sind die Wahrscheinlichkeitsverteilungen der Variablen zu aktualisieren. *Buckets* sind dabei ein organisatorisches Konzept. Es setzt eine Ordnung auf den Variablen des Netzes voraus. Diese ist so zu wählen, daß die Variable, deren Verteilung aktualisiert werden soll, den untersten Rang innerhalb der Ordnung einnimmt.

Der Algorithmus initialisiert dann die *Buckets* in absteigender Reihenfolge korrespondierend zu den Variablen und ihrer Ordnung: In den *Bucket* der Variable mit dem höchsten Rang fallen all diejenigen bedingten Wahrscheinlichkeiten, die diese Variable als Ursache oder als

abhängige Variable enthalten. Dasselbe Prinzip wird jeweils für die höchste noch verbleibende Variable wiederholt, in deren *Bucket* alle noch nicht eingeordneten Wahrscheinlichkeiten fallen, die diese Variable enthalten. Liegen Beobachtungen bezüglich einer Variablen vor, so werden diese in den *Bucket* der betreffenden Variable eingetragen.

Die Variableneliminierung geschieht absteigend durch Summation über alle Zustände, oder im Fall des Vorliegens einer Beobachtung durch alleinige Betrachtung des entsprechenden Zustands. Die mathematischen Umformungen, die dem zugrundeliegen, sind in der Originalarbeit [Dec96] ausführlich dargestellt: Elimination einer Variablen V_i liefert eine Funktion λ_i, die durch die nicht eliminierten Variablen der entsprechenden Verteilung parametrisiert ist, und die die Summenbildung vornimmt (i entspricht dem Rang der Variablen in der gegebenen Ordnung). Sie wird in den *Bucket* der höchsten Parametervariablen eingeordnet und geht dort als Faktor in die im entsprechenden Eliminierungsschritt durchzuführende Summenbildung ein. Im letzten *Bucket* kann zuletzt durch Multiplikation ohne Eliminierung das *Belief Updating* durchgeführt werden, und man erhält die gewünschten *a-posteriori*-Wahrscheinlichkeiten.

Der Fall der *mpe*-Bestimmung, zum Beispiel in der Medizin die Bestimmung der wahrscheinlichsten Diagnose bei gegebenen Symptomen, unterscheidet sich im Verfahren kaum von *Belief Updating*. Es findet lediglich anstelle der Summation eine Maximierung statt. Daneben wird zu jeder berechneten maximalen Wahrscheinlichkeit der ihr zugehörige Zustand gespeichert. Nach der absteigenden Bearbeitung aller *Buckets* werden in einem zweiten, aufsteigenden Durchlauf diese Zustände den betreffenden Variablen zugeordnet und dadurch die gewünschte *mpe* erstellt. Die Berechnung einer *map* ist im Prinzip eine Kombination beider Algorithmen, das heißt einige Variablen werden durch Summation und einige durch Maximierung eliminiert. Gesucht ist die wahrscheinlichste Zustandskombination einer Untermenge der Variablen, und zwar der Hypothesenvariablen. Die betreffenden Variablen müssen in der verwendeten Ordnung die untersten Ränge einnehmen. Nachdem für alle höheren Variablen in den betreffenden *Buckets* wie beim *Belief Updating* die Summation durchgeführt wurde, wird für die Hypothesenvariablen wie bei der *mpe*-Bestimmung maximiert. Anschließend werden ihnen die Zustände zugewiesen, für welche die maximalen Wahrscheinlichkeiten berechnet wurden.

5.2.3 Das BNIF-Format für Bayes-Netze

Die zur Verfügung stehende Implementierung des *Bucket-Elimination*-Algorithmus für *Belief Updating* arbeitet mit Bayes-Netzen im *BNIF-Format*[3]. Daher ist dieses Dateiformat für die Speicherung der Wissensrepräsentation als Bayes-Netz implizit vorgegeben und soll hier kurz vorgestellt werden.

Ähnlich wie beim JCAMP-Format für chemische Daten begann auch hier die Entwicklung damit, daß ein gemeinsames Datenformat es Wissenschaftlern in unterschiedlichen Projekten, Institutionen oder Anwendungsfeldern erleichtern sollte, ihr Wissen und ihre Erfahrungen auszutauschen. Im Zuge der UAI[4]-Konferenz 1996 kondensierten sich die Bestrebungen zur Definition eines gemeinsamen Standards im Zusammenschluß einer Gruppe von Forschern (Cozman, Druzdzel, Garcia; vgl. [Coz]), welche die Weiterentwicklung des erarbeiteten Vorschlags kanalisierten.

Gleichwohl stellte sich im Zuge der darauffolgenden Überlegungen und Diskussionen heraus, daß es mit dem derzeitigen Stand der Forschung schwerlich möglich war, den Bedürf-

[3] *Bayesian Network Interchange Format*
[4] *Uncertainty and Artificial Intelligence*

nissen und Anforderungen jedes einzelnen Bereiches gerecht zu werden. Auf der UAI Konferenz 1998 kam man daher schließlich überein, auf XML[5] zurückzugreifen: Damit war ein bestehender Standard verfügbar, welcher mit seiner Erweiterbarkeit als expliziter Möglichkeit zur Entwicklung von Spezialisierungen nicht nur den Ansprüchen der Verwendung für Bayes-Netze entgegenkam: XML erlaubte es außerdem, unterschiedliche Spezialisierungen für unterschiedliche Ansprüche zu entwickeln, welche jedoch durch ihre XML-Basiertheit über eine gemeinsame Basis verfügten.

Abseits der Entwicklung hin zu XML-basierten Formaten haben sich jedoch auch die ursprünglichen Vorschläge eines BNIF-Formats erhalten und bieten nach wie vor eine Grundlage für die Speicherung und Verarbeitung von Wissen in Form eines Bayes-Netzes. Das Format, wie es etwa in [Coz] beschrieben wird, ist einfach gehalten und sowohl für den Menschen als auch für die automatische Verarbeitung gut lesbar. Die enthaltene Information ist in *Blocks* unterteilt. Ein Block besteht aus einem Schlüsselwort, einem Bezeichner und einer Attributsliste. Die Schlüsselwörter `network`, `variable` und `probability` dienen der Unterscheidung von Informationen, die auf das Bayes-Netz als ganzes, einzelne Variablen oder die einzelnen bedingten Wahrscheinlichkeiten bezogen sind.

Die eigentliche Information befindet sich in den Listen von Attributen der einzelnen Blocks. Es gibt unterschiedliche Typen von Attributen je nach Typ des Blocks: So sind etwa für den durch das Schlüsselwort `network` eingeleiteten Kopf nur Attribute des Typs `property` vorgesehen, welche allgemeine Eigenschaften des vorliegenden Bayes-Netzes wie Autor oder Versionsnummer beschreiben. Durch `variable` eingeleitete Variablen-Blocks enthalten immer ein Typ-Attribut (`type`), welches für diskrete Variablen neben dem Eintrag `discrete` die Zahl und Benennung der Zustände der betreffenden Variablen enthält.

Blocks die durch das Schlüsselwort `probability` eingeleitet werden und welche Information enthalten, die die bedingten Wahrscheinlichkeiten innerhalb des Netzes betrifft, können Attribute der Arten `table`, `default` und sogenannte Eingabeattribute enthalten. Es handelt sich bei allen dreien um Formen der Eingabe von Wahrscheinlichkeitswerten. Darüber hinaus legen `probability`-Blocks die Topologie des Bayes-Netzes fest: Ihre Namen sind nicht frei wählbar, sondern bestehen aus einem Tupel von Variablennamen, das der Schreibweise bedingter Wahrscheinlichkeiten entspricht.

Es ist jedoch zu beachten, daß es sich bei dem beschriebenen Format um einen Vorschlag und nicht um einen endgültigen Standard handelt. Im Detail können daher Abweichungen vorkommen, welche im Rahmen der Implementierung zu berücksichtigen sind. Maßgeblich für die gegenwärtige Arbeit ist die durch die gegebene Implementierung des *Bucket-Elimination*-Algorithmus festgelegte Variante.

Nachdem nun der Überblick über die technischen Rahmenbedingungen in Gestalt zu verwendender Algorithmen und Datenformate abgeschlossen ist, bleibt eine Übersicht der im Rahmen der gegebenen Problemstellung zu bewältigenden Einzelaufgaben zu erstellen.

5.3 Durchzuführende Entwicklungsarbeiten

Die zu lösenden Aufgaben zur Entwicklung eines Systems zur Interpretation von ^{13}C-NMR-Spektren mit dem Ziel der Substitutionsmustererkennung an Benzolderivaten sind nun vom Prinzip her umrissen. Sie sollen an dieser Stelle zusammengefaßt werden; im einzelnen sind sie bei der Realisierung der entsprechenden Systembestandteile gegebenenfalls zu konkretisieren oder zu ergänzen. Wie in Abbildung 5.1 dargestellt sind alle Module über ein steu-

[5] *Extensible Markup Language*

erndes Kontrollmodul mit einander verbunden. Es hat die Aufgabe, Informationen von den übrigen Komponenten anzufordern und diese Information zur weiteren Verarbeitung an andere Module weiterzuleiten. Daher ist es bedeutsam, jede Komponente mit ihren Funktionen und Datenstrukturen im Zusammenhang mit den übrigen Modulen zu betrachten.

Aufgabe des Vorverarbeitungsmoduls ist die Bereitstellung der chemischen Ausgangsdaten. Das Einlesen der Rohdaten, aus welchen sie gewonnen werden, kann unabhängig vom Rest des Systems behandelt werden, da kein anderes Modul in Kontakt mit den Rohdaten tritt. Die Form, in welche spektroskopische und Strukturdaten vom Vorverarbeitungsmodul gebracht werden, muß jedoch auf andere Module Rücksicht nehmen: Spektroskopische Daten und die Summenformel sollen vom Bayes-Netz innerhalb des Wissensmoduls zu struktureller Information weiterverarbeitet werden, die wiederum an die Hypothesengenerierung weitergeleitet wird. Strukturbezogene Daten werden im Anwendungsfall der Evaluation mit beidem in Verbindung gebracht.

Für das Statistikmodul werden ebenfalls Strukturdaten in Verbindung mit Spektraldaten und Summenformelinformation benötigt. Daneben ist auch die Repräsentation des Bayes-Netzes im BNIF-Format relevant, da die Hauptaufgabe des Statistikmoduls in der Adaption der Netzparameter besteht. Nur in diesem Anwendungsfall ist das Statistikmodul aktiv. Die Erklärungskomponente schließlich bringt hinsichtlich der Datenstrukturen keine weiteren Ansprüche mit. Sie dient vorerst nur der Ausgabe von Zwischenergebnissen und des Endergebnisses, muß also die ihr zugeleiteten Daten zwar zu einer entsprechenden informativen Ausgabe verarbeiten können, stellt jedoch keine zusätzlichen Anforderungen an deren Form.

Die auf den genannten Daten zu entwickelnden Funktionen ergeben sich aus den Aufgaben der einzelnen Module. Das Vorverarbeitungsmodul soll JCAMP-Daten einlesen und in einem für die weitere Verarbeitung günstigen Format bereitstellen. Die Aufgabe des Bayes-Netzes, des Kernbestandteils des Wissensmoduls, ist die Klassifikation von Substituenten basierend auf spektralen Befunden. Da die nötigen Funktionen zur Informationspropagierung bereits zur Verfügung stehen, sind lediglich noch Funktionen zum Stellen von Anfragen und zur Weitergabe der Netzantwort nötig. Die Hypothesengenerierung soll anschließend die eingehenden Strukturfragmente zu einem Substitutionsmuster zusammenfügen. Die Verarbeitungsergebnisse werden an den Benutzer ausgegeben.

Aufgabe des Statistikmoduls ist es, durch statistische Untersuchung einer klassifizierten, charakteristischen Stichprobe die bedingten Wahrscheinlichkeiten des Bayes-Netzes anzupassen. Hierbei ist zu berücksichtigen, daß auch die für eine zu bestimmende bedingte Wahrscheinlichkeit relevanten Befunde in den chemischen Daten erkannt werden müssen.

Die Besonderheit des neuen Musteranalysesystems SASCHA ist der Ansatzpunkt einer wissensbasierten Modellierung. Sie soll es ermöglichen, nicht nur den Raum in Frage kommender Lösungen (Substitutionsmuster) möglichst frühzeitig zu beschränken, indem kausale (Summenformel) und diagnostische (chemische Verschiebung) Evidenzen integriert ausgewertet werden, sondern darüber hinaus eine Revision von Teilergebnissen zulassen und somit durch deren wechselweises Zusammenwirken eine starke Verflechtung von Hypothesengenerierung und -validierung zu erreichen.

Hinsichtlich der Evaluierung des Systems ist dabei zu beachten, daß der Schwerpunkt der Entwicklung vorerst ausschließlich auf dem Bereich der Wissensrepräsentation liegt. Insofern ist in erster Linie die Leistung des Klassifikators für die sechs einzelnen Positionen des Benzolringes interessant. Als Klassifikator wird ein Bayes-Netz eingesetzt, welches auf einem Kausalmodell der chemisch-physikalischen Moleküleigenschaften, die für die NMR-Spektroskopie eine Rolle spielen, basiert. Somit hat die Wissensrepräsentation in diesem

Schritt einen unmittelbaren Einfluß. Außerdem liefert er die Ausgangsinformation für die nachfolgende Hypothesengenerierung, welche von einer gewissenhaften und wohl analysierten Realisierung des Bayes-Netzes nur profitieren kann.

Im Vergleich zu bestehenden Systemen sei noch einmal darauf hingewiesen, daß die Aufgabenstellung bewußt wesentlich enger gefaßt ist: Gegenüber der Menge der organischen Moleküle wird nur die Teilmenge der Benzolderivate betrachtet, und gegenüber der Aufklärung der Konstitution ist nur die Beschreibung des Substitutionsmusters gesucht. Die eingeschränkte Aufgabenstellung dient zur Belegung der grundsätzlichen Eignung des gewählten Ansatzes für die automatische Strukturaufklärung und ermöglicht eine detaillierte Analyse einzelner Entwicklungsschritte und Entscheidungen. Die Dokumentation derartiger Erkenntnisse ist für zukünftige Weiterentwicklungen über das exemplarische Szenario der Benzolderivate hinaus von großem Wert.

6 Modellentwicklung

Da der Schwerpunkt der vorliegenden Arbeit auf der expliziten Einbringung von Expertenwissen liegt, ist es sinnvoll, die Entwicklung ebenfalls im Bereich des Wissensmoduls zu beginnen. Es soll ein kausales Modell der Gesetzmäßigkeiten der betrachteten Domäne der Spektroskopie entwickelt und als Bayes-Netz repräsentiert werden. Dieses Bayes-Netz dient im späteren System dazu, basierend auf spektroskopischen Befunden, der Summenformel sowie etwaiger bereits ermittelter Zwischenergebnisse die Positionen des Benzolringes hinsichtlich ihrer Substituenten zu klassifizieren.

In Kapitel 4 wurden neben Grundlagen der Mustererkennung und des Einsatzes von Bayes-Netzen in diesem Bereich insbesondere Grundgedanken der Modellentwicklung vorgestellt. Die folgenden Überlegungen, die hieran anknüpfen, sollen sich dabei auf eine qualitative Wiedergabe der Kausalzusammenhänge der Domäne beschränken, die Quantifizierung des so repräsentierten Wissens ist Gegenstand von Kapitel 8.

Abschnitt 6.1 gibt einen kurzen Überblick über das Vorgehen bei der Modellentwicklung und beschäftigt sich dann damit, die kausalen Zusammenhänge zwischen gesuchter und verfügbarer Information qualitativ zu erfassen. Abschnitt 6.2 vervollständigt die Entwicklung mit einer detaillierten Ausarbeitung der betrachteten Ereignisse hinsichtlich ihrer möglichen Ausprägungen. Abschnitt 6.3 faßt schließlich die Ergebnisse betreffend das entwickelte Modell zusammen.

6.1 Kausale Betrachtung der Domäne

Für die Entwicklung eines kausalen Modells kann kein allgemeines Schema angegeben werden, da viele Modellierungsentscheidungen nicht nur von der jeweiligen Domäne, sondern auch von der gegebenen Aufgabe innerhalb der Domäne und von den Ansprüchen des Entwicklers abhängen. Gleichwohl sind einige grundlegende Prinzipien zu nennen.

Im einzelnen ist es von besonderer Bedeutung festzustellen, welche Ereignisse im Kontext der konkreten Problemstellung interessant sind. Neben der gesuchten und der verfügbaren Information sind dies in der Regel mehrere weitere Ereignisse, da beobachtbare Symptome und gesuchte Hypothesen zumeist nicht direkt, sondern mittelbar in einem kausalen Zusammenhang stehen. Für alle betrachteten Ereignisse stellt sich die Frage, wie sie in diskrete Zufallsvariablen gefaßt werden können, und mit der Etablierung gerichteter Kausalverknüpfungen spielt auch die Zugänglichkeit der dementsprechend benötigten bedingten Wahrscheinlichkeiten eine Rolle. Der Anspruch, dem das zu entwickelnde Kausalmodell dabei im vorliegenden Fall gerecht werden soll, ist nicht nur seine grundsätzliche Eignung für die gestellte Aufgabe. Vor allem soll eine Grundlage geschaffen werden, von der ausgehend erst weitere Entwicklungen und vor allem Detailuntersuchungen stattfinden sollen (vgl. Kapitel 10), da das entstehende Gesamtsystem SASCHA in erster Linie ein Forschungssystem ist.

Für die Domäne der Strukturaufklärung durch Spektroskopie im Kontext der organischen Chemie, und darin für die gegebene Aufgabe der Substitutionsmustererkennung an Benzolderivaten durch Auswertung von ^{13}C-NMR-Spektren, lassen sich die folgenden für die Modellentwicklung relevanten Feststellungen treffen:

- Die gesuchte Information ist offensichtlich die Klasse von Kohlenstoffatomen innerhalb des Benzolrings. Sie soll in einer oder mehreren Hypothesenvariablen erfaßt werden, wobei die systematische Beschreibung von Molekülstrukturen eine Rolle spielt.

- Die zugängliche Information ist die Summenformel und die chemische Verschiebung der sechs Ringatome. Beides ist in diskrete Zufallsvariablen zu fassen.

- Es besteht ein Zusammenhang zwischen der Summenformel und der Struktur des Moleküls. Diese hat ihrerseits einen kausalen Einfluß auf die spektralen Befunde: Sie wird von der Struktur der Valenzelektronen charakterisiert, welche durch ihre abschirmende Wirkung auf den Atomkern auch die chemische Verschiebung prägt. Es wird ein geeignetes Konzept benötigt, um diesen Zusammenhang treffend wiederzugeben.

- Die Übertragung des genannten Konzepts in die kausale Modellierung wird höchstwahrscheinlich die Einbringung von Zwischenvariablen nötig machen.

- Neben der rein qualitativen Modellierung der kausalen Zusammenhänge ist es auch von Interesse, ob alle zu deren Quantifizierung benötigten bedingten Wahrscheinlichkeiten zugänglich sind. Gegebenenfalls muß eine alternative Modellierung gewählt werden.

Im folgenden werden als erstes die relevanten Ereignisse identifiziert und auf einer konzeptionellen Ebene entsprechende Variablen eingeführt. Deren Zustandsgestaltung erfolgt dann im Detail im folgenden Abschnitt 6.2, während zunächst die Zusammenhänge und Abhängigkeiten innerhalb der Domäne gegenüber den möglichen Ausprägungen der Ereignisse im Vordergrund stehen.

6.1.1 Gewünschte und verfügbare Information

In der gegenwärtigen Strukturaufklärungsaufgabe sind die sechs chemischen Verschiebungen der Kohlenstoffatome eines Benzolringes sowie die Summenformel der untersuchten Substanz gegeben. Gesucht ist im aktuellen Schritt der Verarbeitung die Klasse jeder einzelnen der sechs Ringpositionen.

Wie in Kapitel 2 dargestellt wurde, besteht ein kausaler Zusammenhang zwischen der Struktur eines organischen Moleküls und seinem zugehörigen ^{13}C-NMR-Spektrum, welcher sich genauer betrachtet bis auf die Ebene einzelner Kohlenstoffatome und ihrer zugehörigen Peaks bringen läßt: Die chemische Verschiebung eines Kohlenstoffatoms hängt von seiner Elektronenumgebung ab. Je höher die Elektronendichte, desto stärker wird der Atomkern von dem beim NMR-Experiment angelegten Magnetfeld abgeschirmt.

Die Elektronenverteilung wiederum hängt vom Typ des Kohlenstoffatoms sowie den Typen seiner Bindungspartner ab. Typen sollen dabei nach chemischem Element, Hybridisierung und, quasi rekursiv, nach den Typen der Nachbaratome unterschieden werden. Auf diese Weise wird berücksichtigt, daß Elektroneneinflüsse nicht lokal sind, sondern sich auch über mehrere Bindungen im Molekül hinweg auswirken.

Somit wird also als erstes eine Variable benötigt, welche den gesuchten Typ des betrachteten Kohlenstoffatoms repräsentiert. Sie wird in Anlehnung an die gegebene Vorannahme, daß es sich um ein sp^2-hybridisiertes Kohlenstoffatom in einem aromatischen System handelt, mit `C_arom` bezeichnet. Weitere Details zur Repräsentation von Atomtypen im Sinne der Molekülstruktur, im besonderen die Zustände dieser Variablen, bleiben auszuarbeiten; in diesem Zusammenhang sei insbesondere auf Abschnitt 6.1.2 verwiesen.

Fest steht jedoch bereits, daß es einen kausalen Zusammenhang zur Lage des korrespondierenden Peaks gibt. Auch die chemische Verschiebung soll in einer Variablen repräsentiert werden, welche mit `Peakpos` („Peakposition“) bezeichnet wird. Neben der Lage des Peaks wird auch die Summenformel der untersuchten Verbindung als gegeben angenommen. Sie enthält Informationen über die beteiligten chemischen Elemente und die Häufigkeit ihrer Vorkommen. Es bietet sich an, eine Variable für jedes chemische Element einzuführen, um das Vorhandensein des betreffenden Elements zu erfassen. Aus den bisherigen Überlegungen ergibt sich der in Abbildung 6.1 dargestellte initiale Entwurf der Kausalzusammenhänge.

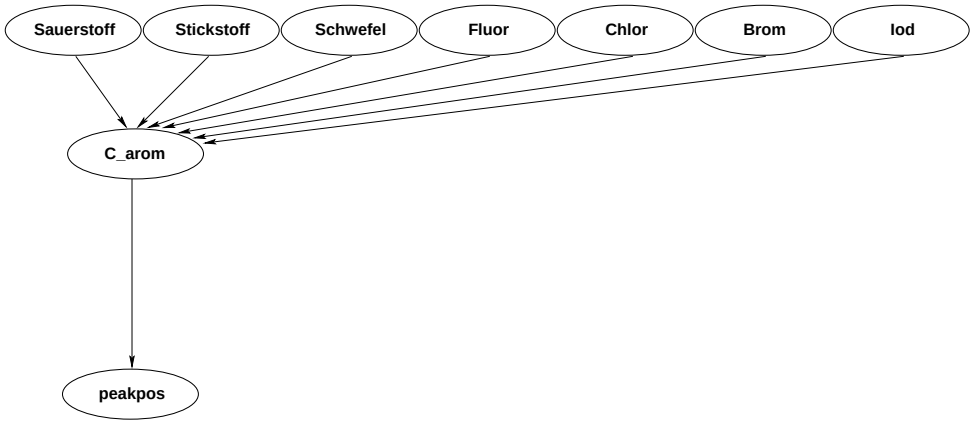


Abb. 6.1: *Erste grobe Repräsentation der kausalen Zusammenhänge für die gegebene Aufgabenstellung. Die Zustände der einzelnen Variablen sowie Details bezüglich des Zusammenhangs zwischen Struktur und Spektrum bleiben auszuarbeiten.*

Neben der Gestalt des Moleküls können auch externe Faktoren bei der Spektrenaufnahme die Positionen der Peaks beeinflussen. Ähnlich wie die in Abschnitt 2.1.2 (Seite 28) dargestellte Einbringung eines Chloratoms in einen reinen Kohlenwasserstoff intern wie ein Magnet auf die Elektronen des Moleküls wirkt, haben z.B. auch benachbarte Moleküle des Lösungsmittels einen Einfluß auf die Elektronenverteilung des untersuchten Moleküls. Weitere Faktoren sind etwa Wasserstoffbrücken, der pH-Wert, usw.. Da jedoch die spektralen Befunde möglichst nur die Elektronenverteilung aufgrund der Molekülstruktur widerspiegeln sollen, werden derartige Einflüsse, welche die gewünschte Information verschleiern würden, bei der Spektrenaufnahme so weit wie möglich vermieden. Aus diesem Grund werden sie auch an dieser Stelle im Modell nicht berücksichtigt, sondern es wird angenommen, daß bei der Spektrenaufnahme geeignete Bedingungen geschaffen und in jeder Hinsicht sorgfältig gearbeitet wurde, so daß die Varianz der einzelnen Wahrscheinlichkeitsverteilungen genügt, um etwaige Abweichungen vom Ideal infolge nicht idealer Aufnahmebedingungen zu kompensieren.

6.1.2 Systematische Beschreibung chemischer Strukturen

Hinsichtlich der Strukturrepräsentation scheint zunächst die Tatsache, daß das spätere Bayes-Netz zur Klassifikation der einzelnen Positionen des Benzolringes eingesetzt werden soll, dafür zu sprechen, daß es genügt, eine einzige die Struktur betreffende Hypothesenvariable (wie `C_arom` in Abbildung 6.1) einzubringen. In einer naiven Realisierung, welche einen Zustand

je möglichem Substituent vorsieht, wird jedoch rasch der Schwachpunkt dieses Ansatzes offenbar: Es ist praktisch unmöglich, eine erschöpfende Liste aller betrachteten Substituenten anzugeben, da jeder Substituent durch Einbau zusätzlicher Skelettatome prinzipiell beliebig erweitert werden kann. Die Aufzählung der Zustände jeder Variablen innerhalb eines Bayes-Netzes muß jedoch erschöpfend sein, da anderenfalls der Satz von der totalen Wahrscheinlichkeit keine Gültigkeit besitzt. Daher scheint es unvermeidlich, bei der Definition der Klassen Substituenten systematisch zusammenzufassen.

Bei der naiven Herangehensweise liefert die große Ähnlichkeit der durch einzelne Zustände repräsentierten Ereignisse (etwa Substituenten, die sich nur um ein zusätzliches Skelettatom unterscheiden) einen Ansatzpunkt für ein alternatives Modell: Betrachtet man zwei Substituenten, welche sich nur geringfügig unterscheiden, so wäre es vorteilhaft, ihre Repräsentation derart auf mehrere strukturbezogene Variablen aufzuteilen, daß die gemeinsame Substruktur durch denselben Zustand der einen und die unterschiedlichen Teile durch zwei verschiedene Zustände einer anderen Variablen repräsentiert werden. Ein Konzept hierzu legt der HOSE-Code nahe, der bereits in Abschnitt 3.1.5 als eine Möglichkeit der systematischen Beschreibung von Molekülstrukturen vorgestellt wurde. Ihm liegt die Idee zugrunde, Moleküle oder Substrukturen ausgehend von einem Fokuspunkt in Sphären untergliedert zu betrachten, wie Abbildung 6.2 noch einmal verdeutlicht.

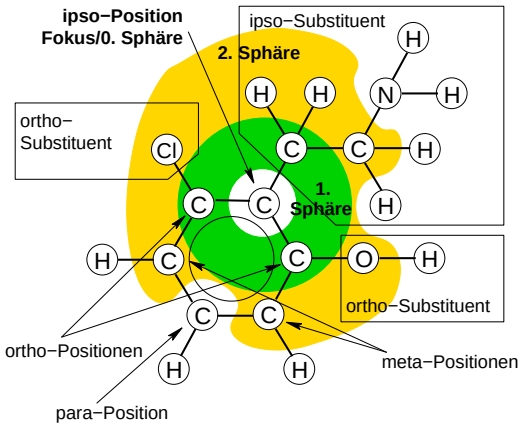


Abb. 6.2: *Der HOSE-Code beschreibt Molekülstrukturen untergliedert nach Sphären. An einem exemplarischen Benzolderivat sei hier außerdem noch einmal auf die Bezeichnung der einzelnen Positionen des Ringes verwiesen.*

Wie in Abbildung 6.2 außerdem deutlich wird, gehören bei der Betrachtung von aromatischen Kohlenstoffatomen in Benzolringen stets drei Atome zur ersten Sphäre, und zwar die beiden *ortho*-Ringatome sowie das erste Atom des an den *ipso*-Kohlenstoff gebundenen Substituenten. Die Ringatome sind jedoch invariant gegeben die Betrachtung von Benzolderivaten, und sie müssen somit nicht explizit modelliert werden. Das verbleibende Atom ermöglicht seinerseits eine grobe strukturelle Kategorisierung des Substituenten in der *ipso*-Position. Eine diesem Ansatz folgende graduelle Verfeinerung der Unterscheidung von Substituentenklassen ist in Abschnitt 4.1.3 bereits vorgestellt worden und wird in Abbildung 6.5 (Seite 89) noch einmal rekapituliert. Auch mit Blick auf den Zusammenhang zwischen Spektrum und Struktur erscheint diese Gliederung sinnvoll, da der Einfluß ansonsten gleichartiger Atome auf die chemische Verschiebung des *ipso*-Kohlenstoffs in der Regel abnimmt, je weiter das Atom entfernt ist, das heißt je höher die Sphäre, zu welcher es gehört.

Dem entsprechend wird nun die Modellierung angepaßt. Als erstes dient die Variable `C_arom` nicht mehr der Beschreibung des gesamten Substituenten, sondern nur noch seiner Klassifikation nach der Gestalt der ersten Sphäre. Neben dem direkten Bindungspartner sollen jedoch auch die Einflüsse entfernterer Atome auf die Elektronendichte in der Umgebung des *ipso*-Kohlenstoffs modelliert werden. Betrachtet man die zweite Sphäre, so trifft man in den beiden *meta*-Positionen wiederum auf zwei Ringatome, welche invariant sind, sowie auf die beiden ersten Atome der *ortho*-Substituenten, welche im folgenden kurz als *s2ortho*-Atome[1] bezeichnet werden. Außerdem gehören der oder die Bindungspartner des ersten Atoms des *ipso*-Substituenten zur zweiten Sphäre und sollen kurz als *s2ipso*-Atome[2] bezeichnet werden.

Bei der Repräsentation in Variablen werden die invarianten *meta*-Ringatome wiederum nicht explizit wiedergegeben. Betreffend die verbleibenden Atome der 2. Sphäre ist es sinnvoll, diese in zwei getrennten Variablen für die *s2ipso*- und *s2ortho*-Atome zu erfassen: Zwischen ersteren und den durch die Variable `C_arom` erfaßten *ipso*-Atomen der ersten Sphäre besteht eine Abhängigkeit, da sie an diese gebunden sind, während die *s2ortho*-Atome von beiden unabhängig sind. Im besonderen ist die Zahl der *s2ipso*-Atome von der Valenz des Atoms in der ersten Sphäre abhängig.

Abhängigkeiten bestehen darüber hinaus zwischen allen die Struktur betreffenden Variablen und der Summenformelinformation. Ist etwa die Abwesenheit eines bestimmten chemischen Elements bekannt, so können alle Ausprägungen von Ereignissen, welche ein Vorkommen des betreffenden Elements beinhalten, automatisch ausgeschlossen werden. Abbildung 6.3 zeigt das bis hierher erarbeitete Modell.

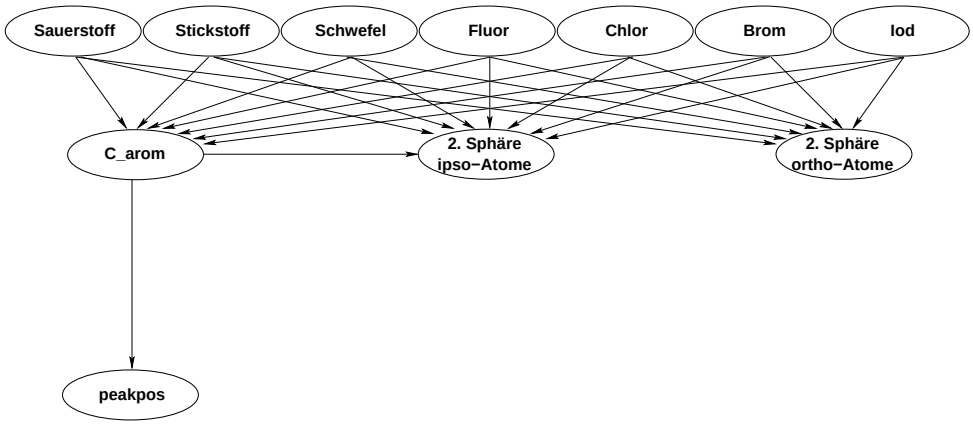


Abb. 6.3: *Kausalmodell der Domäne für die gegebene Aufgabenstellung. Die Molekülstruktur in der näheren Umgebung des Fokusatoms wird nun nach Sphären untergliedert betrachtet und in mehreren Variablen erfaßt.*

Der Deutlichkeit halber sei an dieser Stelle darauf hingewiesen, daß nicht nur `C_arom`, sondern auch jede andere der strukturbezogenen Variablen die Rolle der Hypothesenvariablen innehaben kann. Dabei können die übrigen Variablen jeweils als Informationsvariablen dienen: Wie in Abschnitt 5.1.4 angedeutet, sind Klassifikationsresultate in Gestalt dreigliedriger Segmente eines Benzolringes angestrebt. Diese beinhalten neben der (*ipso*-)Klasse eines

[1] *ortho*-Atome der *2. S*phäre
[2] *ipso*-Atome der *2. S*phäre

bestimmten Substituenten auch die wahrscheinlichste Kombination seiner Nachbargruppen. Diese *ortho*-Klassen werden gegeben die bereits gefundene wahrscheinlichste *ipso*-Klasse ermittelt, `C_arom` dient hier also als Informationsvariable. Im Zusammenhang einer etwaigen Revision von Klassifikationsergebnissen (vgl. Kapitel 9) ist auch die umgekehrte Rollenverteilung denkbar.

Ebenso wie bei der Untergliederung in Atome der ersten und zweiten Sphäre ist natürlich eine weitere Verfeinerung der Unterscheidung von Substituentenklassen möglich, indem die Atome bis zur dritten, vierten oder fünften Sphäre betrachtet werden, was auch eine explizite Modellierung der Einflüsse der Substituenten in den *meta*-Positionen und der *para*-Position erlauben würde. Obwohl diese Gruppen über ihre Verbindung mit dem aromatischen System einen nicht unerheblichen Einfluß auf die betrachtete chemische Verschiebung ausüben, soll jedoch vorerst die Modellierung bei einer Genauigkeit von zwei Sphären enden.

Das Gesamtsystem, dessen Teil das hier entwickelte Kausalmodell ist, ist zum gegenwärtigen Zeitpunkt in erster Linie ein Forschungssystem. Es gibt bislang noch keine Erfahrungen darüber, wie sich einzelne Modellierungsentscheidungen auswirken oder wie gut der verfolgte Modellierungsansatz überhaupt für die Klassifikation eines Substituenten geeignet ist. An dem System sollen Untersuchungen durchgeführt werden, die zu gerade solchen Erkenntnissen führen, und diese können dann wiederum genutzt werden, um das Kausalmodell weiterzuentwickeln, zu präzisieren und nötigenfalls Veränderungen in der Repräsentation einzelner Aspekte vorzunehmen.

Aus diesem Grund erscheint es vom Standpunkt der Modellentwicklung her ratsam, das Kausalmodell zunächst möglichst einfach zu halten. Jegliche Verfeinerungen (etwa betreffend die Einflüsse der *meta*- und *para*-Substituenten, aber auch andere Faktoren, vgl. Kapitel 10), die später daran vorzunehmen sind, können nur davon profitieren, wenn zuvor Erkenntnisse über vorteilhafte oder weniger günstige Modellierungsentscheidungen gesammelt wurden: Diese Erfahrungen sind für ein zielgerichtetes Vorgehen wertvoll und werden somit den Aufwand zukünftiger Weiterentwicklungen reduzieren. Gleichwohl muß die Tatsache Eingang in die an das System zu richtenden Erwartungen finden, daß auch weiter als zwei Sphären entfernte Atome einen erheblichen Einfluß auf die chemische Verschiebung des untersuchten Ringatoms haben und daß das als Ausgangspunkt weiterer Entwicklungen dienende Basismodell somit viel Raum für Verbesserungen läßt.

Die genaue Festlegung der Zustände aller Variablen bleibt zu diesem Zeitpunkt noch offen; sie ist Gegenstand von Abschnitt 6.2. Zur Fertigstellung des Kausalmodells sind zuvor noch einige weitere Schritte nötig. Insbesondere gilt es, die Unterteilung der Strukturbeschreibung in Sphären auch auf den Zusammenhang zur chemischen Verschiebung des betrachteten Kohlenstoffatoms zu übertragen. Dem widmet sich der folgende Abschnitt.

6.1.3 Der Inkrement-Ansatz

Neben der günstigen Realisierung in mehreren Variablen ist die in Sphären untergliederte Strukturrepräsentation in Anlehnung an den HOSE-Code, wie bereits erwähnt, aus einem weiteren Grund sinnvoll: Die Stärke des Einflusses auf die Elektronenstruktur hängt in der Regel neben chemischem Element und Hybridisierung auch von der Entfernung des betreffenden Atoms vom untersuchten Kern ab, das heißt je höher die Sphäre, in welcher ein Atom zu finden ist, desto geringer sein Einfluß. Eine ebenfalls an den Sphären orientierte Aufschlüsselung der chemischen Verschiebung ist daher der nächste durchzuführende Modellierungsschritt.

In der Literatur sind an verschiedener Stelle ausgewertete ^{13}C-NMR-Spektren veröffentlicht [Ewi79, Die92, Str89]. Hervorzuheben ist hier Ewings Arbeit [Ewi79], in welcher er monosubstituierte Benzolderivate hinsichtlich der auf die einzelnen Kohlenstoffatome wirkenden Substituenteneffekte untersucht. Ausgehend von der bekannten chemischen Verschiebung von 128,5 ppm, die die Ringatome von Benzol aufweisen, wird darin untersucht, inwiefern die sechs chemischen Verschiebungen des aromatischen Ringes bei der Anwesenheit unterschiedlichster Substituenten in einer Position des Ringes von diesem Wert abweichen. Dadurch ergibt sich eine ausführlich tabellierte Liste von *Substituenteninkrement*en, welche wiedergibt, auf welche Weise ein bestimmter Substituent die chemische Verschiebung seines Bindungspartners wie auch der übrigen Ringatome beeinflußt.

Derartige Informationen können genutzt werden, um Inkrementtabellen für die Vorhersage von ^{13}C-NMR-Spektren zu erstellen. Die Grundidee der Inkrementmethode zur Vorhersage von Spektren wurde in Abschnitt 3.1.2 bereits vorgestellt. Neben den dort angeführten Arbeiten, welche sich stark auf den Einsatz innerhalb eines größeren Strukturaufklärungssystems beziehen, läßt sich aus [Die92] und [STK94] ein Einblick in die Motivation sowie mögliche Schwierigkeiten bei der Realisierung gewinnen.

Wie in Abschnitt 3.1.2 beschrieben kann die chemische Verschiebung s eines Kohlenstoffatoms berechnet werden als die Summe seiner Grundverschiebung σ_0, der Summe über alle Substituenteninkremente σ_i und einem Korrekturterm K für sterische Effekte:

$$s = \sigma_0 + \sum_i \sigma_i + K \tag{6.1}$$

Für das gegenwärtige Modell wird Gleichung (6.1) in zweierlei Hinsicht modifiziert: Zum einen werden die Substituenteninkremente σ_i durch Inkremente s_i bezogen auf die Einflüsse der Atome in den durch i indizierten Sphären ersetzt. Zum anderen wird auf einen expliziten Korrekturterm verzichtet. Implizit bleibt seine Funktion jedoch im späteren Bayes-Netz durch die bedingten Wahrscheinlichkeiten erhalten.

$$s = s_0 + \sum_i s_i \tag{6.2}$$

Für das Modell bedeutet dies die Einführung zweier Variablen `Sphäre1` und `Sphäre2` für die Teilinkremente der ersten und zweiten Sphäre. Die bereits bestehende Variable `Peakpos` für die Position des betrachteten Peaks übernimmt die Summenbildung gemäß Gleichung (6.2). Dabei muß sie die Möglichkeit berücksichtigen, daß Abweichungen vom gemäß dem Modell zu erwartenden Wert beobachtet werden können, die ursprünglich durch den expliziten Korrekturterm erfaßt wurden. Mit einer gewissen Wahrscheinlichkeit kann also ein leicht höherer oder niedrigerer Wert als die exakte Summe der Inkremente der ersten und zweiten Sphäre beobachtet werden. Dies wird durch eine „weiche Summenbildung“ über die entsprechende bedingte Wahrscheinlichkeit realisiert; genaueres ist Abschnitt 8.3.3 zu entnehmen. Eine zusätzliche Abweichung wird darüber hinaus dadurch entstehen, daß nur bis zu den Atomen der zweiten Sphäre die zugehörigen Inkremente repräsentiert sind und inbesondere die Einflüsse der Substituenten in den *meta*- und *para*-Positionen nicht explizit modelliert werden. Überlegungen zu diesem Zusammenhang finden sich im Zusammenhang der Evaluation in Abschnitt 10.3, während das möglichst einfach strukturiert zu haltende Basismodell dies zunächst in Kauf nimmt.

In Abschnitt 3.1.2 wurde bezüglich des Inkrementansatzes jedoch erwähnt, daß eine Ungenauigkeit des Verfahrens im Sinne von Gleichung (6.1) darin besteht, daß eine Unabhängigkeit der einzelnen Substituenten betreffend des Zusammenwirkens ihrer Inkremente vorausgesetzt wird. Eine Weiterentwicklung ist daher die Verwendung von Kreuztermen, über

welche für alle Substituenten*kombinationen* ebenfalls summiert wird, jedoch ist es gerade beim Zusammenwirken von weniger häufigen oder mehreren Substituenten unter Umständen schwierig, diese Kreuzterme verläßlich zu bestimmen.

Durch konsequente Orientierung an der Strukturmodellierung können die Inkrementvariablen des Kausalmodells ähnliches wie die Kreuzterme leisten. Da die zweite Sphäre nach *ipso*- und *ortho*-Atomen unterschieden wurde, wird auch das zugehörige Inkrement in zwei Teilinkremente aufgeteilt: Die Variable `i2ipso` gibt den *ipso*-Anteil des Inkrements der zweiten Sphäre wieder, `i2ortho` analog den *ortho*-Anteil. Die bedingte Wahrscheinlichkeit $P(\texttt{Sphäre2}|\texttt{i2ipso},\texttt{i2ortho})$ steht dann prinzipiell in einem eben solchen Verhältnis zu den einzelnen Wahrscheinlichkeiten $P(\texttt{Sphäre2}|\texttt{i2ipso})$ und $P(\texttt{Sphäre2}|\texttt{i2ortho})$ wie Kreuzterme zu den Einzelinkrementen. Analoges gilt für $P(\texttt{peakpos}|\texttt{Sphäre1},\texttt{Sphäre2})$ hinsichtlich des Zusammenwirkens der ersten und zweiten Sphäre. Abbildung 6.4 zeigt nun das um Variablen für die Inkremente und Teilinkremente ergänzte Modell.

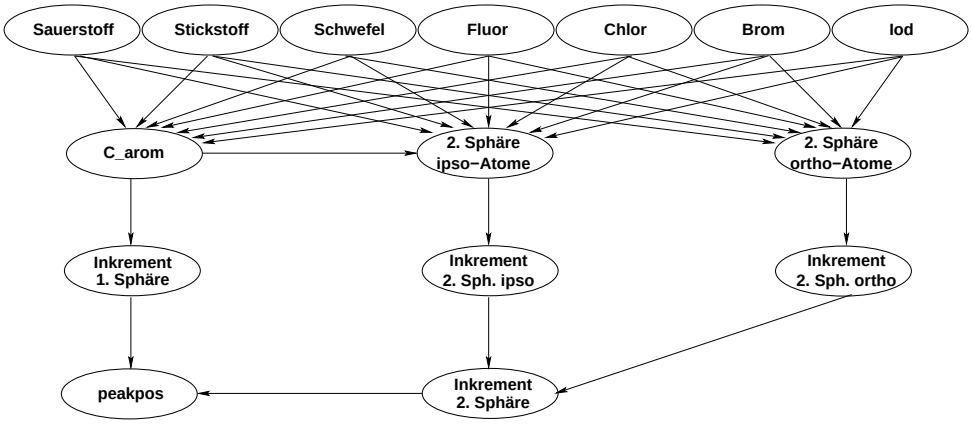


Abb. 6.4: *Das Kausalmodell erfaßt nun sowohl die Struktur als auch ihre Einflüsse auf die chemische Verschiebung des untersuchten Kohlenstoffatoms nach Sphären untergliedert. Aufmerksamkeit bedarf nun noch die Gestaltung der Zustände der einzelnen Variablen.*

Es stellt sich jedoch die Frage, ob die Information zur Bestimmung von auf die einzelnen Sphären bezogenen Inkremente sowie von *ipso*- und *ortho*-Anteilen der zweiten Sphäre zugänglich ist, da Inkrementtabellen [wul04, ucl02] typischerweise Terme für komplette Substituenten bzw. Substituentenarten enthalten. Im folgenden wird ein Verfahren vorgestellt, welches die Gewinnung der gewünschten Information beschreibt. Es läßt sich prinzipiell sowohl auf ausgewertete NMR-Spektren wie in [Ewi79] als auch auf Tabellen von Substituenteninkrementen anwenden, doch scheint das letztere nicht ratsam, da diese ebenfalls aus experimentellen Daten gewonnen wurden, wobei jedoch andere Voraussetzungen betreffend die weitere Verwendung des Resultats zugrundelagen.

Auch in der Literatur findet man jedoch nur Angaben der gesamten chemischen Verschiebung und nicht der Teilinkremente. Um zu veranschaulichen, wie man dennoch die gewünschten Teilinkremente erhält, sei an dieser Stelle in Abbildung (Abbildung 6.5) noch einmal der Baum von Substituentenklassen aus Abschnitt 4.1.3 dargestellt. Darin sind abhängig von der ersten Sphäre einatomige und mehratomige Substituenten zu unterscheiden.

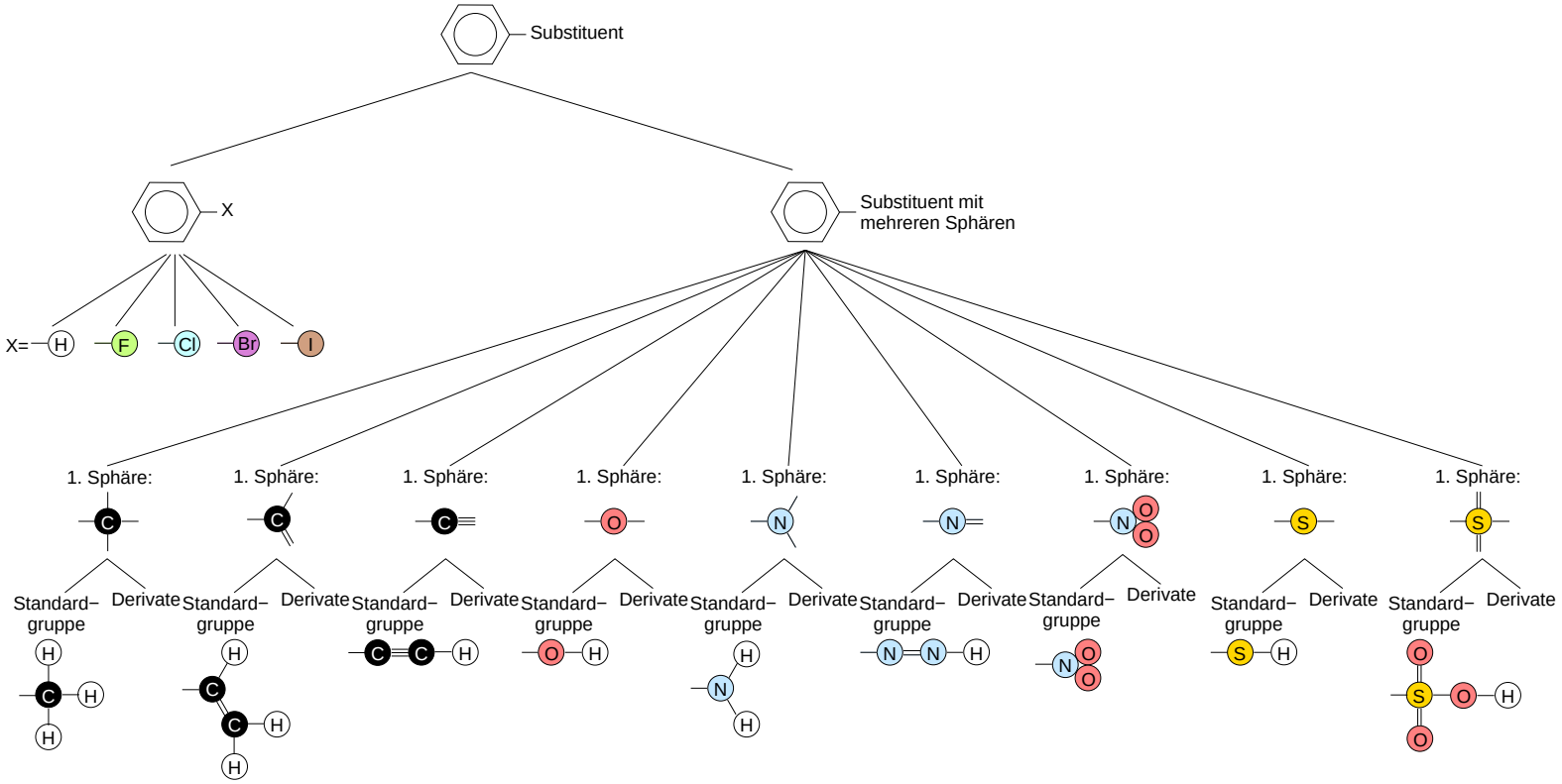


Abb. 6.5: *Klassifikationsbaum zur Unterscheidung von Substituenten an Benzolringen nach strukturellen Gesichtspunkten. Um ein Benzolderivat einem bestimmten Substitutionsmuster zuzuordnen, müssen jedoch alle sechs Positionen des Ringes und ihre relative Stellung zueinander betrachtet werden.*

Die Grundverschiebung s_0 für Kohlenstoffatome in einem Benzolring beträgt 128.5 ppm. Betrachtet man nun monosubstituierte Benzolderivate, so kann für einatomige Substituenten das Inkrement s_1 der ersten Sphäre direkt der Literatur entnommen werden. Sei s die chemischen Verschiebung des *ipso*-Kohlenstoffs mit einem einatomigen Substituenten, so ist dessen Einfluß s_1 gemäß Gleichung (6.3) zu bestimmen:

$$s_1 = s - 128.5 \tag{6.3}$$

Dabei sollte möglichst nicht nur eine Quelle herangezogen werden. Bei unterschiedlichen Angaben in den verschiedenen Quellen kann z.B. der Durchschnitt gebildet, der am häufigsten auftretende Wert oder der Wert der „vertrauenswürdigsten“ Quelle gewählt werden. Vorgehen und mögliche Alternativen sind für die spätere Evaluation zu dokumentieren.

Für die mehratomigen Substituenten im rechten Teil des Baumes wird betreffend der zweiten Sphäre unterschieden, ob es sich um eine „Standardgruppe“ oder ein Derivat derselben handelt. Standardgruppen besitzen in der zweiten Sphäre ausschließlich Wasserstoffatome, oder im Fall von Mehrfachbindungen ein Atom desselben Typs wie das der ersten Sphäre. In diesem Fall muß die dritte Sphäre ausschließlich H-Atome enthalten. Eine Ausnahme bilden die Gruppen $–NO_2$ und $–SO_3H$, um ihre typische Charakteristik zu erhalten. Standardgruppen sind somit

- $–CH_3$
- $–CH{=}CH_2$
- $–C{\equiv}CH$
- –OH
- –SH
- $–NH_2$
- –N=NH
- $–NO_2$
- $–SO_3H$

Unter der Annahme, daß Wasserstoffatome (oder bei Mehrfachbindungen Atome des gleichen Typs wie ihr Bindungspartner) grundsätzlich keinen Einfluß auf die chemische Verschiebung haben, können die Standardgruppen genauso behandelt werden wie einatomige Substituenten, so daß das Inkrement der ersten Sphäre ebenfalls nach Gleichung (6.3) berechnet werden kann.

Betrachtet man nun monosubstituierte Benzolderivate mit anderen als den Standardgruppen, so kann das *ipso*-Teilinkrement der zweiten Sphäre gemäß Gleichung (6.4) berechnet werden, sofern das Inkrement s_1 der ersten Sphäre durch Betrachtung einer Verbindung mit dem entsprechenden Standardsubstituenten bereits bestimmt wurde.

$$s_{2\text{ipso}} = s - 128.5 - s_1 \tag{6.4}$$

Die chemische Verschiebung s bezieht sich wiederum auf den *ipso*-Kohlenstoff, und der Substituent darf sich nur in der zweiten Sphäre von der Standardgruppe unterscheiden.

Zur Bestimmung des *ortho*-Anteils der zweiten Sphäre werden wiederum monosubstituierte Benzolderivate betrachtet, dieses Mal jedoch nicht die chemische Verschiebung des *ipso*-, sondern des *ortho*-Kohlenstoffs. Relevant sind dieses Mal nur Spektren von Verbindungen mit einatomigen Substituenten und mit den Standardgruppen für mehratomige Substituenten. Für jeden davon wird der Wert des *ortho*-Teilinkrements der zweiten Sphäre analog zu Gleichung (6.3) gebildet. Durch paarweise Summation erhält man die Gesamtmenge in Frage kommender Werte für das Teilinkrement $s_{2\text{ortho}}$, da für jedes Kohlenstoffatom des Ringes zwei *ortho*-Positionen berücksichtigt werden müssen.

Auf den ersten Blick scheint dieses Vorgehen durch die implizite Annahme der Unabhängigkeit der Einflüsse von einerseits zwei *ortho*-Substituenten in ihrem Zusammenwirken und andererseits dem Zusammenwirken von *ipso*- und *ortho*-Anteilen der zweiten Sphäre der angestrebten Parallele zur Einbringung von Kreuztermen in den Inkrementansatz zu widersprechen. Es sei daher noch einmal darauf hingewiesen, daß die Berücksichtigung des

Zusammenwirkens der Einflüsse über die bedingten Wahrscheinlichkeiten geschehen soll. Ihnen muß diesbezüglich gesondert Aufmerksamkeit gewidmet werden. Dieser Aspekt der Parametrisierung bzw. der Quantifizierung der zu diesem Zeitpunkt noch rein qualitativen Wissensrepräsentation ist Gegenstand von Abschnitt 8.3.1.

Mit den bis hierher angestellten Überlegungen erscheint nunmehr das Kausalmodell bezüglich der darin betrachteten Ereignisse vollständig. Was jedoch nach wie vor zu entwickeln bleibt sind die Zustände der einzelnen Variablen. Ihnen widmet sich der folgende Abschnitt.

6.2 Gestaltung der Zustände

Ausgehend von den Grundkonzepten der einzelnen Variablen, das heißt von den durch sie modellierten Ereignissen, ist zu überlegen, welche und auch wie viele Zustände diese erhalten sollen. Die Zustände geben die möglichen Ausprägungen wieder, in welchen das durch die Variable modellierte Ereignis eintreten kann, und müssen daher paarweise disjunkt sein und in ihrer Gesamtheit den gesamten Raum möglicher Ausprägungen ausschöpfen. Im folgenden soll das Vorgehen bei der abschließenden Definition der Variablen einschließlich ihrer Zustände dargestellt werden.

6.2.1 Strukturbezogene Variablen

Variablen, welche die Molekülstruktur betreffen, sind offensichtlich `C_arom`, `s2ipso` und `s2ortho` sowie im weiteren Sinne auch die Variablen zur Wiedergabe der Summenformelinformation. Mit den letzteren soll die Realisierung nun begonnen werden. Dabei ist zu beachten, daß die Zustände einer Variablen stets disjunkt und erschöpfend sein müssen, das bedeutet, alle möglichen Ausprägungen des durch die Variable modellierten Ereignisses müssen abgedeckt sein, und es darf keine zwei Zustände geben, welche sich ganz oder teilweise auf dieselbe Ausprägung beziehen.

Am Beispiel von Sauerstoff soll die grundsätzliche Realisierungsidee verdeutlicht werden: Es wird eine Variable `Sauerstoff` definiert, die das Ereignis des Vorhandenseins von Sauerstoff im untersuchten Molekül modelliert. Sie erhält zunächst nur die beiden Zustände `ja` (Sauerstoff ist im Molekül vorhanden) und `nein` (Sauerstoff ist nicht im Molekül vorhanden). Eine dritte mögliche Ausprägung desselben Ereignisses gibt es nicht. Es ist jedoch möglich, den Zustand `ja` durch mehrere Zustände zu ersetzen, um auch die Häufigkeit des Vorkommens zu erfassen. Diese beiden Varianten sind in Abbildung 6.6 dargestellt.

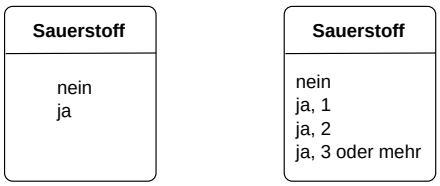


Abb. 6.6: *Mögliche Repräsentationen des Ereignisses „Vorhandensein von Sauerstoff“*

Welche Realisierung am günstigsten ist, kann im Grunde für jedes der betrachteten chemischen Elemente einzeln entschieden werden, es ist jedoch zu beachten, daß zwischen allen Element-Variablen und allen anderen strukturbezogenen Variablen Abhängigkeiten bestehen, so daß selbst im Falle binärer Element-Variablen bei Berücksichtigung von sieben Arten von Heteroatomen bereits $2^7 = 128$ Verteilungen für `C_arom` und `s2ortho` und für `s2ipso` sogar

$128 \cdot 14 = 1792$ Verteilungen benötigt werden. Um dieselben zu bestimmen, muß in entsprechendem Umfang empirisches oder Expertenwissen vorliegen.

Dies ist nicht gegeben, wenn empirische Daten herangezogen werden sollen, da keine beliebig große Stichprobe zur Verfügung steht. Die Verwendung von Expertenwissen stellt ebenfalls keine Alternative dar, da die benötigten Wahrscheinlichkeiten der in einer bestimmten Position jeweils betrachteten Gruppen unter der Voraussetzung, daß eine bestimmte Kombination von Heteroatomen in der Summenformel vorliegt, nicht ohne weiteres exakt zu quantifizieren sind. Insofern sind mit Blick auf die absehbare Verwendung empirischen Wissens Variablen mit möglichst wenigen Zuständen anzustreben.

Dem entgegen könnte man für die Notwendigkeit einer genaueren Unterscheidung argumentieren, daß in ähnlicher Weise, wie die Abwesenheit eines chemischen Elements bestimmte Substituenten ausschließt, auch die Zahl der vorhandenen Atome darüber bestimmt, ob ein gleichzeitiges Auftreten von Substituenten, die das betreffende Element enthalten, in mehreren der im Modell erfaßten Positionen möglich ist. Diese Information der generellen Möglichkeit oder Unmöglichkeit läßt sich unabhängig von der Verwendung von empirischem oder Expertenwissen für die bedingten Wahrscheinlichkeiten beantworten.

Genau betrachtet bedeutet dies, daß es einerseits gültige und andererseits ungültige Konfigurationen gegeben die Zahl der Atome eines bestimmten chemischen Elements gäbe, das heißt es bestünde eine ungerichtete Abhängigkeit zwischen den Strukturvariablen. Damit wäre die Einführung zusätzlicher Zwischenvariablen nach dem in Abschnitt 4.3.2 zur Auflösung ungerichteter Abhängigkeiten vorgestellten Prinzip erforderlich. Um aber das Modell in seiner Eigenschaft als Basis für weitere Entwicklungen zunächst möglichst einfach zu halten, und um zudem im Falle einer späteren Überarbeitung (vgl. Kapitel 10) den tatsächlichen Einfluß der zusätzlichen Information (in diesem Fall die Anzahl von Atomen jedes betrachteten chemischen Elements) auf die Leistung des Systems beurteilen zu können wird die Realisierung an dieser Stelle auf binäre Summenformel-Variablen festgelegt.

Hinsichtlich der übrigen strukturbezogenen Variablen scheint es am sinnvollsten, mit der Betrachtung von `C_arom` zu beginnen. Die Variable repräsentiert die Besetzung der ersten Sphäre. Da die beiden *ortho*-Kohlenstoffatome des Benzolringes invariant sind, bleibt nur zu unterscheiden, was für ein Atom die Bindung des *ipso*-Substituenten zum Benzolring herstellt. Die Zustände von `C_arom` sollen den möglichen Atomen in dieser Position unterschieden nach chemischem Element und Hybridisierung entsprechen. Bei den betrachteten chemischen Elementen O, N, S, F, Cl, Br und I neben Kohlenstoff und Wasserstoff ergibt sich so die in Abbildung 6.7 dargestellte Realisierung. Sie entspricht der zweiten Ebene innerhalb des Klassenbaums in Abbildung 6.5 auf Seite 89.

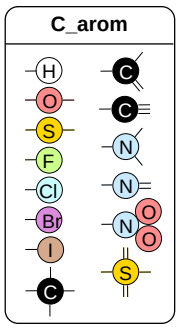


Abb. 6.7: *Gestaltung der Hypothesenvariable* `C_arom`*. Die Zustände geben die möglichen Ausprägungen der ersten Sphäre wieder.*

Da die festgelegten Zustände die unterschiedlichen Typen von Atomen beschreiben, welche grundsätzlich für das Modell interessant sind, können sie darüber hinaus als Basis für die Entwicklung der Zustände der verbleibenden Strukturvariablen dienen. Ein guter Ansatz für diese ist es, zunächst den durch sie jeweils zu beschreibenden Ereignisraum und seine Elementarereignisse zu betrachten und für jedes Elementarereignis einen Zustand anzunehmen. Da die Zustände von `s2ortho` und `s2ipso` den in den betreffenden Positionen möglichen Kombinationen von Atomen entsprechen sollen, gibt es jedoch aus Gründen der Kombinatorik jeweils sehr viele Elementarereignisse. Für die Verteilung von n verschiedenen Atomtypen auf k Positionen gilt die allgemeine Formel

$$x = \binom{n+k-1}{k} \tag{6.5}$$

für die Bestimmung der Anzahl x der unterschiedlichen Kombinationen. Wiederholungen desselben Atomtyps sind dabei zulässig, außerdem wird die Reihenfolge der Nennung nicht berücksichtigt, so daß die einzelnen Bindungen eines Atoms und somit auch etwaige Konfigurationsisomere nicht unterscheidbar sind. Der HOSE-Code würde eine derartige Unterscheidung zwar ermöglichen, jedoch wird bei der gegenwärtigen Aufgabenstellung die Molekülstruktur nur bis zur Ebene der Konstitution betrachtet, so daß eine mengenartige Betrachtung der einzelnen Sphären ausreichend ist.

Betrachtet man dieselben $n = 14$ Typen von Atomen wie bei der Festlegung der Zustände für `C_arom`, so ergeben sich für die $k = 2$ in `s2ortho` betrachteten Positionen $x = 210$ unterschiedliche Besetzungen. Für `s2ipso` ist die Vielfalt noch größer: Abhängig von der Belegung der ersten Sphäre kann die zweite Sphäre kein, ein, zwei, drei oder (im Falle einer Valenzschalenerweiterung des Schwefels) theoretisch sogar vier oder fünf Atome enthalten. Es sind also die möglichen Kombinationen von $n = 15$ (neben den in `C_arom` betrachteten 14 Typen auch das Leerbleiben einer Position) Typen in $k = 5$ Positionen gesucht – das bedeutet, nach Gleichung (6.5) wären $x = 11628$ unterschiedliche Besetzungen denkbar.

Wenngleich es unbestritten generell möglich ist, mit einer solch hohen Zahl unterschiedlicher Ausprägungen eines Ereignisses umzugehen, stellen zwei Aspekte jedoch in Frage, ob es tatsächlich sinnvoll ist, diese feinstmögliche Granularität der Strukturbeschreibung zu wählen: Der erste ist die Frage der Relevanz der Unterscheidung im Kontext der Anwendungsdomäne, und der zweite ist die Zugänglichkeit entsprechenden Wissens.

Bezüglich des ersteren Punktes ist es aus dem Blickwinkel der NMR-Spektroskopie wichtiger zu unterscheiden, wie sich die einzelnen Gruppen hinsichtlich ihrer Eigenschaften im NMR verhalten, als jede einzelne mögliche Variante gegeneinander abzugrenzen. Wenngleich diese Abgrenzung ebenfalls als eine Strukturaufklärungsaufgabe formuliert werden kann, so befaßt sich die gegenwärtige Arbeit doch mit einer anderen Problematik, der Substitutionsmustererkennung, und befindet sich hier an einem Punkt, wo nicht strukturelle Aspekte, sondern Elektroneneinflüsse auf ein bestimmtes, zu klassifizierendes Atom des Benzolringes im Zentrum der Betrachtung stehen.

Die Zugänglichkeit von Wissen spielt ebenfalls eine Rolle, und zwar beim Übergang vom Strukturmerkmal zu den Inkrementen, die jeweils zu den einzelnen Ausprägungen korrespondieren. Ist für ein Strukturmerkmal (ein Elementarereignis) kein zugehöriges Inkrement bekannt, von welchem verläßlich auf das Strukturmerkmal zurückgeschlossen werden könnte, so ist es nicht sinnvoll, gerade dieses Elementarereignis gesondert zu repräsentieren. Es kann vielmehr mit anderen Ereignissen zusammengefaßt werden, die Strukturen repräsentieren, die ihm insbesondere von ihrem Einfluß auf die Elektronenstruktur her ähneln.

Als positiver Nebeneffekt einer geringeren Zahl von Zuständen ist außerdem eine drastische Beschleunigung des Klassifikationsvorganges zu erwarten. Zum einen bestimmt die Zahl der Zustände einer Variablen darüber, über wie viele Zustände im Zuge einer Klassifikationsanfrage an das spätere Bayes-Netz zu summieren oder zu maximieren ist (vgl. [Dec96] bzw. Abschnitt 5.2.2), zum anderen geht sie als Faktor in die Zahl der benötigten Verteilungen für ihre Kinder ein. Über die positive Auswirkung auf die Rechenzeit hinaus sinkt zudem der Speicherplatzbedarf des Bayes-Netzes erheblich.

Abbildung 6.8 zeigt die Gestaltung, die nach diesen Vorüberlegungen für `s2ipso` und `s2ortho` gewählt wurde. Sie ergibt sich aus dem im folgenden beschriebenen Verfahren.

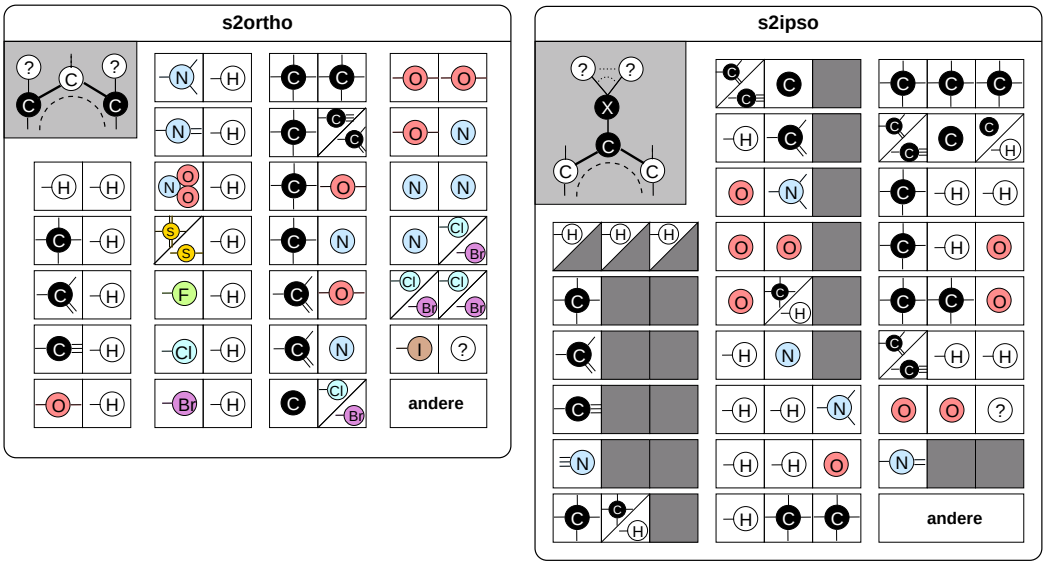


Abb. 6.8: *Gestaltung der strukturbezogenen Variablen* `s2ortho` *und* `s2ipso`*. Geteilte Felder bedeuten, daß in der betreffenden Position zwischen den angegebenen Atomtypen kein Unterschied gemacht wird. Ist nur ein Elementsymbol ohne Bindungen angegeben, wird nicht nach Hybridisierung unterschieden. Graue Felder symbolisieren, daß sich kein Atom in der betreffenden Position befindet.*

Für `s2ortho` wurde nun zunächst festgestellt, welche der theoretisch möglichen 210 Substituentenkombinationen in der Praxis tatsächlich vorkommen, da insbesondere Substituenten, die bereits für sich genommen selten auftreten (etwa Fluor, Iod oder Sulfonate) in den einzelnen möglichen Kombinationen um so seltener zu erwarten sind. Dazu wurde die Häufigkeit aller möglichen Kombinationen auf einer freundlicherweise von der BASF AG zur Verfügung gestellten Stichprobe untersucht (vgl. Kapitel 8). Substituentenkombinationen, die sich strukturell ähneln, wurden gemeinsam auf einen Zustand abgebildet, sofern die strukturelle Verallgemeinerung eine ausreichende Häufigkeit besitzt. Dies führte zu insgesamt 21 Klassen von Kombinationen. Alle übrigen werden in einem Zustand `andere` subsumiert.

Im Fall der Variablen `s2ipso` wurde vor der Untersuchung der Stichprobe zuerst festgestellt, für welche der theoretisch möglichen Belegungen Expertenwissen zur Ermittlung der zugehörigen Teilinkremente zugänglich ist. Bei allen übrigen theoretisch möglichen Strukturen wird angenommen, daß sie ohnehin zu selten auftreten, um sie jeweils in einem eigenen Zustand zu repräsentieren. So konnten etliche „seltene“ Elementarereignisse schon vorab auf den Zustand `andere` abgebildet werden. Außerdem muß ein zusätzlicher Atomtyp, ein dreifachgebundenes Stickstoffatom, eingeführt werden, um Nitrile bzw. Cyanide (die –C≡N Gruppe) zu erfassen. Der zusätzliche Typ kann jedoch nicht direkt an ein aromatisches Koh-

lenstoffatom gebunden auftreten und ist somit für C_arom und s2ortho nicht von Bedeutung. Nach der Untersuchung der Stichprobe wurden schließlich noch alle Kombinationen, die nur Wasserstoff- oder gar keine Atome enthalten, als identisch betrachtet, da in all diesen Fällen ein Einfluß von $0,0$ ppm angenommen wird (vgl. Definition der Standardgruppen). Auf diese Weise gelangt man zu insgesamt 24 Zuständen.

Abschließend bleibt zu bemerken, daß in jedem Fall dokumentiert werden sollte, welche Belegungen jeder Zustand im einzelnen umfaßt (vgl. Abschnitt 8.2.2), um später die bedingten Wahrscheinlichkeiten beim Übergang vom Strukturmerkmal zum Teilinkrement bestimmen zu können. Der Gestaltung der Variablen zur Repräsentation der letzteren widmet sich der folgende Abschnitt.

6.2.2 Signalrepräsentation: Diskretisierung der ppm-Achse

Die Variable Peakpos sowie diejenigen zur Modellierung der Teilinkremente beziehen sich auf Ereignisse, deren Ausprägungen sich auf numerischer Ebene unterscheiden. Ihre Zustände repräsentieren also bestimmte Zahlenwerte, und es stellt sich die Frage, mit welcher Genauigkeit diese wiederzugeben sind, da die ppm-Skala prinzipiell kontinuierlich, die Zufallsvariablen des zu entwickelnden Bayes-Netzes jedoch diskret sind.

Betreffend die chemische Verschiebung eines Kohlenstoffatoms wird prinzipiell bereits im Spektrometer im Zuge der Digitalisierung eine Diskretisierung vorgenommen. Üblich ist dabei die Angabe auf eine, manchmal auch zwei Nachkommastellen genau. Jedoch ist der Wertebereich, in welchem die Absorptionen aromatischer Kohlenstoffatome liegen, sehr groß und reicht etwa von 90–185 ppm (vgl. z.B. [Bre92, BWe91]). Bei einer Diskretisierung mit einer Genauigkeit von 0,1 ppm und einer entsprechenden Gestaltung der die Lage des Peaks repräsentierenden Variablen ergäben sich damit rund 950 Zustände.

Es stellt sich jedoch die Frage, ob das Modell insgesamt überhaupt von einer derartigen Präzision profitieren kann. Zum einen wäre, um die entsprechende Detailtiefe bis zum Übergang zu den strukturbezogenen Variablen zu transportieren, eine vergleichbare Genauigkeit in den (Teil-)Inkrementen erforderlich. Diese würden dadurch ebenfalls sehr viele Zustände erhalten, deren Anzahl sich jeweils als Faktor auf den Umfang der bedingten Wahrscheinlichkeitsverteilungen $P(\texttt{Peakpos}|\texttt{Sphäre1},\texttt{Sphäre2})$ und $P(\texttt{Sphäre2}|\texttt{s2ipso},\texttt{s2ortho})$ auswirkt, wodurch das Bayes-Netz sehr stark aufgebläht würde.

Vor allem aber müßte eine Abweichung um 0,1 ppm auch einen spürbaren Einfluß auf die Rückschlußwahrscheinlichkeit der strukturellen Merkmale haben, damit diese Präzision einen Nutzwert für das Modell hat. Meßungenauigkeiten, abweichende Bedingungen beim NMR-Experiment oder ähnliche Faktoren können jedoch ohne weiteres die Ursache einer Abweichung in dieser Größenordnung sein, so daß sie nicht einmal zwingend auf einen konstitutionellen Strukturunterschied im Molekül zurückzuführen ist. Eine Vergröberung scheint nach diesen Überlegungen sinnvoll.

Als erstes sollen hierbei nun die Teilinkremente der zweiten Sphäre betrachtet werden. Die Variable i2ortho, die das Teilinkrement repräsentiert, das von der Besetzung der *s2ortho*-Atome abhängt, erhält ihren Gesamtwertebereich wie in Abschnitt 6.1.3 beschrieben: Für jede mögliche Klasse von *ortho*-Substituenten wird ein partielles Inkrement ermittelt und aus diesen paarweise die Summe gebildet, um den Wertebereich und die Verteilung der darin konkret vorkommenden Einzelwerte zu erhalten. Das so definierte Intervall reicht etwa von +20 ppm bis -31,5 ppm. Eine Einteilung in 0,1-ppm-Schritten würde in über 500 Zuständen resultieren.

Wählt man Zustände, welche jeweils Intervalle einer Schrittweite von 1 ppm repräsentieren, so gelangt man zu 52 Zuständen. Eine weitere Vergröberung ist zweifellos möglich, ausgehend von der Frage, ob eine Abweichung um 1 ppm auf einen strukturellen Unterschied schließen läßt, und insbesondere, ob dies in allen Bereichen des Gesamtintervalls gleichermaßen der Fall ist, da es deutlich weniger mögliche Substituentenkombinationen gibt, deren zugehörige Teilinkremente in den Randbereichen des Intervalls liegen.

Während von einer allgemeinen Vergröberung (etwa auf 2-ppm-Schritte) abgesehen wird, führt eine Analyse des verfügbaren Literaturwissens sowie der Stichprobe für die Randbereiche des ermittelten Intervalls zu den Schwellwerten -26 ppm und +7 ppm. Jenseits dieser Grenzen ist die Tatsache, daß das `i2ortho`-Teilinkrement oberhalb des Maximal- bzw. unterhalb des Minimalwerts liegt, das relevante Faktum, während der genaue Betrag des Teilinkrements von untergeordneter Bedeutung ist. Einschließlich je eines Zustandes für alle größeren und alle kleineren Inkrementwerte ergeben sich damit 36 Zustände. Für `i2ipso` führt dasselbe Vorgehen zu insgesamt 23 Zuständen innerhalb des Intervalls von -8 ppm bis +10 ppm einschließlich je eines Zustands für größere und kleinere Werte.

Anders ist die Lage im Fall der Variablen `Sphäre1`, welche das Inkrement der ersten Sphäre repräsentiert. Hier das gesamte Intervall der möglichen Einflüsse zu betrachten ist nicht sinnvoll: Es reicht von -6.2 ppm bis +35.1 ppm, so daß sich bei einer Genauigkeit von 0.1 ppm über 400 und bei einer Genauigkeit von 1 ppm immer noch über 40 Zustände ergäben. Interessant sind innerhalb des gesamten Intervalls jedoch nur die 14 Werte, welche zu den 14 betrachteten Atomtypen korrespondieren. Also werden 14 Zustände für `Sphäre1` definiert, die diesen 14 Werten entsprechen.

Dasselbe Prinzip, nur genau die zu den modellierten Atomkombinationen korrespondierenden Werte zu betrachten, ist für die Teilinkremente der zweiten Sphäre nicht möglich. Zum einen subsumieren die Teilinkremente implizit auch die Einflüsse der Atome jenseits der zweiten Sphäre. (In ähnlicher Weise fassen Meiler *et al.* beim Training der neuronalen Netze zur Spektrenvorhersage Informationen über Atome höherer Sphären zusammen [MMW00], da der für die Eingabe verwendeten HOSE-Code die Wiedergabe der Atome aller nicht explizit erfaßten Sphären in einer „Summensphäre“ vorsieht.) Würden diese ganz außen vor gelassen, so wäre das entwickelte Modell nur für Benzolderivate geeignet, deren Substituenten eine maximale Länge von 2 haben. Zum anderen sind, wie in Abschnitt 6.2.1 beschrieben, sehr viele Belegungen der betrachteten Positionen unterscheidbar, so daß das betrachtete Intervall annähernd vollständig abgedeckt wird.

Für die Variable `Sphäre2` schließlich ergeben sich die Zustände durch Summenbildung aus den Teilinkrementen, die von `i2ipso` und `i2ortho` wiedergegeben werden, wobei auch jeweils die beiden Zustände für außerhalb des in 1-ppm-Schritten abgedeckten Intervalls liegende Werte berücksichtigt werden. Sie werden als der größte bzw. kleinste ganzzahlige außerhalb des Intervalls liegende Wert betrachtet. So ergibt sich ein Intervall von -36 ppm bis +19 ppm für `Sphäre2`, wiederum zuzüglich je eines Zustands für alle größeren und alle kleineren Werte (insgesamt 58 Zustände).

Analog kann für `Peakpos` vorgegangen werden – der so erreichte Gesamtwertebereich deckt sich mit der der Literatur zu entnehmenden Angabe von etwa 90 bis 185 ppm. Wählt man Zustände, welche jeweils Intervalle einer Schrittweite von 1 ppm repräsentieren, so erreicht man eine Zahl von 96 Zuständen. Eine weitere Analyse des verfügbaren Literaturwissens hinsichtlich der Randbereiche des Intervalls, wie sie auch für `i2ipso` bzw. `i2ortho` durchgeführt wurde, zeigt, daß insbesondere unterhalb von 95 ppm und oberhalb von 173 ppm kaum Peaks zu erwarten sind. Mit diesen beiden Schwellwerten sowie je einem Zustand für alle größeren und alle kleineren Werte erreicht man eine Zahl von 81 Zuständen.

Hinsichtlich der durch Sphäre2 und Peakpos quasi in zwei Schritten durchgeführten Summation ist es natürlich denkbar, auf die Variable Sphäre2 zu verzichten und statt dessen i2ipso und i2ortho direkt in die Summenbildung zu Peakpos einzubeziehen. Betrachtet man das Modell ohne die Variable Sphäre2, so legt jedoch allein schon die Assoziativität der Summenbildung ein *Divorcing* (vgl. Abschnitt 4.3.2) der Variablen i2ipso und i2ortho von Sphäre1 nahe, was gerade der Einführung der Variablen Sphäre2 entspricht. Der effektive Vorteil wird im Vergleich der Zahl der Zustände von Sphäre2 mit der Zahl der Zustandskombinationen von i2ipso und i2ortho deutlich: Er fällt mit 58 Zuständen gegenüber $23 \cdot 36 = 828$ Kombinationen deutlich zugunsten von Sphäre2 aus.

6.3 Ergebnisse

Es wurde das in Abbildung 6.9 dargestellte Kausalmodell der Betrachtung von ^{13}C-NMR-Spektren zur Strukturaufklärung in der organischen Chemie entwickelt. Es betrachtet der Aufgabenstellung entsprechend die strukturellen Klassen von Substituenten an Benzolringen. Die einzelnen Variablen wurden in den vorangegangenen Abschnitten bezüglich der durch sie modellierten Ereignisse wie auch ihrer Zustände als deren Ausprägungen ausgearbeitet.

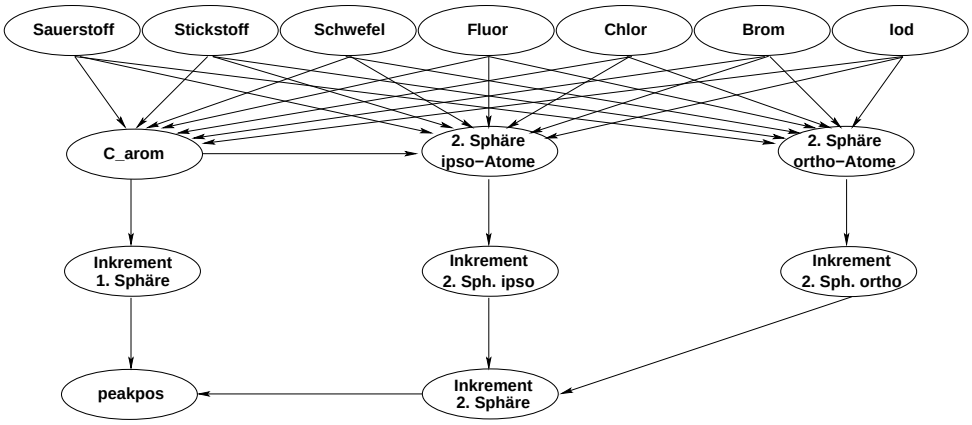


Abb. 6.9: *Kausales Modell zur Klassifikation von Substituenten an aromatischen Kohlenstoffatomen anhand von deren chemischer Verschiebung in ^{13}C-NMR-Spektren innerhalb des Systems* SASCHA *zur Erkennung von Substitutionsmustern an Benzolderivaten.*

Informationsvariablen des Bayes-Netzes, dessen Grundlage das entwickelte Kausalmodell ist, sind neben der Variablen Peakpos zur Eingabe der Peakposition die Variablen Sauerstoff, Stickstoff, Schwefel, Fluor, Chlor, Brom und Iod zur Eingabe der Summenformelinformation. Die Summenformelvariablen unterscheiden dabei nur Vorhandensein und Nichtvorhandensein des betreffenden chemischen Elements. Bei der Abbildung der tatsächlichen Lage eines Peaks von der kontinuierlichen ppm-Skala auf die Zustände von Peakpos ist der geringste Unterschied entscheidend.

Die Hypothesenvariable, die die gesuchte Information wiedergibt, ist C_arom. Ihre Zustände entsprechen den möglichen Klassen eines Substituenten, wobei die Kategorisierung nach dem Typ des Atoms in der ersten Sphäre vorgenommen wird. Die Zustände der Variablen sind im einzelnen Abbildung 6.7 zu entnehmen.

Auch die Variablen s2ortho und s2ipso können jedoch als Hypothesenvariablen dienen. Sie betrachten die in der zweiten Sphäre anwesenden Atomtypen und ihre unterschiedlichen

Kombinationen. s2ipso bezieht sich dabei auf den zum aktuell zu klassifizierenden Substituenten gehörigen Anteil; so ist eine feinere Unterscheidung von Substituentenklassen möglich, als dies allein durch C_arom geschieht. Gegenwärtig ist dies jedoch von untergeordneter Bedeutung. s2ortho beschreibt die Kombination von Substituentenklassen in den *ortho*-Positionen. Zur Gestaltung der Zustände der beiden Variablen s2ipso und s2ortho wurde eine Stichprobe hinsichtlich der Häufigkeit aller jeweils möglichen Kombinationen untersucht. Für s2ortho wurden dieselben Atomtypen wie bei C_arom betrachtet, für s2ipso kommt zusätzlich Stickstoff mit einer Dreifachbindung hinzu. Dies führte zu den in Abbildung 6.8 dargestellten Zuständen.

Außer ihrer jeweiligen Rolle als Hypothesenvariable können C_arom und s2ortho auch wechselseitig als Informationsvariable dienen. Wenn im Rahmen des Gesamtsystems für die Klassifikation eines bestimmten Substituenten Zusatzinformation über die Klassen seiner Nachbargruppen gesammelt werden soll, kann die wahrscheinlichste Substituentenkombination in den *ortho*-Positionen gegeben eine bestimmte *ipso*-Klasse ermittelt werden. Diese strukturelle Zusatzinformation kann zum einen zur Generierung von Substitutionsmusterhypothesen genutzt werden, kann darüber hinaus aber auch dazu dienen, frühere Klassifikationsergebnisse zu untermauern oder nötigenfalls auch zu revidieren.

Schließlich sind noch vier Zwischenvariablen vorhanden, welche Teilinkremente bezogen auf einen bestimmten Anteil der Molekülstruktur repräsentieren. Die Variablen Sphäre1 und Sphäre2 stehen für den Anteil der Atome der ersten bzw. zweiten Sphäre an der chemischen Verschiebung. Da die zweite Sphäre in der Betrachtung der Struktur in einen *ipso*- und einen *ortho*-Anteil untergliedert ist, wird auch das Inkrement der zweiten Sphäre aus zwei entsprechenden Anteilen, i2ipso und i2ortho, zusammengesetzt. Ihre Zustände sind in Abbildung 6.10 dargestellt.

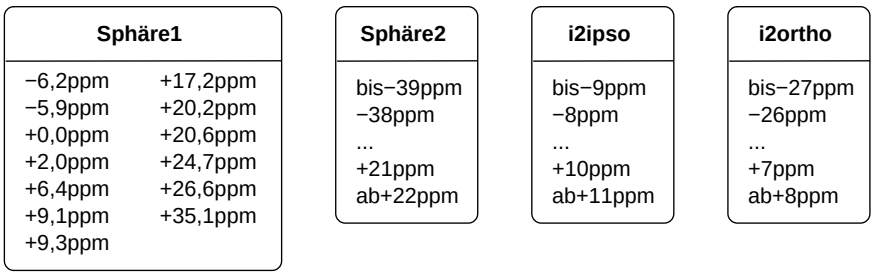


Abb. 6.10: *Gestaltung der auf die Teilinkremente bezogenen Variablen. Für* Sphäre2, i2ipso *und* i2ortho *setzt sich die Unterteilung im nicht dargestellten Bereich in 1-ppm-Intervallen fort.*

Das entwickelte Modell als Element der Wissensrepräsentation ist das Herzstück des in der Entwicklung befindlichen Musteranalysesystems SASCHA zur Erkennung von Substitutionsmustern an Benzolderivaten. Um jedoch zum einsatzfähigen System zu gelangen, sind einige weitere Schritte notwendig. Als erstes müssen die bedingten Wahrscheinlichkeiten zur Gewichtung der Kausalzusammenhänge sowie *a-priori*-Wahrscheinlichkeiten für die übrigen Variablen festgelegt werden, um das Kausalmodell zu einem Bayes-Netz zu vervollständigen. Dies ist Gegenstand von Kapitel 8.

Um die dabei durchzuführenden statistischen Untersuchungen zur Analyse der gegebenen Stichprobe vornehmen zu können, müssen jedoch erst die Grundvoraussetzungen geschaffen werden, um Spektral- und Strukturdaten lesen und verarbeiten zu können. Dies fällt in den Bereich des Vorverarbeitungsmoduls, das in Kapitel 7 ausgearbeitet wird. Es dient im

Gesamtsystem außerdem dazu, die eingehenden JCAMP-Daten für die Verarbeitung, im besonderen für die Eingabe von Evidenzen in das Bayes-Netz, aufzubereiten. Außerdem wird eine zumindest rudimentäre Erklärungskomponente zur Ergebnisausgabe benötigt.

Im Anschluß an die Klassifikation sind die Ergebnisse des Bayes-Netzes jeweils für die sechs Peaks eines Benzolrings der Hypothesengenerierung zu übergeben. Die Ergebnisse sollen, wie in Abschnitt 5.1.4 angedeutet, die Gestalt dreigliedriger Ringsegmente haben, um basierend auf deren Überschneidung den gesamten Ring zusammensetzen zu können. Entsprechend ist die Anfrage an das Bayes-Netz zu formulieren. Falls sich die ermittelten Fragmente nicht mit einander in Einklang bringen lassen, besteht grundsätzlich die Möglichkeit einer Revision der Klassifikationsergebnisse, jedoch erfordert dies, wie in Kapitel 9 näher ausgeführt werden soll, umfangreiche Untersuchungen und Überlegungen zur Erarbeitungen einer geeigneten Kontrollstrategie.

Schließlich bleibt noch zu erwähnen, daß es im Rahmen des Modells, bei der Repräsentation seiner einzelnen Aspekte, der Ausgestaltung der einzelnen Variablen sowie im Bereich der bedingten und der *a-priori*-Wahrscheinlichkeiten viel Raum für Varianten gibt. Die Evaluierung des Bayes-Netzes als Klassifikator, die Gegenstand von Kapitel 10 ist, dient dabei der Beurteilung von Modellierungsentscheidungen. Vor dem Hintergrund der unbestrittenen Notwendigkeit von Anpassungen und Präzisierungen in der Modellierung geben die dabei zu erzielenden Erkenntnisse Aufschluß darüber, in welcher Weise die anzustrebenden Weiterentwicklungen zielgerichtet und erfolgversprechend umgesetzt werden können.

7 Vorverarbeitung und Merkmalsextraktion

Für den praktischen Einsatz eines Musteranalysesystems ist die Aufbereitung der Eingangsdaten der fundamentale Schritt. Je nach Domäne, Anwendung und Art der Daten können dabei unterschiedliche Vorverarbeitungsschritte sowie eine Merkmalsberechnung stattfinden. In jedem Fall muß die zu verarbeitende Information in eine Repräsentation gebracht werden, welche systemintern weiterverarbeitet werden kann.

Im Fall des Musteranalysesystems SASCHA dient das Vorverarbeitungsmodul dem Lesen von spektroskopischen und Strukturdaten im JCAMP-Format sowie der Extraktion derjenigen Information, die durch das System verarbeitet werden soll, das heißt der Summenformel sowie der sechs zu einem Benzolring korrespondierenden Peaks. Darüber hinaus muß auch Strukturinformation bereitgestellt werden: Für die Evaluation des Systems wird das tatsächliche Substitutionsmuster des Benzolrings benötigt, und zur Initialisierung oder Adaption der Systemparameter, das heißt zur Bestimmung der bedingten Wahrscheinlichkeiten, welche innerhalb eines Bayes-Netzes die Kausalverknüpfungen des in Kapitel 6 entwickelten Kausalmodells gewichten, sind statistische Untersuchungen einer Stichprobe betreffend den Zusammenhang zwischen spektroskopischen und strukturellen Merkmalen erforderlich.

Benötigt werden im einzelnen Funktionen zum Lesen von JCAMP-Dateien sowie zur Extraktion der gewünschten Information. Dazu ist eine Beschäftigung mit dem JCAMP-Dateiformat erforderlich, außerdem müssen die Ansprüche der übrigen Systemkomponenten an das Vorverarbeitungsmodul, wie in Kapitel 5, Abschnitt 5.1.2 beschrieben, beachtet werden.

Im folgenden beschäftigt sich Abschnitt 7.1 mit dem JCAMP-Datenformat und seinen an dieser Stelle wichtigen Details. Anschließend widmet sich Abschnitt 7.2 der Entwicklung geeigneter Funktionen und Datenstrukturen, welche den Ansprüchen des in der Entwicklung befindlichen Systems gerecht werden. Abschnitt 7.3 resümiert die Entwicklung des Vorverarbeitungsmoduls und gibt einen Ausblick auf die nächsten zu realisierenden Arbeiten im Rahmen des angestrebten Gesamtsystems.

7.1 Verarbeitung von JCAMP-Daten

Die Eingangsinformation für SASCHA wird in Gestalt von JCAMP-Dateien zur Verfügung gestellt. Dieses Dateiformat hat den Vorteil, daß ein direkter Bezug zwischen Struktur- und Spektraldaten möglich ist, wie einführend in Abschnitt 5.2.1 erwähnt. Im folgenden soll nun eine genauere Betrachtung der für diese Arbeit relevanten Details der Formate JCAMP-DX 5.00 für Spektraldaten [DL93] sowie JCAMP-CS 3.7 für Strukturdaten [GHH+91] erfolgen. Darauf aufbauend werden anschließend Datenstrukturen und Funktionen zur Nutzbarmachung der enthaltenen Information für das System entwickelt.

7.1.1 Aufbau von JCAMP-Dateien

JCAMP-Dateien sind grundsätzlich sehr einfach aufgebaut. Sie bestehen aus einer Folge sogenannter *LDR*s (engl. *labeled data record*: „etikettiertes Datenfeld“), die sich weiter in den *Bezeichner* und dessen zugehörigen *Eintrag* untergliedern lassen. Für beides wird der ASCII Zeichensatz benutzt.

Der Bezeichner eines LDRs gibt an, welche Information der zugehörige Eintrag enthält. Er besteht aus dem Bezeichnernamen, der von den Zeichen ## und = eingefaßt ist (z.B. ##Title=) und beginnt, bis auf Zwischenraum, am Zeilenanfang. Groß- und Kleinschreibung spielt für Bezeichner keine Rolle, ferner werden alle Zwischenräume, Bindestrich (-), Schrägstrich (/) und Unterstrich (_) ignoriert. Drei Arten von LDRs sind anhand der Gestalt ihrer Bezeichner zu unterscheiden: Beginnt der Bezeichnername unmittelbar nach ##, so enthält der Eintrag globale Information. Folgt auf ## ein Punkt vor dem Bezeichnernamen, so handelt es sich beim zugehörigen Eintrag um datentypspezifische Angaben, die nur für die aktuelle Untersuchungsmethode (z.B. NMR) sinnvoll sind. Außerdem kann dem Bezeichnernamen $ vorausgestellt sein, in diesem Fall handelt es sich um einen benutzerdefinierten LDR.

Die automatische Verarbeitung von JCAMP-Dateien wird durch diese benutzerdefinierten LDRs erschwert: Sie sind zwar ohne weiteres erkennbar, die Bedeutung der eingetragenen Daten ist jedoch unklar, auch wenn sie sich dem menschlichen Leser anhand des Bezeichners leicht erschließen mag. Sollen benutzerdefinierte LDRs ausgewertet werden, so ist es unumgänglich, einen systeminternen Standard zu definieren (wie z.B. für das Datenbanksystem SPECINFO [Spe98]), welcher festlegt, welche benutzerdefinierten LDRs mit welcher Bedeutung verwendet werden.

Für die Einträge der LDRs sind im wesentlichen textuelle und numerische Einträge als elementare Datentypen zu unterscheiden. Betreffend numerische Einträge ist an dieser Stelle nur das *AFFN*-Format[1] von Bedeutung. Beginnt ein Eintrag mit +, -, einem Dezimalpunkt oder einer Ziffer, so wird er als Zahl interpretiert. Neben diesen Zeichen darf lediglich E (für die Exponentenschreibweise) im Eintrag vorkommen. Außerdem ist ein einzelnes ? als Eintrag für unbekannte oder außerhalb des Meßbereichs liegende Werte zulässig.

Von textuellen Einträgen sind zwei Varianten zu unterscheiden: *String*-Einträge sind alphanumerische Einträge, die für die automatische Verarbeitung vorgesehen sind und somit bestimmten Regeln folgen müssen, die dem durch den Bezeichner des LDRs angegebenen Typ entsprechen. Dies kann eine Liste zulässiger Einträge sein, etwa für ##DataType= zur Bezeichnung der Art der spektroskopischen Daten, oder ein bestimmtes Schema, nach welchem die enthaltene Information anzuordnen ist. *Text*-Einträge sind dagegen nicht für die automatische Verarbeitung vorgesehen und haben natürlichsprachliche Form.

Die Datenfelder einer JCAMP-Datei können darüber hinaus durch Kommentare ergänzt werden, welche durch $$ eingeleitet werden und sich bis zum Zeilenende erstrecken. Sie beenden jedoch den LDR bzw. seinen Eintrag nicht. Ein LDR endet mit dem Beginn des nächsten LDRs oder mit EOF (*end of file*). Wird beides nicht vorgefunden, so setzt sich der Eintrag nach dem Ende des Kommentars fort, es sei denn die folgende Zeile beginnt mit $$: In diesem Fall setzt sich dort der Kommentar fort. Kommentare sind vom Datentyp *Text* und somit nicht für die automatische Auswertung geeignet.

Neben den durch ihre Bezeichner unterschiedenen Arten von einzelnen LDRs lassen sich auch Folgen von LDRs, sogenannte *Block*s, zusammengefaßt betrachten. Ein Block ist stets von LDRs mit den Bezeichnern ##Title= und ##End= eingefaßt. Man unterscheidet zunächst nach dem JCAMP-DX-Standard *Link-Block*s und *Datenblock*s, JCAMP-CS Blocks sind ebenfalls als Datenblocks zu verstehen.

JCAMP Dateien, die nur aus einem einzigen Datenblock bestehen, werden *Simple JCAMP* genannt. *Compound JCAMP* Dateien enthalten dagegen mehrere Blocks, die über *Link*-Blocks hierarchisch verschachtelt sein können. Alternativ können sie als Aneinanderreihung von (verschachtelten oder einfachen) Blocks die Gestalt eines Archivs haben.

[1] *ASCII Free Format Numeric*

Alle Blocks besitzen bestimmte gemäß dem Standard obligatorische LDRs, die als *Kern* bezeichnet werden. Dem können in *Link*- und JCAMP-DX-Datenblocks Informationen in weiteren LDRs vorangestellt sein, die als *Notes* bezeichnet werden und dazu dienen, die Beschreibung der durchgeführten Untersuchung angemessen zu vervollständigen, jedoch nicht obligatorisch sind. In JCAMP-CS-Datenblocks kann der Kern von weiteren, wiederum nicht obligatorischen LDRs gefolgt werden, die als *Shell* bezeichnet werden und Zusatzangaben betreffend die Molekülstruktur enthalten. Abgesehen von der Unterscheidung zwischen Kern- und zusätzlicher Information ist die absolute Position eines LDRs innerhalb eines Blocks jedoch irrelevant, er wird ausschließlich über seinen Bezeichner identifiziert, über welchen auch die Art der enthaltenen Information kenntlich ist.

7.1.2 Zugang zur enthaltenen Information

Nicht alle obligatorischen LDRs enthalten Information, welche für SASCHA von Bedeutung ist. Umgekehrt gibt es auch nichtobligatorische LDRs, die wichtige Information enthalten, z.B. Angaben über besondere Bindungssysteme innerhalb der Molekülstruktur oder Visualisierungsinformation, die in zukünftigen Arbeiten genutzt werden könnte. Im folgenden werden die auszuwertenden LDRs und das Format ihrer Einträge näher betrachtet. Die in den übrigen LDRs enthaltene Information ist ausführlich in [DL93] bzw. [GHH+91] beschrieben.

Die Auswahl der auszuwertenden LDRs erfolgt selbstverständlich ausgehend von der gemäß der Aufgabenstellung benötigten Information (vgl. Abschnitt 5.1.2). Für jeden davon wird das Format der betreffenden Einträge gemäß der Definition betrachtet sowie die zur Verfügung stehende Stichprobe betreffend des Einhaltens des JCAMP-Standards untersucht. Davon ausgehend können Funktionen entwickelt werden, welche die gewünschte Information verfügbar machen. Ferner sind geeignete Datenstrukturen Voraussetzung für die weitere Nutzung. Mit der Entwicklung von beidem befaßt sich Abschnitt 7.2.

Seitens der Spektraldaten werden prinzipiell nur die Positionen der sechs zu den Kohlenstoffatomen des Benzolringes korrespondierenden Peaks benötigt. Peakpositionen können grundsätzlich aus `##Peak Table=` oder `##Peak Assignments=` LDRs entnommen werden. Nur im letzteren Fall ist jedoch eine Zuordnung zwischen einzelnen Peaks und einzelnen Kohlenstoffatomen gegeben, so daß die zum Benzolring gehörigen Peaks identifiziert werden können. Ein Datenblock darf nur einen der beiden genannten LDRs enthalten; welche Art von Daten enthalten ist, gibt der LDR `##DataType=` an. In dem betreffenden Block könnten außerdem die Einträge von `##JCAMP-DX=` (Versionsnummer) und `##.ObserveNucleus=` (untersuchtes chemisches Element) überprüft werden. Darüber hinaus sind die LDRs `##XFactor=` und `##YFactor=` von Bedeutung, welche Skalierungsfaktoren für die Angaben bezüglich der Peaks beinhalten.

Die Einträge von `##PeakAssignments=` LDRs haben folgende Gestalt[2]: Zunächst beschreibt die Angabe `(XYMA)` den Aufbau der Elemente der auf sie folgenden Liste – chemische Verschiebung *x* und Intensität *y* des Peaks, seine Multiplettstruktur *m* und eine Referenz *a* auf das verursachende Atom. Die einzelnen Elemente der Liste sind in Klammern gefaßt, ihre Komponenten *x*, *y*, *m* und *a* sind durch Kommata getrennt, und *a* ist durch Einfassung in spitze Klammern als *String* gekennzeichnet. Leerzeichen innerhalb des Eintrags sind ohne Bedeutung. In der vorliegenden Stichprobe ist *a* stets der Index des verursachenden Atoms gemäß `##AtomList=`.

[2] Die gemäß [DL93] zulässigen Alternativen kommen in der Stichprobe nicht vor.

Dieser LDR ist Teil der benötigten Strukturinformation. Mit ihrer Hilfe werden die sechs Benzolpeaks innerhalb des Spektrums identifiziert, und zu Zwecken der Evaluation und der Parameteradaption werden statistische Untersuchungen der Stichprobe an ihr vorgenommen. Auch die als Eingabedatum erforderliche Summenformelinformation kann den Strukturdaten entnommen werden. Die Datenfelder `##AtomList=`, `##BondList=` sowie das gemäß [Spe98] definierte Feld `##$BondTypes=` enthalten die benötigten Angaben. Außerdem könnte der LDR `##JCAMP-CS=` (Versionsnummer) überprüft werden.

Der `##AtomList=` LDR enthält eine numerierte Liste der Atome des Moleküls. Gemäß des JCAMP-Standards sollte die Auflistung aus Tripeln bestehend aus einer Zahl *n*, einem Elementsymbol *s* und einer weiteren Zahl *h* bestehen. Dabei gibt *n* den Index des Atoms an, *s* sein chemisches Element, und *h* die Anzahl nicht explizit im Verlauf der Liste angegebener Wasserstoffatome, die an dieses Atom gebunden sind. In den Dateien der Stichprobe fehlt jedoch durchgängig die Angabe von *h*, obwohl keinerlei Wasserstoffatome explizit aufgelistet werden, was beim Lesen der Information berücksichtigt werden muß.

Es kann jedoch versucht werden, die fehlende Information zu rekonstruieren. Der LDR `##BondList=` enthält eine Liste der Bindungen im Molekül. Sein Eintrag wird ohnehin für die Repräsentation der Molekülstruktur für die Anwendungsfälle der Evaluation und der Parameteradaption benötigt. Darüber hinaus kann überprüft werden, ob Atome freie Valenzen besitzen; diese sind durch Bindungen zu Wasserstoffatomen zu besetzen.

In den meisten Fällen ist dieses Vorgehen erfolgreich, besitzt jedoch ein chemisches Element die Möglichkeit einer Valenzschalenerweiterung, so ist seine aktuelle Valenz unsicher. Im Rahmen der vorliegenden Arbeit trifft dies nur im Fall von Schwefel zu. Hier wird davon ausgegangen, daß Schwefel nur mit den beiden Valenzen 2 (Thioether und Disulfide) und 6 (Sulfonate) vorkommt. Werden weniger als zwei Bindungen gefunden, wird bis zur Gesamtzahl von zwei Bindungen ergänzt; nur bei mindestens drei explizit aufgezählten Bindungen werden etwaige freie Valenzen durch Wasserstoffatome bis zu einer Gesamtzahl von sechs Bindungen besetzt.

Die Gestalt des `##BondList=` LDRs ist eine Liste von Tripeln bestehend aus zwei Zahlen und einem Buchstaben. Die beiden Zahlen sind Atomindizes, der Buchstabe gibt die Art der Bindung zwischen den indizierten Atomen an (`S`, `D`, `T`, `Q` für Ein- bis Vierfachbindungen, `A` für andere). Die Bindungsliste selbst ist nicht explizit indiziert, sie ist jedoch in Zeilen untergliedert, deren implizite Numerierung als Indizes zu verstehen ist. Dies ist wichtig, da sich der Eintrag von `##$BondTypes=` darauf bezieht.

Der LDR `##$BondTypes=` enthält eine zeilenweise Liste von Zahlentupeln, die Bindungssysteme beschreiben. Die erste Zahl gibt jeweils den Typ des Systems an (z.B. steht Typ 5 für aromatische Systeme), die darauffolgenden Zahlen sind die Indizes der an diesem System beteiligten Bindungen. Obwohl `##$BondTypes=` ein benutzerdefinierter LDR ist, wird die hier verfügbare Information genutzt: Anderenfalls müßte in der durch Atom- und Bindungsliste gegebenen Molekülstruktur aufwendig nach Benzolringen gesucht werden (Suche nach Zyklen in ungerichteten Graphen mit gefärbten Kanten und Knoten, wobei nur Zyklen mit einer bestimmten Größe, Knoten- und Kantenfärbung gewünscht sind). In `##$BondTypes=` sind jedoch aromatische Systeme und die zu ihnen gehörigen Bindungen bereits vermerkt.

Mit Blick auf zukünftige Entwicklungen des Erklärungsmoduls ist schließlich der LDR `##XYRaster=` interessant. Er gehört nicht zu den obligatorischen LDRs, ist aber in der Stichprobe durchgängig vorhanden. Sein Eintrag enthält Zahlentripel, die aus einem Atomindex sowie *x*- und *y*-Koordinate bestehen. Diese Information kann später zur Visualisierung benutzt werden.

7.2 Funktionen und Datenstrukturen

Nach diesem Überblick über die für SASCHA wichtigen Datenfelder, ihren Aufbau und die enthaltene Information stellt sich als nächstes die Frage der Speicherung in einer für die weitere Verarbeitung günstigen Form. Einige Grundgedanken wurden bereits in Abschnitt 5.1.2 angerissen. An dieser Stelle sollen sie weiter ausgearbeitet werden.

Dabei sind im Vorverarbeitungsmodul zwei Sichtweisen von Bedeutung: zum einen der Anwendungsfall der Spektrenauswertung und zum anderen die Anwendungsfälle der Evaluation und des Lernens, welche zusätzliche, strukturbezogene Information benötigen. Daher wird das Modul in zwei Teilschritte untergliedert: Der erste Teilschritt orientiert sich eng an der in JCAMP-Daten unmittelbar enthaltenen Information, während im zweiten Teilschritt mit Blick auf Evaluierung und Parameteradaption mehr Gewicht auf eine günstige Repräsentation von Strukturinformation gelegt wird. Die Entwicklung folgt im Prinzip einer objektorientierten Sichtweise.

Die stark JCAMP-bezogene Ausrichtung des ersten Teilschritts ist des weiteren für eine spätere Wieder- und Weiterverwendung des Vorverarbeitungsmoduls von Belang. Dieser Schritt dient dazu, zunächst einmal die in einzelnen LDRs enthaltene Information zugänglich zu machen. Wenn die für die gegenwärtige Problemstellung optimierte Repräsentation des zweiten Teilschritts sich für andere Aufgaben als ungeeignet erweist, kann die JCAMP-orientierte Repräsentation ungeachtet dessen verwendet und bedarfsweise in eine neue Repräsentation überführt werden.

7.2.1 In JCAMP-Daten enthaltene Information

Die gegebene Stichprobe besteht aus Dateien, die jeweils einen *Link*-Block mit drei weiteren, inneren Blocks enthalten. Dies sind ein JCAMP-CS Block mit Strukturdaten sowie zwei Datenblocks mit spektralen Daten, zum einen ohne und zum anderen mit Zuordnung der Peaks zu den einzelnen Atomen. Bei der Verarbeitung muß jedoch berücksichtigt werden, daß viele Benutzer (und Systeme) sich nicht streng an den JCAMP-Standard halten, so daß möglichst wenige Voraussetzungen hinsichtlich eines bestimmten erwarteten Aufbaus der Eingabedateien gemacht werden sollten. Es sollten nur die unbedingt benötigten LDRs betrachtet werden, und auch bei der Extraktion von deren Information sollte das Vorverarbeitungsmodul möglichst tolerant sein.

Es wäre z.B. grundsätzlich denkbar, die hierarchische Blockstruktur der LDRs auszunutzen, um etwa durch Auswertung des Feldes `##DataType=` zu prüfen, daß in der Tat mehrere Blocks gegeben sind, oder uninteressante Blocks zu übergehen. Durch ähnliche Überprüfungen wäre eine Standardisierung der Experimentbedingungen oder eine Überprüfung des JCAMP-Standards realisierbar. Nichts davon wird jedoch durchgeführt, um das Vorverarbeitungsmodul wie oben erwähnt in jeder Hinsicht so tolerant wie möglich zu halten.

Im Falle der Standardisierung von Experimentbedingungen kommt hinzu, daß viele der in Frage kommenden LDRs nicht obligatorisch sind, so daß oftmals das Einhalten oder nicht Einhalten der Normbedingungen gar nicht überprüft werden kann, und daß außerdem die Definition der Gestalt der Einträge recht große Freiheiten läßt, was einen automatischen Abgleich erschwert. Dennoch ist die Behandlung derartiger Informationen, sei es zur Schaffung einer Norm bezüglich der Bedingungen des NMR-Experiments oder als Zusatzinformation bei der Auswertung, ein interessanter Aspekt, der Gegenstand späterer Arbeiten sein kann.

Derzeit reicht es jedoch aus, wenn diejenige Information zur Verfügung steht, die SASCHA unbedingt benötigt, damit die Eingabe akzeptiert und die Verarbeitung fortgesetzt wird. Dazu wird ein Abgleich der Bezeichner der einzelnen Datenfelder durchgeführt. Der Bezeich-

ner definiert die Art des Datenfeldes, durch welche feststeht, welche Information in welcher regulären Form enthalten ist. Der Aufbau der zu diesem Zeitpunkt interessanten LDRs ##PeakAssignments=, ##AtomList=, ##BondList= und ##$BondTypes= wurde bereits in Abschnitt 7.1.2 beschrieben. Darüber hinaus wird der LDR ##XFactor= für die Skalierung der gegebenen *x*-Werte der Peaks benötigt.

Hinsichtlich der Spektraldaten könnte ein Objekt vom Typ Peak definiert werden, dessen Eigenschaften seine Lage, Intensität und Multiplettstruktur sowie eine Referenz auf das verursachende Atom sind. Eine Liste solcher Objekte würde dem LDR ##PeakAssignments= wie in der Stichprobe gegeben entsprechen. Es werden jedoch für SASCHA lediglich die chemische Verschiebung und der Bezug zum verursachenden Atom benötigt, und dies sind auch gemäß Definition [DL93] die einzigen obligatorischen Angaben in ##PeakAssignments=. Daher wird eine andere Realisierung gewählt, welche Peakzuordnungsliste in die Betrachtung von Atomen mit einbezieht und nicht unabhängig davon prozessiert: Dem Inhalt von ##AtomList= folgend sind die grundsätzlichen Eigenschaften eines Atoms sein Index und sein chemisches Element, ##PeakAssignments= liefert darüber hinaus die chemische Verschiebung als eine zusätzliche Eigenschaft.

Betreffend Bindungen sind gemäß ##BondList= Referenzen auf die beiden beteiligten Atome sowie der Typ der Bindung als Attribute erforderlich. Außerdem erhält jedes einzelne Bindungs-Objekt einen Index, welcher der Zeilennummer in ##BondList= entspricht, da diese implizite Information für die Betrachtung von Bindungssystemen im Feld ##$BondTypes= benötigt wird. Ein Bindungssystem besteht aus einer Liste von Bindungen und besitzt einen Typ, darüber hinaus erhält es ebenfalls einen Index, der sich aus der Zeilennummer ergibt. Die beschriebenen Datenstrukturen sind in Abbildung 7.1 dargestellt. darüber hinaus wird ein Typ CSInfo (JCAMP-*CS Info*rmation) definiert, welcher Atomliste, Bindungsliste und Liste der Bindungssysteme einer Verbindung gesammelt zur Verfügung stellt. Mit diesen Objekten sind die im folgenden aufgezählten Funktionen assoziiert.

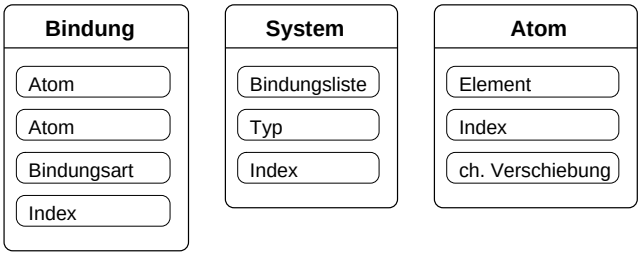


Abb. 7.1: *Datenstrukturen zur Repräsentation der in JCAMP-Dateien enthaltenen Information.*

- Lesen der Atomliste. Die Funktion erhält den ##AtomList= LDR aus einer JCAMP-Datei als Zeichenkette und liefert eine Liste von Atom-Objekten. Die chemische Verschiebung bleibt undefiniert.
- Lesen der Bindungsliste. Die Funktion erhält den ##BondList= LDR aus einer JCAMP-Datei als Zeichenkette sowie eine Atomliste und liefert eine Liste von Bindungs-Objekten.
- Lesen der Liste von Bindungssystemen. Die Funktion erhält den ##$BondTypes= LDR aus einer JCAMP-Datei als Zeichenkette sowie eine Bindungsliste und liefert eine Liste von (Bindungs-)System-Objekten.

- Lesen der Liste von Peakzuordnungen. Die Funktion erhält eine Atomliste sowie den `##PeakAssignments=` LDR aus einer JCAMP-Datei. Die Atom-Objekte der gegebenen Liste werden dem LDR-Eintrag entsprechend modifiziert.

- Lesen des Skalierungsfaktors für die chemische Verschiebung. Die Funktion erhält den `##XFactor=` LDR aus einer JCAMP-Datei als Zeichenkette sowie eine Atomliste. Die Atom-Objekte der gegebenen Liste werden dem Eintrag des LDRs entsprechend modifiziert.

- Nachbearbeiten der zusammengehörigen Atom- und Bindungsliste eines CSInfo-Objekts. Wie in Abschnitt 7.1.2 beschrieben müssen diese Listen um die nicht explizit aufgeführten Wasserstoffatome ergänzt werden. Die Funktion erhält eine Atomliste und eine Bindungsliste als Parameter. Für jedes Atom wird basierend auf seinem chemischen Element und der Bindungsliste die Zahl seiner freien Valenzen bestimmt. Entsprechend werden Wasserstoffatome der Atomliste und Bindungen zu diesen der Bindungsliste hinzugefügt.

Mit dem Abschluß des ersten Teilschritts, das heißt dem Raffinieren der Rohdaten zu Information, sind alle nötigen Voraussetzungen gegeben, um die für den Anwendungsfall der Spektrenauswertung (gegenüber Parameteradaption und Evaluation) nötigen Merkmale zu liefern. Dies sind die Evidenzen, die mit dem Ziel der Klassifikation der einzelnen Positionen des Benzolrings in das in Abschnitt 6 entwickelte Bayes-Netz eingegeben werden sollen.

Es wird also Information betreffend die einzelnen chemischen Elemente sowie die Lage der sechs Benzolpeaks benötigt. Diese soll in einer für das Bayes-Netz verarbeitbaren Form zur Verfügung gestellt werden. Anfragen an das Bayes-Netz erfolgen in Form einer in sich untergleiderten Zeichenkette. Der erste Abschnitt enthält die Anfragevariable, alle folgenden geben an, in welchem Zustand sich jeweils eine Informationsvariable befindet und haben die Form `Variablenname=Zustand`. Die folgenden Funktionen stehen in diesem Zusammenhang innerhalb des ersten Teilschritts der Vorverarbeitung zusätzlich zur Verfügung:

- Feststellen, ob ein bestimmtes chemisches Element an der Verbindung beteiligt ist. Die Funktion erhält ein CSInfo-Objekt und das betreffende chemische Element als Parameter und liefert die Zahl der Vorkommen des Elements zurück. Es wird nicht nur das Vorhandensein oder Fehlen des Elements, sondern die Zahl der Vorkommen ermittelt, um die Wiederverwendbarkeit im Falle einer Erweiterung des Modells zu gewährleisten.

- Feststellen der Lage der Benzolpeaks. Die Funktion erhält ein CSInfo-Objekt als Eingabe. In der Bindungssystemliste werden aromatische Systeme, die aus sechs Bindungen bestehen, gesucht. Die beteiligten Atome werden, sofern alle davon Kohlenstoffatome sind, in einer Liste zusammengefaßt. Zurückgeliefert werden alle gefundenen Ringe in Gestalt solcher Atomlisten. Die Lage der Peaks ist als Eigenschaft den einzelnen Atom-Objekten zu entnehmen.

- Ausgabe der Peak- und Summenformelinformation in einer Form, die vom Bayes-Netz als Eingabe verarbeitet werden kann. Die Funktion erhält ein CSInfo-Objekt als Eingabe und liefert eine Liste von Zeichenketten in der gewünschten Form. Die vorgenannten Funktionen werden benutzt, um die benötigte Information festzustellen.

7.2.2 Repräsentation von Molekülstrukturen

Die Adaption der bedingten Wahrscheinlichkeiten des Bayes-Netzes sowie die Evaluierung der Klassifikationsleistung und des Gesamtsystems erfordern eine tiefgreifendere Betrachtung chemischer Strukturen. Darauf zielt der zweite Teilschritt der Vorverarbeitung ab und ist somit hauptsächlich für die beiden Anwendungsfälle der Evaluation und der Parameteradaption relevant. Die in Abschnitt 5.1.2 angerissenen problemspezifischen Überlegungen hinsichtlich der Repräsentation von Strukturdaten sollen zu diesem Zweck im folgenden genauer ausgearbeitet werden.

Im Kontext einer stark auf Molekülstrukturen bezogenen Sichtweise wird die chemische Verschiebung, wie bereits in Abschnitt 7.2.1 beschrieben, lediglich als Eigenschaft des betrachteten Atoms gesehen. Objekte des Typs Atom können auch im zweiten Teilschritt der Vorverarbeitung verwendet werden. Darüber hinaus wird jedoch ein Objekttyp benötigt, der so gestaltet ist, daß zum einen die Zuordnung zwischen spektraler und strukturbezogener Information erhalten bleibt, und welcher zugleich eine Kategorisierung struktureller Merkmale in Anlehnung an den HOSE-Code ermöglicht.

Ein solches Objekt wäre als Strukturelement zu bezeichnen. Es besteht aus einem Atom, seinem Typ (im Sinne seiner Bindungen) sowie seinen Nachbarn. Die Nachbarn sind Verweise auf andere Strukturelemente. Abbildung 7.2 veranschaulicht diese Realisierung.

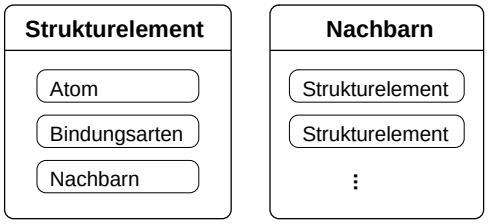


Abb. 7.2: *Datenstrukturen zur Repräsentation von Molekülstrukturen. Durch die Liste von Referenzen auf benachbarte Strukturelemente wird eine nach Sphären untergliederte Betrachtung ähnlich der Darstellung im HOSE-Code ermöglicht.*

Ein Molekül ist im Sinne dieser Repräsentation eine Menge von Strukturelementen (einatomigen Strukturbausteinen). Die Beziehung zwischen ihnen, die sich auf der Bindungsstruktur im Molekül begründet, wird jeweils über die Liste der Nachbarn hergestellt, die Referenzen auf die benachbarten Strukturelemente enthält. So ist ausgehend von jedem beliebigen Atom nachvollziehbar, welche anderen Atome eine bestimmte Anzahl von Bindungen weit entfernt sind, was eine nach Sphären untergliederte Betrachtung in Anlehnung an den HOSE-Code erleichtert. Darüber hinaus ist die Repräsentation nicht auf die in der Modellierung berücksichtigte Anzahl von Sphären rund um den Benzolring beschränkt, sondern es kann das ganze Molekül repräsentiert werden. Im Falle einer Erweiterung des Modells sind somit keine Anpassungen nötig.

Für die Evaluation des Klassifikators sind natürlich die möglichen Klassen interessant, das sind im einzelnen für jede Position des Benzolrings das Atom der ersten Sphäre in der *ipso*-Position sowie die *ipso*- und *ortho*-Atome der zweiten Sphäre. Dieselben Atome sind auch hinsichtlich der Parameteradaption interessant, da sie bzw. ihre Kombination sich in den Variablen des Bayes-Netzes und deren Zuständen widerspiegeln.

Vor diesem Hintergrund ist es von Bedeutung, den Benzolring innerhalb der Molekülstruktur zu erkennen. Dies ist auf der Repräsentation als Strukturelement-Objekte schwierig, je-

doch sind der JCAMP-Information des ##$BondTypes= LDRs etwaige besondere Bindungssysteme, wie Benzolringe als aromatische Systeme, zu entnehmen. Daher werden die Objekte der JCAMP-orientierten Repräsentation nach ihrer Überführung in die Strukturelement-Repräsentation nicht zerstört, sondern ebenfalls bereitgehalten. Dieser Umstand unterstreicht ihre Vorteilhaftigkeit hinsichtlich Wieder- und Weiterverwendung. Im Rahmen des zweiten Vorverarbeitungsteilschrittes sind daher auch einige Funktionen definiert, welche Objekte des ersten Teilschritts verarbeiten, wie sich in der folgenden Auflistung zeigt:

- Repräsentation eines Moleküls als Liste von Strukturelementen. Die Funktion erhält ein CSInfo-Objekt (also die gesamte im ersten Teilschritt gelesene JCAMP-Information) als Eingabe und liefert eine Liste von Strukturelement-Objekten, welche untereinander den Bindungen entsprechend referenziert sind.
- Identifizieren des zu einem Atom korrespondierenden Strukturelements. Die Funktion erhält eine Liste von Strukturelementen (ein Molekül) sowie ein Atom-Objekt und liefert das diesem entsprechende Strukturelement der Liste.
- Erstellen einer Abstandskarte der Atome innerhalb eines Moleküls. Die Funktion erhält ein Molekül als Eingabe und liefert eine Matrix der minimalen paarweisen Abstände.
- Aufzählen aller Atome in gegebener Entfernung von einem Fokusatom. Die Funktion erhält ein Molekül, eine Abstandskarte, den gewünschten Abstand und das Strukturelement des Fokusatoms. Sie liefert eine Liste von Strukturelementen.
- Aufzählen aller Atome in einer bestimmten Entfernung und relativen Lage von einem Fokusatom in einem Benzolring. Anders als das bereits beschriebene Aufzählen aller Atome in gegebener Entfernung erhält diese Funktion als zusätzlichen Parameter eine textuelle Angabe der relativen Position, z.B. `IPSO` oder `ORTHO`. Es werden nur diejenigen Atome zurückgeliefert, die sich in dem dadurch bezeichneten Zweig des Moleküls befinden.
- Ermitteln der korrekten Klassifikation einzelner Positionen. Ausgehend von einem CSInfo-Objekt, welches in die Repräsentation als Strukturelemente überführt wird, werden mit Hilfe der oben beschriebenen Funktionen die für die strukturelle Kategorisierung maßgeblichen Atome bestimmt. Da diese einzeln bzw. in ihrer Kombination auf die Zustandsbezeichnungen der entsprechenden Variablen abzubilden sind, gibt es jeweils spezialisierte Funktionen für die im einzelnen betrachteten Position (*ipso* und *ortho*).

7.3 Ergebnisse

Für einen reibungslosen Arbeitsablauf ist ein solides Fundament der Datenverarbeitung unerläßlich. Darüber hinaus sollten die entwickelten Funktionen und Datenstrukturen gut wiederzuverwenden und im besonderen im Zuge zukünftiger Weiterentwicklungen des Systems unproblematisch weiterzuverwenden sein. Daher wurde auf eine sorgfältige und vorausschauende Entwicklung des Vorverarbeitungsmoduls wertgelegt.

Innerhalb der Aufgaben des Vorverarbeitungsmoduls sind zwei Arten erforderlicher Funktionalitäten zu unterscheiden, die, wie in Abbildung 7.3 dargestellt, zwei unterschiedlichen Teilschritten der Vorverarbeitung zuzuordnen sind: einerseits das grundsätzliche Zugänglichmachen der in den Rohdaten enthaltenen Information und andererseits das Verfügbarmachen dieser Information in einer für die Sichtweise der gegebenen Problemstellung günstigen

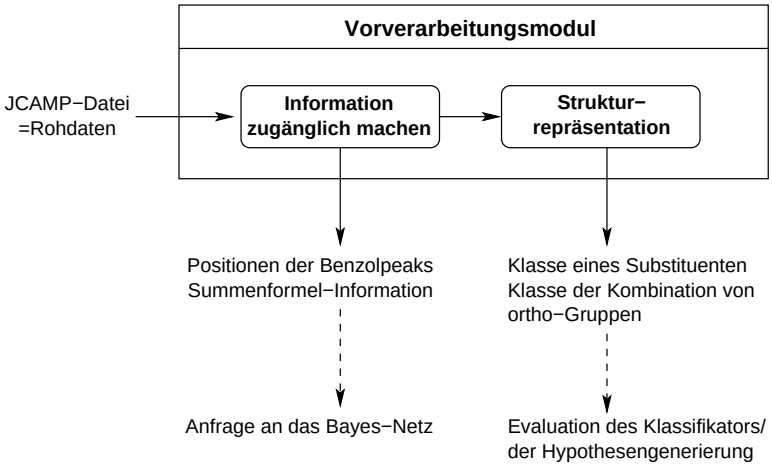


Abb. 7.3: *Verlauf der Vorverarbeitung auf Objektebene: JCAMP-Rohdaten werden zu Information raffiniert und in einem zweiten Schritt in eine für die Betrachtung von Molekülstrukturen besonders günstige Repräsentation überführt, die auf die spezielle Betrachtungsweise des gegenwärtigen Systems zugeschnitten ist.*

Form. Für beide Teilschritte wurden geeignete Objekttypen und Funktionen entwickelt. Sie reichen vom Lesen von JCAMP-Dateien und Erfassen der in bestimmten LDRs enthaltenen Information über das Feststellen der Lage der Benzolpeaks und das Gewinnen der Summenformelinformation mit anschließender Ausgabe in Gestalt einer durch das Bayes-Netz verarbeitbaren Anfrage bis hinzu strukturbezogenen Aufgaben. Hierunter fällt in erster Linie das Aufzählen aller Atome in einer bestimmten Entfernung und relativen Lage von einem Fokusatom des Benzolrings, worauf basierend die korrekte Klassifikation einzelner Positionen des Rings ermittelt werden kann. Bei der Evaluation kann dann das Klassifikationsresultat mit dieser Angabe verglichen werden.

Im Rahmen des Gesamtsystems SASCHA bildet das Vorverarbeitungsmodul die Basis aller weiteren Verarbeitungsschritte. Es liefert die Daten, die andere Module benötigen (Evidenzen für die Klassifikation oder strukturelle Merkmale für Evaluation und Parameteradaption). Somit kann nun die Entwicklung mit den für die Folgeschritte benötigten Modulen fortgesetzt werden.

Das Bayes-Netz wurde in Kapitel 6 bereits weitgehend entwickelt, lediglich die Gestaltung der bedingten Wahrscheinlichkeiten ist offen geblieben. Diese Parameter des Bayes-Netzes können grundsätzlich explizit vorgegeben, empirisch ermittelt oder gemäß wohlfundierter Theorien bestimmt werden. Das empirische Ermitteln der Wahrscheinlichkeiten kann als Lernen der Charakteristik einer durch die Stichprobe gegebenen Umgebung aufgefaßt werden und wird in Kapitel 8 näher beschrieben. Liegen die Wahrscheinlichkeiten vor, so kann das Bayes-Netz zur Klassifikation eingesetzt und dahingehend evaluiert werden. Außerdem soll im Rahmen des Gesamtsystems eine Weiterverarbeitung der Klassifikationsresultate durch die Hypothesengenerierung zu einem Substitutionsmuster erfolgen.

8 Parametrisierung und Lernen von Umgebungscharakteristika

Die elementaren Voraussetzungen zur automatischen Auswertung von ^{13}C-NMR-Spektren, wie sie SASCHA leisten soll, sind die grundsätzliche Fähigkeit zur Handhabung entsprechender Eingangsdaten und vor allem ein geeignetes Modell, welches die Kausalzusammenhänge zwischen strukturellen Merkmalen und spektroskopischen Befunden treffend wiedergibt. Diese beiden Aspekte wurden in den vorangegangenen Kapiteln bereits eingehend behandelt. Das in Gestalt des Kausalmodells eingebrachte qualitative Expertenwissen muß jedoch quantifiziert werden; hierzu dienen bedingte Wahrscheinlichkeiten, welche zusammen mit dem Modell in Gestalt eines Bayes-Netzes gespeichert werden. Erst dann können Berechnungen auf dieser Grundlage durchgeführt werden.

Es ist also ein Modul vonnöten, welches die bedingten Wahrscheinlichkeiten des Bayes-Netzes ermittelt. Geschieht dies durch Approximierung mittels relativer Häufigkeiten auf einer gegebenen Stichprobe, so kann dies als Lernen der Charakteristik einer bestimmten Einsatzumgebung aufgefaßt werden. Somit kann auch das im folgenden entwickelte Modul als Lernmodul verstanden werden, seine Funktionalitäten sind jedoch in dieser Hinsicht sehr rudimentär, so daß die Bezeichnung Statistikmodul treffender erscheint. Sie wird zudem auch der zweiten Aufgabe dieses Moduls gerecht: Basierend auf statistischen Untersuchungen einer Stichprobe wird die Gestaltung der Zustände einzelner Variablen des Kausalmodells, wie in Kapitel 6 bereits angedeutet, noch einmal näher betrachtet.

Abschnitt 8.1 beschreibt zunächst die Aufgaben des Statistikmoduls im Kontext des Gesamtsystems und betrachtet dabei auch die Voraussetzungen, welche vor der Aktivierung des Moduls gegeben sind und berücksichtigt werden müssen oder welche das Modul erfordert, ehe es aktiviert werden kann. Im Anschluß sind in Abschnitt 8.2 statistische Untersuchungen der Stichprobe vor dem Hintergrund der Zustandsgestaltung einzelner Modellvariablen Gegenstand der Betrachtung. Abschnitt 8.3 führt schließlich die Gewinnung der zur Vervollständigung des Bayes-Netzes benötigten bedingten Wahrscheinlichkeitsverteilungen näher aus. Die erhaltenen Ergebnisse werden in Abschnitt 8.4 zusammengefaßt.

8.1 Einbettung ins Gesamtsystem

Im Rahmen des Gesamtsystems hat das Statistikmodul insofern eine zentrale Rolle, daß es Berührungspunkte mit allen anderen Modulen besitzt. Auch seine Funktionalität ist von zentralem Interesse, obwohl es während der Auswertung eines Spektrums durch SASCHA inaktiv bleibt. Es kommt jedoch zum Einsatz, um einerseits die Wahrscheinlichkeiten und andererseits auch die Zustände der Variablen des Bayes-Netzes festzulegen, und ohne dies wäre keinerlei Auswertung möglich.

Seinerseits ist das Statistikmodul auf die Funktionalitäten des Vorverarbeitungsmoduls zur Handhabung der Daten der Stichprobe angewiesen. Darüber hinaus muß es bei der Parameteradaption mit dem bereits im BNIF-Format für Bayes-Netze (vgl. Abschnitt 5.2.3) niedergelegten Kausalmodell umgehen und in dieses die benötigten bedingten Wahrscheinlichkeiten einfügen.

Voraussetzungen bei der Aktivierung des Statistikmoduls ergeben sich unmittelbar aus diesen Zusammenhängen: Die Funktionen des Vorverarbeitungsmoduls, wie sie in Kapitel 7 (insbesondere Abschnitt 7.2.2) beschrieben wurden, sind als Ausgangsbasis gegeben. Sie betreffen die Untersuchung des Vorliegens bestimmter spektroskopischer und struktureller Merkmale. Ähnlich der dort beschriebenen Handhabung von JCAMP-Rohdaten ist hier nun zusätzlich der Umgang mit Bayes-Netz-Dateien im BNIF-Format relevant. Ein entsprechender Objekttyp und zugehörige Funktionen werden im folgenden beschrieben. Darüber hinaus werden theoretische Überlegungen zur Bestimmung relativer Häufigkeiten im allgemeinen dargestellt.

8.1.1 Lesen von BNIF-Dateien

Für die Zusammenarbeit mit dem Wissensmodul in Gestalt eines als BNIF-Datei gespeicherten Kausalnetzes müssen derartige Dateien gelesen, gezielt modifiziert und wieder gespeichert werden können. In diesem Zusammenhang ist zunächst ein geeigneter Objekttyp von Interesse. Orientiert man sich an den Vorgaben des Dateiformats, so läßt er sich verhältnismäßig leicht entwickeln: Ein BNIF-Objekt besitzt eine Bezeichnung (einen Namen) und besteht aus einer Liste von Eigenschaften, einer Liste von Variablen und einer Liste von Wahrscheinlichkeiten. Eine Eigenschaft ist eine einfache Zeichenkette, Variablen und Wahrscheinlichkeiten sind wiederum Objekte. Eine Variable besitzt einen Namen, eine Liste von Eigenschaften und eine Liste von möglichen Zuständen. Alle diese Bestandteile sind Zeichenketten. Eine Wahrscheinlichkeit besitzt eine Referenz auf die abhängige Variable und eine Liste von Referenzen auf die Variablen, von denen sie abhängt. Außerdem besitzt sie eine Liste von Eigenschaften sowie eine Tabelle von reellen Zahlen zwischen 0 und 1, welche die entsprechende Wahrscheinlichkeitsverteilung wiedergibt. Die genannten Objekte sind in Abbildung 8.1 dargestellt. Mit ihnen sind die folgenden Funktionen assoziiert.

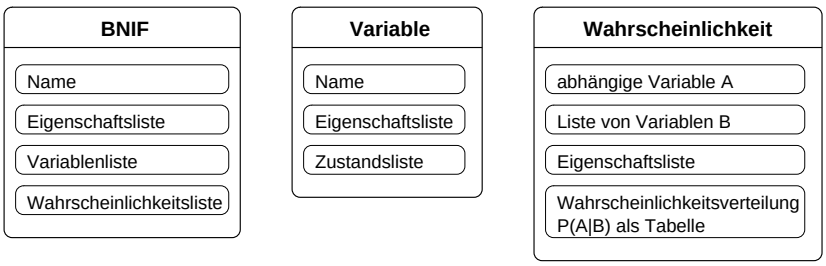


Abb. 8.1: *Objekte zur Repräsentation von Bayes-Netzen analog zum BNIF-Format.*

- Lesen einer BNIF-Datei. Die Funktion erhält den Dateinamen als Parameter und liefert ein BNIF-Objekt. Wird beim Lesen der Datei der Beginn einer Variablen- oder Wahrscheinlichkeits-Deklaration gefunden (vgl. Abschnitt 5.2.3, Schlüsselworte `variable` und `probability`), so wird der betreffende Ausschnitt an eine der folgenden beiden Funktionen weitergegeben, welche ihn zu einem entsprechenden Objekt verarbeiten, das dann dem BNIF-Objekt hinzugefügt wird.

- Überführen einer Variablen-Deklaration in ein Variablen-Objekt.

- Überführen einer Wahrscheinlichkeits-Deklaration in ein Wahrscheinlichkeits-Objekt.

- Ermitteln einer Variablenreferenz anhand des Namens. In Wahrscheinlichkeits-Deklarationen sind die beteiligten Variablen über ihre Namen referenziert. Die Funktion erhält ein BNIF-Objekt und den Variablennamen als Parameter und liefert ein Variablen-Objekt.
- Schreiben einer BNIF-Datei. Die Funktion erhält ein BNIF-Objekt als Parameter. Dessen Bestandteile (Eigenschaften, Variablen und Wahrscheinlichkeiten) werden in entsprechende Deklarationen umgewandelt und im BNIF-Format in eine Datei geschrieben. Dazu werden die folgenden Funktionen benutzt:
- Überführen eines Variablen-Objekts in eine Variablen-Deklaration.
- Überführen eines Wahrscheinlichkeits-Objekts in eine Wahrscheinlichkeits-Deklaration.

8.1.2 Relative Häufigkeiten

Hauptfunktionalität des Statistikmoduls ist es, *relative Häufigkeiten* zu ermitteln, mit dem Ziel, die für das Bayes-Netz benötigten *a-priori-* und bedingten Wahrscheinlichkeiten empirisch zu ermitteln. Als relative Häufigkeit bezeichnet man den Quotienten $\frac{n}{N}$ der Anzahl n von Beispielen, die eine bestimmte Eigenschaft besitzen, und der Gesamtzahl N der Beispiele einer Stichprobe. Sie kann nur Werte zwischen 0 und 1 annehmen und wird verwendet, um Wahrscheinlichkeiten im allgemeinen abzuschätzen, wie in Gleichung (8.1) dargestellt: P_A dient als Näherung der Wahrscheinlichkeit $P(A)$.

Auch bedingte Wahrscheinlichkeiten können auf diese Weise approximiert werden: Da sie einen Übergang in einen anderen Wahrscheinlichkeitsraum darstellen (vgl. Abschnitt 4.2.2, [Bos95] S. 31 ff.) wird zuerst die Grundmenge der Stichprobe angepaßt und dann die Zahl der Beispiele, welche die durch das abhängige Ereignis gegebene Eigenschaft besitzen, ermittelt. Durch den in Gleichung (8.2) bestimmten Wert P_{AB} kann die bedingte Wahrscheinlichkeit $P(A|B)$ angenähert werden. Dabei ist n_{AB} die Zahl der Beispiele *aus der reduzierten Stichprobe*, für welche Ereignis A zutrifft und n_B die Zahl der Beispiele *der betrachteten Grundmenge*, für welche Ereignis B zutrifft, also die Größe der reduzierten Stichprobe.

$$P_A = \frac{n_A}{N} \tag{8.1}$$

$$P_{AB} = \frac{n_{AB}}{n_B} \tag{8.2}$$

Auch zur empirischen Bestimmung der für Bayes-Netze benötigten Wahrscheinlichkeitsverteilungen lassen sich relative Häufigkeiten verwenden. Eine Wahrscheinlichkeitsverteilung $P(A) = (P(A = a_1), P(A = a_2), \ldots, P(A = a_k))^T$ läßt sich approximieren, indem die ihren Einzelkomponenten entsprechenden relativen Häufigkeiten P_{a_j} wie in Gleichung (8.3) ermittelt werden. Die *a-priori*-Wahrscheinlichkeiten sind Verteilungen dieser Gestalt.

$$P_{a_j} = \frac{n_{a_j}}{N} \tag{8.3}$$

$$P^i_{a_j} = \frac{n_{a_j}}{N_i} \tag{8.4}$$

In ähnlicher Weise können die bedingten Wahrscheinlichkeitsverteilungen über die Betrachtung ihrer einzelnen Komponenten approximiert werden. $P(A|B_1 \ldots B_m)$ ist eine Matrix, deren I einzelne Spalten unterschiedliche Verteilungen $P_i(A)$ enthalten. I ist das Produkt der

Zahl der Zustände von $B_1 \ldots B_m$, das heißt die Zahl möglicher Zustandskombinationen. Die Komponenten von $P_i(A)$ werden gemäß Gleichung (8.4) angenähert; dabei ist N_i die Zahl der für die i-te Spalte relevanten Beispiele, das heißt desjenigen Ausschnitts der Stichprobe, für den die Ereignisse $B_1 \ldots B_m$ jeweils in einer bestimmten Ausprägung vorliegen.

Welcher Zustand jeder Variablen (welche Ausprägung des betreffenden Ereignisses) für eine bestimmte Spalte relevant ist, hängt von der Zahl der Zustände jeder Variablen ab. Sind alle Variablen binär, so ergeben sich die Indizes 1 oder 0 der Zustände jeder Variablen aus der Binärschreibweise der Spaltennummer i. Sei z.B. $P(A|B_1,B_2,B_3,B_4)$ gesucht, und $Z(B_k)$ bezeichne den Index des Zustands, in welchem sich B_k befindet. Für $i = 5$ ist N_5 die Zahl aller Beispiele der Stichprobe, für die $Z(B_1) = 0$, $Z(B_2) = 1$, $Z(B_3) = 0$ und $Z(B_4) = 1$, denn die (vierstellige) Binärdarstellung von $i = 5$ lautet

$$(0101) = (Z(B_1), Z(B_2), Z(B_3), Z(B_4))$$

Die k-te Stelle der Binärdarstellung gibt also den Zustandsindex von B_k wieder. Verallgemeinert man dies auf Variablen mit einer beliebigen Zahl von Zuständen, so ist gewissermaßen eine Umrechnung der Spaltennummer i in ein Zahlensystem mit Stellen unterschiedlicher Wertigkeit erforderlich. Auch hier gibt dann wiederum die k-te Stelle den Index des Zustands der Variablen B_k an.

Essentiell für die Bestimmung jeglicher relativer Häufigkeiten ist es außerdem festzustellen, welchen spektralen oder strukturellen Merkmalen ein bestimmter Zustand einer Variablen entspricht. Erst dadurch werden die Eigenschaften festgelegt, die zur Feststellung eines Wertes P_{a_j} bzw. $P^i_{a_j}$ zu prüfen sind. Es wurde eine systematische Notation entwickelt, die es erlaubt, den Zusammenhang zwischen einem bestimmten Zustand und spektralen oder strukturellen Merkmalen in den Eigenschaften der betreffenden Variablen festzuhalten. Sie wird neben der grundsätzlichen Festlegung der Zustände der einzelnen Variablen im folgenden Abschnitt näher erläutert.

8.2 Betrachtete Ereignisse

Während der Entwicklung des Kausalmodells wie in Kapitel 6 beschrieben wurden die im Kontext der gegebenen Aufgabenstellung und der betrachteten Domäne interessanten Ereignisse ausfindig gemacht und in diskrete Zufallsvariablen gefaßt. Wenngleich die Definition der Ereignisse, welche die einzelnen Variablen modellieren, leicht zu geben war, erforderte es einige zusätzliche Überlegungen, die Ausprägungen, in welchen ein Ereignis jeweils auftreten kann, in Gestalt disjunkter Zustände erschöpfend zu erfassen.

Dies gilt vor allem für die auf die Struktur des Moleküls bezogenen Variablen `s2ipso` und `s2ortho`. Die Kombination von zwei Atomen im Falle von `s2ortho` führt hier wie in Abschnitt 6.2.1 dargestellt zu 210 theoretisch möglichen Ausprägungen, für `s2ipso` sind es sogar über 11000. Das Vorgehen im Fall der dritten strukturbezogenen Variablen `C_arom` ist dagegen unproblematisch, da hier in der betreffenden Position nur einzelne Atome zu betrachten sind und keine Kombinationen derselben.

Jede der möglichen Atomkombinationen von `s2ipso` und `s2ortho` ist eine individuelle Ausprägung des jeweils betrachteten Ereignisses; sie sind im mathematischen Sinne als Elementarereignisse in dem durch die Betrachtung einer bestimmten Position des Moleküls aufgespannten Wahrscheinlichkeitsraum zu sehen. Somit würde eine Realisierung mit einem Zustand je möglicher Atomkombination sowohl der Forderung einer erschöpfenden Darstellung als auch der paarweisen Disjunktheit der einzelnen Zustände gerecht. Aufgrund der Vielzahl der Möglichkeiten ist jedoch eine Zusammenfassung anzustreben. Dem widmet sich der

folgende Abschnitt. Außerdem ist eine Beschreibung der zu den Zuständen im einzelnen korrespondierenden Atomkombinationen von Interesse, um die bedingten Wahrscheinlichkeiten empirisch ermitteln zu können, in welchen sie eine Rolle spielen.

Auch die Festlegung der Zustände derjenigen Variablen, die die einzelnen Inkremente bzw. Teilinkremente modellieren, war nicht ohne einige Vorüberlegungen vorzunehmen. Die grundsätzliche Entscheidung für eine auf 1,0 ppm genaue Wiedergabe wurde in Kapitel 6 bereits motiviert, dort wurde jedoch ebenso aufgezeigt, daß sich die Frage stellt, ob die Extreme der einzelnen Inkrementwerte tatsächlich erreicht werden.

Die Festlegung geeigneter Intervallgrenzen kann gleichwohl in einem Schritt mit der empirischen Gewinnung der bedingten Wahrscheinlichkeiten, wie in Abschnitt 8.3.1 beschrieben, geschehen: Finden sich in der so aufgestellten Tabelle Zeilen, in denen kein Wert $\neq 0,0$ ist, so bedeutet dies, daß der betreffende Zustand der abhängigen Variablen nie angenommen wird. Trifft dies auf Zustände an den Grenzen des repräsentierten Intervalls zu, so kann das Intervall durch die Festlegung von Schwellwerten, die einen Ausschluß der betreffenden nicht benötigten Zustände zur Folge haben, verkleinert werden.

8.2.1 Zustände der strukturbezogenen Variablen

Die Abbildungen 8.2 und 8.3 stellen die Ereignisse dar, die durch die Variablen `s2ipso` und `s2ortho` modelliert werden. Im Falle von `s2ipso` (Abbildung 8.2) sei darauf hingewiesen, daß die Zahl der dabei zu betrachtenden Atome von der Valenz des Nachbaratoms in der ersten Sphäre abhängt. Somit können weniger, aber auch mehr als die in der Abbildung markierten drei Atome angetroffen werden.

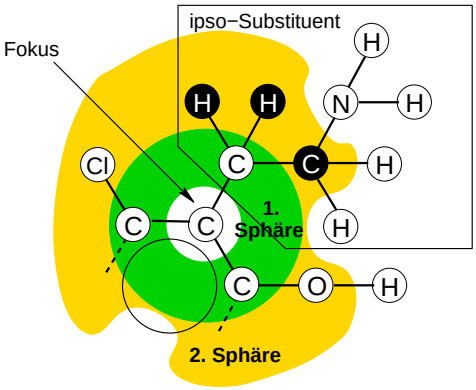


Abb. 8.2: *Das durch die Variable* `s2ipso` *modellierte Ereignis ist die Belegung der Atome in der hervorgehobenen Position.*

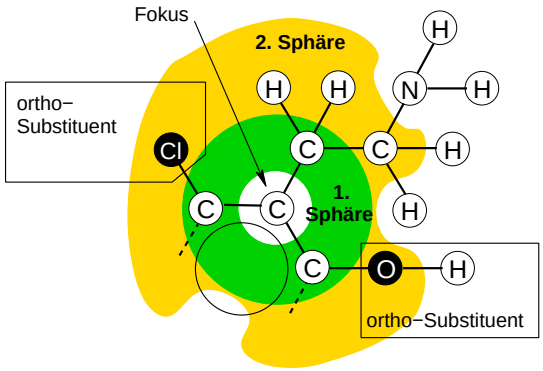


Abb. 8.3: *Durch die Variable* `s2ortho` *wird die Kombination der beiden hervorgehobenen Atome modelliert.*

Das Vorverarbeitungsmodul stellt Funktionalitäten zur Verfügung, um festzustellen, welche Atome in einer gegebenen Entfernung und relativen Position von einem bestimmten Fokusatom vorliegen. Darauf aufbauend definiert das Statistikmodul eine Funktion, welche ein Histogramm über die in einer bestimmten relativen Position und Entfernung vorkommenden Atome erstellt. Dabei werden zunächst für alle Beispielverbindungen die sechs Benzol-Kohlenstoffatome identifiziert. Jedes davon fungiert nacheinander als Fokusatom im Sinne von Abbildung 8.2 bzw. 8.3, so daß die Stichprobe sechsmal so viele Elemente umfaßt, wie Beispielverbindungen gegeben sind. Für jedes Fokusatom werden die Atome in der gewünschten relativen Position und Entfernung identifiziert. Aus der so gefundenen Atommenge wird ei-

ne systematische Bezeichnung generiert, welche die beteiligten Atomspezies beschreibt. Für jeden so generierten Schlüssel wird die Zahl der Vorkommen gezählt.

Für diese Bezeichnung werden die üblichen Symbole der chemischen Elemente, C, H, O, N, S usw. verwendet. Werden von einem Element mehrere Varianten unterschieden, z.B. unterschiedlich hybridisierte Kohlenstoffatome, so wird dem Elementsymbol durch einen Unterstrich verbunden eine genauere Bezeichnung nachgestellt. Die unterschiedenen Atomspezies erhalten die Priorität

```
H>C_sp3>C_sp2>C_sp>O>N_amin>N_imin>N_5>S_2>S_6>F>Cl>Br>I.      (8.5)
```

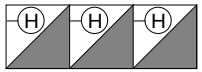
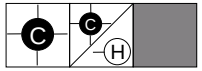
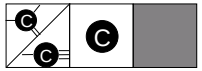
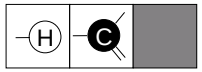
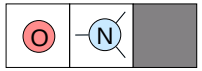
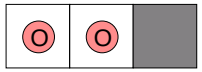
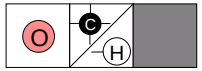
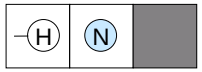
Dies genügt für eine Systematik betreffend die Betrachtung von `s2ortho`. Für `s2ipso` wird eine zusätzliche Möglichkeit in die Ordnung eingereiht, und zwar `N_nitril` für Stickstoff mit einer Dreifachbindung zwischen `N_imin` (Stickstoff mit einer Doppelbindung) und `N_5` (Stickstoff in der Gestalt der NO_2-Gruppe, mit der scheinbaren Valenz 5). Die Systematik ist eine durch Kommata getrennte Aneinanderreihung der vorgefundenen Atomtypen gemäß ihrer Priorität. Beinhaltet also z.B. eine Atomkombination Wasserstoffatome, so werden diese immer vor allen anderen Atomen aufgezählt. Die Kombination eines Sauerstoff- und eines sp^2-hybridisierten Kohlenstoffatoms wird z.B. `C_sp2,O` geschrieben.

Innerhalb des Statistikmoduls sind dazu die folgenden Funktionen definiert:

- Generieren einer systematischen Beschreibung der in einer bestimmten Position vorgefundenen Atomkombination.
- Erstellen eines Histogramms der in einer bestimmten Position vorkommenden Atomkombinationen. Die Funktion erhält die gewünschte relative Lage (*ipso* oder *ortho* und Distanz) als Parameter, generiert für jedes Beispiel der Stichprobe die systematische Beschreibung der dort vorgefundenen Atomkombination als Schlüssel und zählt die Vorkommen jedes generierten Schlüssels.
- Auswahl derjenigen Teilmenge der Beispiele, die in einer bestimmten Position eine bestimmte Atomkombination aufweisen. Die Funktion erhält die gewünschte relative Lage sowie die dort erlaubten möglichen alternativen Typen als Parameter.

Die folgenden Tabellen 8.4 und 8.5 beschreiben die Ergebnisse der Analyse für `s2ipso` und `s2ortho`. Weist eine einzelne Atomkombination etwa 15 bis 20 Vorkommen im Falle von `s2ipso` bzw. von 20 bis 25 Vorkommen im Falle von `s2ortho` in der Stichprobe (5418 Substituenten) auf, so kommt sie für einen eigenen Zustand in Frage. Anderenfalls werden strukturell ähnliche Kombinationen untereinander oder mit ähnlichen bereits für sich ausreichend häufigen Kombinationen zusammengefaßt. Es wird dabei beachtet, daß diejenigen Atomkombinationen auf jeden Fall einen eigenen Zustand erhalten sollen, welche mit den „Standardgruppen“ assoziiert sind (vgl. S. 90); dies ist z.B. bei dem `s2ipso`-Zustand `C_sp` der Fall (Standardgruppe –C≡C-H). Darüber hinaus ist auf eine besondere chemische Charakteristik der Gruppen zu achten.

Bei einer solchen Zusammenfassung von Elementarereignissen ist es zudem wichtig zu dokumentieren, welche einzelnen Atomkombinationen sich hinter dem gemeinsamen Zustand verbergen: Zum einen muß für den Entwickler transparent gehalten werden, welche Aussage hinter einem bestimmten Zustand einer bestimmten Variablen steht, und zum anderen muß für die empirische Ermittlung der Wahrscheinlichkeiten zur Quantifizierung der Kausalverknüpfungen auch für ein automatisches Verfahren feststellbar sein, welche Struktureigenschaften in den Beispielen der Stichprobe zu überprüfen sind.

Zustand	Vorkommen	Kommentar
	3907	Standardgruppen, die ausschließlich H−Atome in der 2. Sphäre tragen (395) sowie einatomige Substituenten (3512)
	396	ein gesättigter Kohlenwasserstoffrest
	147	ungesättigter Kohlenwasserstoffrest; kann auch über die Doppelbindung an N_imin gebunden auftreten.
	11	Standardgruppe Alkine
	19	Nitrile. Da dies die einzige Alternative zu C_sp für den Fall von C_sp in der 1. Sphäre ist, gibt es keine sinnvolle Möglichkeit der Zusammenfassung
	67	Zwei gesättigte Kohlenwasserstoffreste oder ein solcher Rest und ein Wasserstoffatom.
	20	Verzweigung mit zwei Kohlenwasserstoffresten, von denen mindestens einer ungesättigt ist.
	172	kann in Kombination mit C_sp2 oder N_imin in der ersten Sphäre vorkommen
	17	Amide
	175	Carbonsäuren und Ester
	87	vor allem Aldehyde und Ketone
	28	vor allem Imine, jedoch theoretisch auch die Kombinationen H,N_5 und H,N_amin

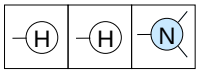
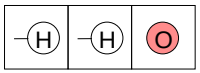
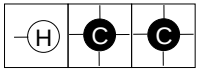
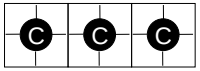
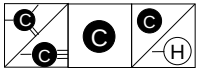
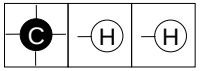
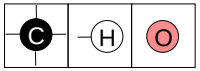
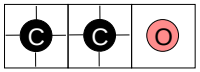
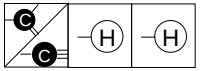
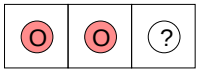
Zustand	Vorkommen	Kommentar
	28	Aminogruppe an C_sp3
	21	Ether oder Alkoholgruppe an C_sp3
	30	einfache Verzweigung an C_sp3 mit zwei gesättigten Kohlenwasserstoffresten
	43	tertiärer Butylrest oder ähnlich stark verzweigte Kohlenwasserstoffreste
	18	Verzweigung mit zwei oder drei Kohlenwasserstoffresten und mindestens einem ungesättigten Anteil.
	72	unverzweigte gesättigte Kohlenwasserstoffe
	22	Ether oder Alkohol zusammen mit einem gesättigten Kohlenwasserstoffrest an C_sp3
	10	Alkohol oder Ether in verzweigtem
	50	unverzweigter ungesättigter Kohlenwasserstoffrest
	12	Standardgruppe Sulfonate sowie alle weiteren Kombinationen mit zwei Sauerstoff− und einem weiteren Atom.
	18	Standardgruppe Imine
andere		alle übrigen Fälle

Abb. 8.4: *Nach statistischer Analyse einer Stichprobe von 5418 Substituenten in 903 Verbindungen wurden die Zustände von* `s2ipso` *festgelegt.*

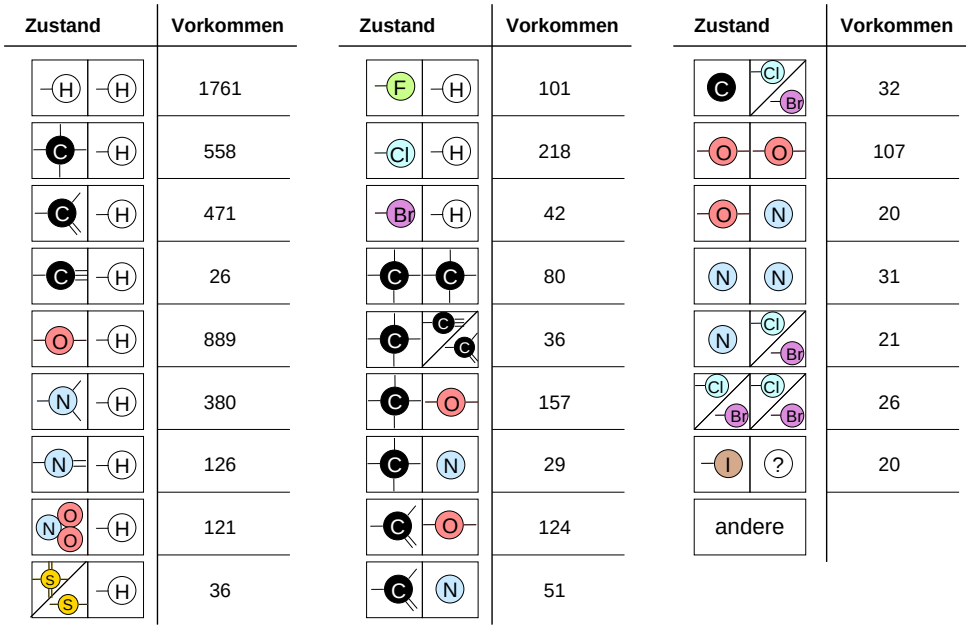


Abb. 8.5: *Nach statistischer Analyse einer Stichprobe von 5418 Substituenten in 903 Verbindungen wurden die Zustände von* `s2ortho` *festgelegt.*

Diese beiden Ansprüche der Lesbarkeit für den Menschen wie auch der maschinellen Verarbeitbarkeit erwiesen sich jedoch im Verlauf der Entwicklung als schwerlich mit einander vereinbar. Auf der einen Seite sind etwa die auf der in Aufzählung (8.5) beschriebenen Hierarchie basierenden systematischen Bezeichnungen zur Wiedergabe der einzelnen Atomkombinationen sowohl für den Menschen verständlich als auch für die automatische Verarbeitung geeignet, auf der anderen Seite sind sie jedoch gerade im Fall der Zusammenfassung mehrerer Elementarereignisse zu einem Zustand nicht anwendbar. Somit ist es nicht ohne weiteres möglich, direkt aus der Benennung des Zustands auf die strukturellen Eigenschaften zu schließen, für welche er steht.

Für die maschinelle Verarbeitung kann man sich sehr gut der Möglichkeit des BNIF-Dateiformats bedienen, frei Eigenschaften von Variablen zu definieren: Es wird ein Typ von Eigenschaft eingeführt, welcher eine Beschreibung der zu jedem Zustand korrespondierenden Elementarereignisse enthält. Dieses Prinzip kann nicht nur verwendet werden, um im Falle zusammengefaßter Elementarereignisse zu dokumentieren, welche Zusammenfassung vorgenommen wurde, sondern kann grundsätzlich zur Anwendung kommen, um diejenigen Merkmale zu beschreiben, auf welche hin die Stichprobe bei der Betrachtung eines bestimmten Zustands zu untersuchen ist. Diese Form der Beschreibung wird im folgenden Abschnitt vorgestellt. Durch ein solches Vorgehen können alle Zustandsbezeichner für den Menschen möglichst informativ gewählt werden, was auch bei der Entwicklung und Weiterentwicklung des Modells von Vorteil ist.

8.2.2 Beschreibung von Merkmalen

Um innerhalb von Variablenbeschreibungen Eigenschaften kenntlich zu machen, welche wiedergeben, für welches Ereignis die betreffende Variable und für welche Ausprägung ein bestimmter Zustand steht, wird das Schlüsselwort `CSINFO` (JCAMP-*CS Inf*ormation) benutzt.

Es hat seinen Ursprung darin, daß die Beschreibung sich im Zuge der Definition der Zustände der strukturbezogenen Variablen entwickelte. Dasselbe Schlüsselwort wird nun jedoch für alle, auch nicht-strukturbezogene, Variablen benutzt.

Innerhalb der Eigenschaftsliste einer Variablen wird das Schlüsselwort `CSINFO` zumindest einmal erwartet. Die betreffende Eigenschaft kann zum einen `CSINFO=EXPERT` lauten, wie etwa für die Variable `Sphäre1` der Fall. Dies bedeutet, daß die dieser Variablen zugehörige Wahrscheinlichkeitsverteilung nicht empirisch ermittelt werden soll, sondern durch Expertenwissen festgelegt wird. Für das Statistikmodul ist diese Variable also uninteressant, und es treten keine weiteren Eigenschaften mit dem Schlüsselwort `CSINFO` auf.

Im zweiten Fall wird `CSINFO=` von dem Wort `Atomlist` und einem weiteren Begriff gefolgt. Dieser Begriff gibt an, welche Eigenschaft innerhalb der Atomliste zu prüfen ist. Dabei kann es sich um das chemische Element (`Elem`) oder die chemische Verschiebung (`Shift`) der Atome handeln. In diesem Fall folgen weitere Eigenschaften, in welchen dem Schlüsselwort `CSINFO` durch einen Unterstrich verbunden der Index des beschriebenen Zustands folgt. Durch ein Gleichheitszeichen verbunden ist anschließend eine Liste von Parametern gegeben, welche angibt, wohingehend die gewählte Eigenschaft innerhalb der Atomliste überprüft werden soll, um festzustellen, ob ein gegebenes Beispiel ein Repräsentant des durch den Index referenzierten Zustandes ist. Zwei Beispiele seien zur Verdeutlichung gegeben:

1. Für die Variable `Stickstoff` lautet eine Eigenschaft `CSINFO=Atomlist Elem`. Die darauffolgende Eigenschaft `CSINFO_0=N FALSE` gibt an, daß der Zustand mit dem Index 0 für das Nichtvorhandensein (`FALSE`) von Stickstoff (Elementsymbol `N`) steht.

2. Für die Variable `peakpos` lautet eine Eigenschaft `CSINFO=Atomlist Shift`. Die Zustände Nummer 0 und 77 sind wie folgt beschrieben:
   `CSINFO_0=94.0 SMALLER`
   `CSINFO_77=173.9 GREATER`
   Das bedeutet, eine chemische Verschiebung mit einem kleineren (`SMALLER`) Wert als 94.0 ppm wird auf Zustand 0, Werte größer (`GREATER`) als 173.9 ppm werden auf Zustand 77 abgebildet. Die übrigen Zustände mit den Nummern $i = 1 \ldots 76$ sind in der Form `CSINFO_`i`=`x_i `INTERVAL` y_i beschrieben: Werte in dem durch x_i und y_i gegebenen Intervall werden auf den Zustand mit dem Index i abgebildet.

Die dritte mögliche Form, die die erste Eigenschaft mit dem Schlüsselwort `CSINFO` haben kann, ist eine Folge strukturbezogener Angaben. Hier werden die Begriffe `NEIGHBOUR`, `IPSO`, `ORTHO` und `BENZOL` benutzt. Für die Variable `s2ortho` beispielsweise ist das durch sie modellierte Ereignis mit `CSINFO=NEIGHBOUR NEIGHBOUR ORTHO` umschrieben. Zu betrachten sind also die Nachbarn der Nachbarn (`NEIGHBOUR NEIGHBOUR`) von Benzolringatomen, jedoch nur diejenigen, die sich in der *ortho*-Position befinden. `ORTHO` restringiert dabei nicht nur den zu betrachtenden Molekülausschnitt auf die *ortho*-Position, sondern stellt zugleich den Bezug zu einem Benzolringatom als Fokuspunkt her.

Die Beschreibung der einzelnen Zustände erfolgt wiederum mittels `CSINFO_`i`=` unter Angabe einer Parameterliste. Diese besteht aus Atomtypen oder Listen von Atomtypen alternierend mit Zahlen. Auch hier sollen Beispiele zur Verdeutlichung dienen:

1. Die Eigenschaft `CSINFO_0=H 2` beschreibt das Vorhandensein von 2 Wasserstoffatomen (`H`). In derselben Weise können alle übrigen in der Hierarchie in (8.5) beschriebenen Typen von Atomen sowie zusätzlich die chemischen Elementsymbole derjenigen Elemente wie C oder S, von welchen mehrere Typen unterschieden werden, verwendet werden.

2. Es können natürlich mehrere derartige Angaben gemacht werden. Beispielsweise steht `CSINFO_1=H 1 C_sp3 1` für das Vorhandensein eines Wasserstoff- und eines sp^3-hybridisierten Kohlenstoffatoms.

3. Wird der Beschreibung des Atomtyps ein Komma nachgestellt, so ist es Teil einer Aufzählung alternativer erlaubter Typen. Die Liste der Alternativen wird in Klammern eingefaßt. `CSINFO_20=(Cl,Br,) 2` steht für das Vorhandensein von insgesamt zwei Chlor- und Bromatomen.

Der Angabe der Anzahl entsprechender Atome kann außerdem ein Fragezeichen vorangestellt werden. ?*n* ist als „höchstens *n*“ zu verstehen. Außerdem wird die Anzahl entsprechender Atome für die Überprüfung der nächsten Bedingung übernommen und zu den dort gefundenen Atomen hinzugezählt, so daß die Angaben `(H,) ?1 (C_sp3,) 1` und `(H,C_sp3,) 1` äquivalent sind. Als weiteres Beispiel lautet die Beschreibung des 17. Zustandes von `s2ipso`
`CSINFO_17= (H,) ?1 (C_sp3,) ?2 (C_sp2,C_sp,) 3 (F,Cl,Br,N,O,S,) 0`
Dies besagt: *Es darf höchstens ein Wasserstoffatom vorkommen. Zählt man die gefundenen sp^3-hybridisierten Kohlenstoffatome hinzu, darf die Anzahl von beidem zusammen höchstens 2 betragen. Zählt man nun die vorhandenen sp^2- und sp-hybridisierten Kohlenstoffatome hinzu, so müssen insgesamt 3 Atome gefunden worden sein. Atome der Elemente F, Cl, Br, N, O und S dürfen nicht vorkommen.*

Schließlich gibt es noch die Möglichkeit, anstelle einer Parameterliste das Schlüsselwort `OTHER` anzugeben. Alle Beispiele, die nicht positiv für einen der bis dahin untersuchten Ausprägungen des betrachteten Ereignisses zu zählen sind, werden als positiv für die mit `OTHER` gekennzeichnete Ausprägung gezählt. Da die Ausprägungen eines Ereignisses in Gestalt der Zustände einer Variablen disjunkt sowie in ihrer Gesamtheit erschöpfend sein müssen ist dies sinnvoll, jedoch nur für den jeweils letzten Zustand in der Zustandsliste einer Variablen zu verwenden.

Mithilfe dieser Beschreibungen können nun die Dateien der Stichprobe zur empirischen Gewinnung von Wahrscheinlichkeiten genutzt werden. Zunächst wird das Bayes-Netz im BNIF-Format gelesen, dann werden für jede empirisch zu ermittelnde Wahrscheinlichkeitstabelle relevanten Variablen festgestellt. Jede Position der Tabelle wird einzeln gemäß Gleichung (8.4) (vgl. Abschnitt 8.1.2) durch einen Wert $P^i_{a_j}$ angenähert. Die benötigten Zahlenwerte N_i und n_{a_j} werden mit Hilfe der in Abschnitt 7.2.1 beschriebenen Funktionen des Vorverarbeitungsmoduls ermittelt, wobei die `CSINFO`-Merkmalsbeschreibungen über die Auswahl der richtigen Funktion und deren Parametrisierung bestimmen. Für die Untersuchung der Stichprobe stehen zusammengefaßt die folgenden Funktionen zur Verfügung:

- Erstellen einer Zuordnung von Variablennamen zu Ereignissen.
- Feststellen der zur Überprüfung des Vorliegens einer bestimmten Ausprägung benötigten Parameter aus den durch `CSINFO_`*i* gekennzeichneten Eigenschaften.
- Prüfen, ob ein Beispiel ein positives Beispiel ist.
- Bestimmen absoluter Häufigkeiten. Die Funktion erhält die Beschreibung des Ereignisses aus der Eigenschaft `CSINFO=...`, die Parameterliste einer Ausprägung *i* entsprechend `CSINFO_`*i*`=...` sowie eine Stichprobe und liefert die Anzahl positiver Beispiele.
- Entfernen negativer Beispiele aus der Stichprobe. Mithilfe derselben Angaben wie für die Bestimmung absoluter Häufigkeiten liefert die Funktion eine Kopie der Stichprobe, aus welcher alle negativen Beispiele entfernt wurden. Dies ist bei der Bestimmung aller $P^i_{a_j}$ einer Spalte *i* hilfreich.

8.3 Quantifizierung der Kausalstruktur

Sind die Zustände aller Variablen festgelegt, so sind mit den beschriebenen Funktionen zur Prüfung der mit einem Zustand assoziierten spektralen oder strukturellen Eigenschaften alle Voraussetzungen erfüllt, um bedingte sowie *a-priori*-Wahrscheinlichkeiten des Bayes-Netzes mit Hilfe relativer Häufigkeiten zu approximieren und auf diese Weise die Charakteristik einer gegebenen Umgebung zu lernen.

Betrachtet man jedoch die im einzelnen beteiligten Variablen, so stellt man fest, daß nicht alle der bedingten Wahrscheinlichkeiten geeignet sind, um empirisch bestimmt zu werden. Im besonderen sind die Teilinkremente `Sphäre1`, `i2ipso` und `i2ortho` problematisch, da sie den gegebenen JCAMP-Daten nicht direkt entnommen werden können: Dort ist nur die resultierende Peakposition aufgeführt, in die jedoch alle drei Inkremente sowie zusätzlich nicht explizit modellierten Einflüsse, z.B. durch entferntere Atome, eingehen.

Die Verteilung $P(\texttt{Sphäre1}|\texttt{C\_arom})$ für das Teilinkrement der ersten Sphäre kann jedoch ohne große Schwierigkeiten mit Hilfe von Expertenwissen ermittelt werden, indem man auf Standardgruppen (vgl. Seite 90) zurückgreift. Diese wurden zur Definition der Zustände von `Sphäre1` herangezogen, so daß zu jedem Zustand von `C_arom` ein korrespondierendes Inkrement der entsprechenden Standardgruppe unter den Zuständen von `Sphäre1` zu finden ist. Sei i das zur Standardgruppe einer Klasse c der ersten Sphäre korrespondierende Inkrement, dann gilt:

$$P(\texttt{Sphäre1}|\texttt{C\_arom}) = \begin{cases} 1 \text{ wenn } \texttt{Sphäre1} = i \wedge \texttt{C\_arom} = c \\ 0 \text{ sonst} \end{cases} \tag{8.6}$$

Bei den Verteilungen $P(\texttt{i2ipso}|\texttt{s2ipso})$ und $P(\texttt{i2ortho}|\texttt{s2ortho})$ ist dasselbe Vorgehen nicht ohne weiteres möglich, da sich, wie im Zusammenhang des Inkrementansatzes in Abschnitt 6.1.3 erwähnt, in den Verteilungen das Zusammenwirken der Einflüsse unterschiedlicher Atome widerspiegeln soll. Auch bei Betrachtung ein und derselben Gruppe kann zudem, je nach der Gestalt ihrer Umgebung im Molekül, durch sterische Einflüsse oder andere Wechselwirkungen die effektive Beeinflussung der chemischen Verschiebung variieren.

Das Vorgehen zur Gewinnung dieser beiden bedingten Wahrscheinlichkeiten wird in Abschnitt 8.3.1 vorgestellt. In der Folge sind auch bedingte Wahrscheinlichkeiten, die die genannten Teilinkremente als Bedingungen enthalten, nicht empirisch zu bestimmen. Dies sind im einzelnen $P(\texttt{Sphäre2}|\texttt{s2ipso},\texttt{s2ortho})$ und $P(\texttt{peakpos}|\texttt{Sphäre1},\texttt{Sphäre2})$. Für diese Fälle ist ein anderes Konzept vorgesehen, nämlich wie in Abschnitt 6.2.2 angedeutet eine „weiche Summenbildung“. Näheres ist Abschnitt 8.3.3 zu entnehmen. Abschnitt 8.3.2 widmet sich dem auf relativen Häufigkeiten basierenden Verfahren, das für die strukturbezogenen Variablen `C_arom`, `s2ipso` und `s2ortho` zur Anwendung kommt.

8.3.1 Teilinkremente der zweiten Sphäre

Bei der Gewinnung der Wahrscheinlichkeiten $P(\texttt{i2ipso}|\texttt{s2ipso})$ und $P(\texttt{i2ortho}|\texttt{s2ortho})$ besteht das Problem, daß zwar die Überprüfung der zu den Zuständen der strukturbezogenen Variablen `s2ipso` und `s2ortho` korrespondierenden Merkmale automatisch überprüft werden können, nicht jedoch die zu `i2ipso` und `i2ortho` korrespondierenden Teilinkremente.

Ein grundsätzlicher Ansatz zu seiner Lösung besteht darin, sich auf solche Verbindungen zurückzuziehen, für welche die Anteile aller Einflüsse außer dem der aktuell betrachteten Position bekannt sind. Für $P(\texttt{i2ipso}|\texttt{s2ipso})$ kommen hier alle monosubstituierten Benzolderivate in Frage: Die insgesamt beobachtete chemische Verschiebung setzt sich gemäß Gleichung (6.2) aus der Grundverschiebung von 128,5 ppm, dem Beitrag der ersten und dem

Beitrag der zweiten Sphäre zusammen; letzterer ist wiederum die Summe der Anteile der *ipso*- und der *ortho*-Position. *Per definitionem* ist das zu einem Wasserstoffatom korrespondierende Inkrement stets 0, also ist bei monosubstituierten Benzolderivaten der Anteil der *ortho*-Position am Inkrement der zweiten Sphäre 0. Demnach setzt sich die chemische Verschiebung s_{mono} des *ipso*-Kohlenstoffs wie in Gleichung (8.7) beschrieben zusammen:

$$s_{\text{mono}} = 128,5ppm + s_1(\text{Standardgruppe}) + s_{2\text{ipso}} \tag{8.7}$$

Das Inkrement s_1 der ersten Sphäre ist dabei eine von der Gestalt der ersten Sphäre abhängige Konstante und kann für ein gegebenes Beispiel aus der Struktur des *ipso*-Substituenten hergeleitet werden. Damit ist der anteilige Einfluß $s_{2\text{ipso}}$ der Atome der zweiten Sphäre in der *ipso*-Position als Differenz der gegebenen chemischen Verschiebung und der bekannten Bestandteile $(128,5ppm + s_1(\text{Standardgruppe}))$ berechenbar.

Der Literatur (z.B. [Ewi79]) sind interpretierte Spektren monosubstituierter Benzolderivate zu entnehmen. Betreffend das gewünschte Teilinkrement der zweiten Sphäre sind jeweils all diejenigen Spektren gemeinsam zu betrachten, die denselben Atomtyp in der ersten Sphäre des Substituenten aufweisen (z.B. ein Stickstoffatom). Innerhalb dieser Menge sind, in Analogie zu der in Abbildung 6.5 auf Seite 89 dargestellten Kategorisierung von Substituenten, zwei Fälle zu unterscheiden: Handelt es sich um die Standardgruppe (im Beispiel $–NH_2$), so wird aus der zugehörigen chemischen Verschiebung das Inkrement $s_1(\text{Standardgruppe})$ berechnet. Handelt es sich nicht um die Standardgruppe (z.B. $–NH–CH_3$), so wird die Klasse der *s2ipso*-Atome (im Beispiel `H,C_sp3`) bestimmt und deren zugehöriges Teilinkrement $s_{2\text{ipso}}$ gemäß Gleichung (8.7) ermittelt.

Die so erhaltenen Teilinkremente werden für die Bestimmung der benötigten bedingten Wahrscheinlichkeit $P(\texttt{i2ipso}|\texttt{s2ipso})$ in Parameterdateien gespeichert. Diese können für ein und dieselbe Atomkombination (`s2ipso`-Zustand) potentiell mehrere Inkremente (`i2ipso` Zustände) enthalten. In dieser Varianz spiegelt sich implizit das Zusammenwirken der Einflüsse der *ipso*-Atome entfernterer Sphären mit denen der zweiten Sphäre wider.

Es wird nun eine Funktion definiert, welche derartige Parameterdateien liest und verarbeitet, um die bedingte Wahrscheinlichkeit $P(\texttt{i2ipso}|\texttt{s2ipso})$ zu ermitteln. Dabei wird zunächst die gesamte Wahrscheinlichkeitstabelle mit 0 initialisiert. Dann wird aus der Parameterdatei die Bezeichnung des Zustandes der Variablen `s2ipso` und damit die zu betrachtende Spalte i bestimmt. Der Bezeichnung ist eine Zahlenliste zugeordnet; die einzelnen Zahlen sind als Zustände von `i2ipso` zu verstehen. Zu jedem Zustand korrespondiert eine Zeile j innerhalb von $P(\texttt{i2ipso}|\texttt{s2ipso})$.

Nun gibt es zwei mögliche Vorgehensweisen: Entweder nur die betreffende Position (i, j) in der Tabelle wird bearbeitet, oder auch die benachbarten Einträge werden beeinflußt. Im ersteren Fall findet eine harte Initialisierung statt, bei welcher nur der (i, j)-te Eintrag um 1,0 erhöht wird. Im zweiten Fall einer weichen Initialisierung wird das Gesamtgewicht von $1,0$ auf den Eintrag selbst und die Einträge in seiner Nachbarschaft innerhalb der i-ten Spalte aufgeteilt. Dadurch kommt eine Glättung der Gesamtverteilung zustande.

Das für den Zusammenhang zwischen dem *ipso*-Teilinkrement und den *ipso*-Atomen der zweiten Sphäre benutzte und in den Parameterdateien niedergelegte Expertenwissen wird darüber hinaus auch für den Zusammenhang des *ortho*-Teilinkrements mit den *ortho*-Atomen der zweiten Sphäre herangezogen. Wie im folgenden beschrieben wird so das nicht streng additive Zusammenwirken der betreffenden Einflüsse erfaßt.

Grundsätzlich ähnelt der Ansatz dem für den *ipso*-Anteil: Es wird ein Szenario betrachtet, in welchem alle Teilinkremente mit Ausnahme des gesuchten bekannt sind. Es werden für den *ortho*-Anteil jedoch nicht in der Literatur erfaßte interpretierte Spektren betrachtet, sondern die gegebene Stichprobe wird untersucht. Sie wird zunächst nach dem Typ des Atoms in

der ersten Sphäre der *ipso*-Position in Teilmengen untergliedert und jede Teilmenge einzeln durchlaufen. Dabei steht für jede Teilmenge das Inkrement s_1 der ersten Sphäre für die betreffende Klasse fest, und das Inkrement der zweiten Sphäre kann als $s_2 = 128,5 - s_1$ berechnet werden.

Das Inkrement s_2 ist wiederum aus dem *ipso*- und dem *ortho*-Anteil zusammengesetzt. Die zu einzelnen Atomkombinationen korrespondierenden *ipso*-Anteile sind jedoch bereits in den oben beschriebenen Parameterdateien gegeben, so daß für ein gegebenes Beispiel der *ortho*-Anteil gemäß folgender Gleichung berechnet werden kann:

$$s_2 = s_{2\text{ipso}} + s_{2\text{ortho}} \tag{8.8}$$

Da aufgrund des Zusammenwirkens zwischen zweiter Sphäre und höheren Sphären in der *ipso*-Position mehrere Wertalternativen für $s_{2\text{ipso}}$ aufgezählt sein können, ergeben sich auf diese Weise auch mehrere Wertalternativen für den *ortho*-Anteil. Im Verlauf der Stichprobe ist sogar durch denselben Effekt, jedoch innerhalb der *ortho*-Position, noch mehr Vielfalt in den Werten möglich. Gerade diese Vielfalt ist jedoch erwünscht und wurde zudem erwartet, um das nicht notwendigerweise streng additive Verhalten der Inkremente bei Variation mehrerer Strukturmerkmale in unterschiedlichen Positionen wiederzugeben („Kreuzterme“, vgl. Seite 87–88).

Für jede Klasse von Atomkombinationen in der *s2ortho*-Position, das heißt für jeden Zustand von `s2ortho`, werden alle so ermittelten Wertalternativen des betreffenden Inkrements (Zustände von `i2ortho`) gespeichert, und es wird ein Histogramm über die Zahl ihrer Vorkommen erstellt. Diese Information wird in einer Parameterdatei gespeichert. Die endgültige Bestimmung der benötigten bedingten Wahrscheinlichkeitsverteilungen erfolgt mit Hilfe dieser Parameterdateien weitgehend analog zum Vorgehen im Falle von $P(\texttt{i2ipso}|\texttt{s2ipso})$.

Betreffend die Bestimmung der bedingten Wahrscheinlichkeiten im Übergang von der Strukturbeschreibung zu den Teilinkrementen der zweiten Sphäre werden also folgende Definitionen vorgenommen:

- Funktion zur Bestimmung der bedingten Wahrscheinlichkeitsverteilung $P(\texttt{i2ipso}|\texttt{s2ipso})$. Die Funktion greift auf eine mithilfe von Literaturwissen erstellte Parameterdatei zurück und kann in den beiden oben beschriebenen Varianten für eine harte oder Weiche Initialisierung verwendet werden.

- Das Literaturwissen zur Bestimmung von $P(\texttt{i2ipso}|\texttt{s2ipso})$ ist in den Parameterdateien jeweils gegeben eine bestimmte Besetzung der ersten Sphäre gespeichert. Am Beispiel eines *sp*-hybridisierten Kohlenstoffatoms in dieser Position soll ihr Aufbau verdeutlicht werden:
  ```
  C_sp 0.0 0.0 0.0 0.0 0.0
  N_nitril -9.2 -9.5 -9.53 -9.53 -9.9
  ```
  Jede Zeile enthält als erstes die Bezeichnung eines Zustandes der Variablen `s2ipso`. Es folgt eine durch Leerzeichen getrennte Liste von in Frage kommenden Werten für das Teilinkrement $s_{2\text{ipso}}$. Die Listen in jeder Zeile müssen gleich lang sein.

- Bestimmen der bedingten Wahrscheinlichkeitsverteilung $P(\texttt{i2ortho}|\texttt{s2ortho})$. Dies geschieht weitgehend analog zur Bestimmung von $P(\texttt{i2ipso}|\texttt{s2ipso})$. Die verwendeten Parameterdateien enthalten jedoch anstelle von Literaturangaben Analyseergebnisse basierend auf einer Stichprobe.

- Analyse einer Stichprobe hinsichtlich des Teilinkrements $s_{2\text{ortho}}$. Für jedes Beispiel wird unter Berücksichtigung des zur Gestalt der ersten Sphäre korrespondierenden Inkrements s_1 das Inkrement s_2 der zweiten Sphäre aus der beobachteten chemischen

Verschiebung berechnet. Aus s_2 werden mit Hilfe von Expertenwissen betreffend den *ipso*-Anteil entsprechende Werte für $s_{2\text{ortho}}$ berechnet und Histogramme über die dabei erreichten Werte in systematisch benannten Dateien abgelegt.

- Die Parameterdateien mit den Analyseergebnissen für die Bestimmung von $P(\texttt{i2ortho}|\texttt{s2ortho})$ sind zeilenweise aus Zahlenpaaren aufgebaut. Die erste Zahl gibt jeweils einen Zustand von `i2ortho`, das heißt einen ganzzahlig gerundeten Wert für das Teilinkrement $s_{2\text{ortho}}$, an. Die zweite Zahl ist die Anzahl der beobachteten Vorkommen dieses Teilinkrements in der Stichprobe. Der zu betrachtende Zustand von `s2ortho` ist dem Dateinamen zu entnehmen.

8.3.2 Relative Häufigkeiten struktureller Eigenschaften

Die bedingten Wahrscheinlichkeiten der strukturbezogenen Variablen `s2ipso`, `s2ortho` und `C_arom` in Abhängigkeit von den Summenformelvariablen (vgl. Abbildung 8.6) sowie die *a-priori*-Wahrscheinlichkeiten der letzteren können mit Hilfe relativer Häufigkeiten empirisch bestimmt werden. Für die Zustände der betreffenden Variablen sind die zu untersuchenden Merkmale in der in Abschnitt 8.2.2 eingeführten Notation beschrieben. Funktionen zur Feststellung des Vorliegens solcher Merkmale sind Teil des Vorverarbeitungsmoduls (vgl. Abschnitt 7.2.2). Es werden also lediglich Prozeduren benötigt, welche die entsprechenden Anfragen für alle Beispiele der Stichprobe iterieren und neben der Größe der Stichprobe anhand der Rückgabewerte die Zahl der positiven Beispiele bestimmen.

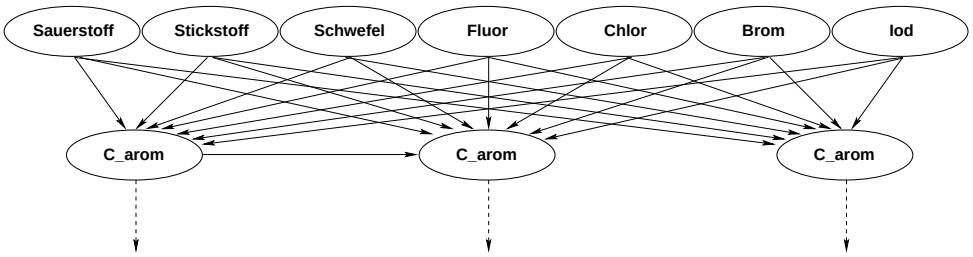


Abb. 8.6: *Übersicht über die strukturbezogenen sowie die Summenformelvariablen des dem System zugrundeliegenden Kausalmodells.*

Darüber hinaus jedoch kann Fachwissen eingesetzt werden, um den Aufwand zu verringern und das Vorgehen somit schneller und effizienter zu gestalten. Dies geschieht bei den bedingten Wahrscheinlichkeiten: Für jeden Zustand der abhängigen strukturbezogenen Variable wird zunächst geprüft, ob dieser das Vorhandensein eines chemischen Elements erfordert, das gemäß der Zustandskombination der gegebenen Spalte der Tabelle jedoch nicht vorhanden ist. Ist dies der Fall, kann die betreffende Komponente der bedingten Wahrscheinlichkeit ohne Untersuchung der Stichprobe den Wert 0 erhalten.

Aufmerksamkeit erfordern darüber hinaus in der Stichprobe nicht beobachtete Ausprägungen von Ereignissen, da offensichtlich ein Unterschied besteht, ob ein Ereignis in einer bestimmten Ausprägung gegeben die mit ihm verbundenen Ursachen prinzipiell möglich ist, aber in der Stichprobe nicht beobachtet wurde, oder ob es prinzipiell unmöglich ist. Würden beide Fälle gleich behandelt, käme dies dem Schluß gleich, daß jede Kombination von Ausprägungen, die in der Stichprobe nicht beobachtet wird, auch außerhalb der Stichprobe nicht vorkommt. Da dieser Schluß nicht ohne weiteres vertretbar ist, müssen entsprechende Maßnahmen getroffen werden.

Ähnliche Situationen sind auch im Zusammenhang mit anderen Techniken der Informatik, welche sich statistischer Methoden bedienen, bekannt, z.B. im Bereich von n-gramm-Modellen in der Sprachverarbeitung. Dort stehen Verfahren zur Verfügung (vgl. [Fin03], S. 105 ff. sowie [Kat87], [NEK94]), die auch hier zur Anwendung kommen können, etwa das intuitive *Adding-One*-Verfahren, *Discounting*-Methoden oder *Backing-Off*.

Der Einfachheit halber wurde *Adding-One* gewählt; das heißt alle prinzipiell möglichen Ausprägungen erhalten betreffend die Häufigkeit ihres Vorkommens bereits vor der Untersuchung der Stichprobe einen festen Sockelbetrag von 1. Eine Untersuchung im Vergleich mit anderen Verfahren, welche die Wahrscheinlichkeit der nicht beobachteten Ereignisse weniger stark überschätzen, kann Gegenstand zukünftiger Arbeiten sein.

Für die empirische Ermittlung der *a-priori*- und bedingten Wahrscheinlichkeiten der strukturbezogenen und der Summenformelvariablen werden folgende Funktionen benötigt:

- Bestimmen der relevanten Ausprägungen der Bedingungen in Gestalt der Zustandsindizes der betreffenden Variablen für eine bestimmte Spalte der Tabelle bedingter Wahrscheinlichkeiten. Dies ist notwendig, um bei bedingten relativen Häufigkeiten die zu prüfenden Eigenschaften zu ermitteln.

- Initialisierung der strukturbezogenen Variablen mit Hilfe von Fachwissen. Erfordert ein Zustand das Vorhandensein eines chemischen Elements, das gemäß der aktuell betrachteten Zustandskombination der Summenformelvariablen nicht vorhanden ist, so wird der betreffende Eintrag mit 0 initialisiert, anderenfalls mit 1. Die mit 0 initialisierten Zustände brauchen bei der empirischen Gewinnung der Wahrscheinlichkeitsverteilung nicht berücksichtigt zu werden.

- Bestimmung von *a-priori*-Wahrscheinlichkeiten. Für jeden Zustand einer gegebenen Variablen wird jeweils die Beschreibung der betreffenden Ausprägung des betrachteten Ereignisses aus den `CSINFO_`*i*-Eigenschaften entnommen. Mit Hilfe dieser Beschreibungen wird die Zahl der positiven Beispiele der Stichprobe festgestellt. Durch Division durch die Stichprobengröße erhält man die benötigte relative Häufigkeit.

- Bestimmung von bedingten Wahrscheinlichkeiten. Für jede Spalte der gesuchten Wahrscheinlichkeitstabelle wird zunächst die gemäß der Zustandskombination der Bedingungen relevante Teilmenge der Gesamtstichprobe zusammengestellt. Für jede einzelne Spalte wird dann so wie im Fall der *a-priori*-Wahrscheinlichkeiten beschrieben verfahren.

Gesondert ist der Fall zu betrachten, daß für eine ganze Spalte der Tabelle einer bedingten Wahrscheinlichkeit keine Beispiele vorliegen ($N_i = 0$), das heißt die Stichprobe enthält keine Fälle, in welchen die gerade betrachteten Ursachen in der für die betreffende Spalte relevanten Kombination von Ausprägungen vorkommen. Unter dieser Voraussetzung kann nichts über die Wahrscheinlichkeit der einzelnen Ausprägungen des abhängigen Ereignisses ausgesagt werden, da keinerlei Informationen darüber zur Verfügung stehen. In diesem Fall wird für die betreffende Spalte eine Gleichverteilung angenommen, da dies die am wenigsten informationshaltige Verteilung ist.

8.3.3 Weiche Summenbildung

Nach den bis hierher beschriebenen Schritten ist das Bayes-Netz bis auf die beiden bedingten Wahrscheinlichkeiten $P(\texttt{peakpos}|\texttt{Sphäre1},\texttt{Sphäre2})$ und $P(\texttt{Sphäre2}|\texttt{s2ipso},\texttt{s2ortho})$ vollständig. Diese beiden Wahrscheinlichkeitsverteilungen werden, wie in Abschnitt 6.1.3 kurz angedeutet, über eine „weiche Summenbildung“ festgelegt.

Das Prinzip wird hier für $P(\texttt{Sphäre2}|\texttt{i2ipso},\texttt{i2ortho})$ erläutert: Für den Fall, daß sich `i2ipso` im Zustand x ppm und `i2ortho` im Zustand y ppm befindet, soll die Wahrscheinlichkeit, daß sich `Sphäre2` im Zustand $x+y$ ppm befindet, hoch sein. Bezieht man außerdem mögliche Rundungsfehler, Meßungenauigkeiten und Wechselwirkungen der Einzeleinflüsse mit ein, so kommen jedoch auch Zustände $x+y+d$ ppm oder $x+y-d$ ppm mit $d \geq 1$ für positive Wahrscheinlichkeitswerte in Betracht.

Die Grundidee erscheint in Kausalrichtung wie ein systematisches „sich mit einer bestimmten Wahrscheinlichkeit Verrechnen“, was nicht unbedingt vorteilhaft wirkt. In Rückschlußrichtung wird die Sinnhaftigkeit jedoch deutlich: Wird ein Peak beobachtet, welcher ein wenig von der streng nach Addition der Inkremente zu erwartenden Position abweicht, so ist es dennoch möglich, ihn auf die korrekten Inkremente und somit auf die korrekten Strukturmerkmale zurückzuführen, so daß die entsprechende Position des Benzolringes zutreffend klassifiziert werden kann, obwohl die Einzelinkremente aufgrund interner Abhängigkeiten und Wechselwirkungen in Wirklichkeit nicht rein additiv zusammenwirken.

Außerdem werden durch dieses Verfahren Rundungsfehler in den Elternvariablen ausgeglichen: Da eine Rundung auf 1 ppm vorgenommen wurde, sind Abweichungen um 1 ppm nach oben oder nach unten gegenüber einer exakten Summation möglich. Beispielsweise werden die Teilinkremente $+0,4$ppm und $+0,3$ppm jeweils auf den Zustand `+0ppm` der Variablen `i2ipso` und `i2ortho` abgebildet. Als Zielzustand von `Sphäre2` ergibt sich aus der Summe der gerundeten Werte `+0ppm`, die tatsächliche Summe von $+0,7$ppm wäre jedoch auf den Zustand `+1ppm` abzubilden.

Um die weiche Summenbildung automatisiert zur Anwendung zu bringen erhalten die beiden betreffenden Wahrscheinlichkeiten in der BNIF-Datei in Analogie zur Beschreibung der mit einzelnen Variablenzuständen zu assoziierenden Merkmale ebenfalls eine Eigenschaft mit dem Schlüsselwort `CSINFO`. Im Falle von $P(\texttt{peakpos}|\texttt{Sphäre1},\texttt{Sphäre2})$ lautet sie

```
CSINFO=SOFT SUM 2.0 128.5
```

`SOFT SUM` bezeichnet die „weiche Summenbildung“. Die beiden Parameter geben die Größe des Intervalls, innerhalb dessen das Ergebnis von der exakten Summe abweichen darf, um dennoch als exakt zu gelten, und eine zu jedem Ergebnis zu addierende Konstante an. Die Intervallgröße ist dabei auch im Zusammenhang mit den Diskretisierungsschritten der betreffenden Variablen zu wählen. Eine Intervallgröße von 2.0 entspricht gerade der zum Ausgleich von Rundungsfehlern benötigten Toleranz von ± 1ppm bei Diskretisierungsschritten von 1 ppm. Zur Bestimmung der gesuchten Wahrscheinlichkeitsverteilungen dient die folgende Funktion:

- Für jede Spalte der betrachteten Wahrscheinlichkeitstabelle wird die Zustandskombination der Ursachenvariablen bestimmt. Jedem Zustand ist ein Zahlenwert zugeordnet, der sich aus dem Zustandsnamen herleitet. Aus diesen und der innerhalb der `CSINFO=SOFT SUM`-Eigenschaft gegebenen Konstante wird die exakte Summe gebildet. Aus den Zustandsbezeichnern der abhängigen Variablen werden nun Zielwerte hergeleitet, und mit jedem davon wird die exakte Summe verglichen. Liegt sie innerhalb des (ebenfalls innerhalb der `CSINFO=SOFT SUM`-Eigenschaft gegebenen) Toleranzintervalls, erhält der zugehörige Eintrag das Gewicht $1,0$. Liegt sie außerhalb des Toleranzintervalls, jedoch innerhalb der über einen zusätzlichen Parameter anzugebenden „Weichheit“, so erhält der zugehörige Eintrag des Zielwerts das Gewicht $\frac{1}{\Delta^2}$, wobei Δ die Abweichung von der näheren Intervallgrenze bezeichnet. Ist $\Delta < 1$, wie es bei Beteiligung von `Sphäre1` an der Summenbildung der Fall sein kann, so wird das Gewicht $\frac{1}{\Delta^2}$ nach oben auf den Wert 1 begrenzt. Abschließend wird eine Normierung zur Spaltensumme 1 durchgeführt.

8.4 Ergebnisse

Ausgehend von einem Kausalmodell, wie in Kapitel 6 beschrieben, welches bereits im BNIF-Dateiformat gegeben, aber noch um die Zahlenwerte der Wahrscheinlichkeitsdefinitionen zu ergänzen ist, wurden Ansätze und Funktionen entwickelt, welche die Vervollständigung des Bayes-Netzes erlauben. Diese greifen neben der BNIF-Datei auf Funktionalitäten des in Kapitel 7 entwickelten Vorverarbeitungsmoduls zurück.

Die im Statistikmodul zusammengefaßten neuen Entwicklungen umfassen das Lesen und Schreiben von BNIF-Dateien sowie die Modifikation der enthaltenen Information, im besonderen der Wahrscheinlichkeitstabellen. Die Parametrisierung des Bayes-Netzes durch Initialisierung dieser Tabellen ist die Hauptaufgabe des Statistikmoduls, und dabei werden alle drei im Zusammenhang mit Bayes-Netzen möglichen Herangehensweisen zur Quantifizierung der Kausalverknüpfungen genutzt: Der Zusammenhang zwischen bestimmten Substituenten und ihren korrespondierenden Inkrementen wird teils basierend auf Expertenwissen, teils semiempirisch hergestellt, Zusammenhänge zwischen bestimmten Strukturanteilen und dem Vorhandensein der einzelnen chemischen Elemente werden empirisch quantifiziert, und betreffend das Zusammenwirken unterschiedlicher Teileinflüsse spielen mathematische Gesetze in Gestalt einer „weichen Summenbildung“ eine Rolle.

Darüber hinaus dient das Statistikmodul zur Analyse einer gegebenen Stichprobe von annotierten ^{13}C-NMR-Spektren, um die Zustände der strukturbezogenen Variablen `s2ipso` und `s2ortho` festzulegen. Es wurden hierzu die Atome der zweiten Sphäre in der *ipso*- und *ortho*-Position untersucht und jeweils das Vorkommen der beobachteten Atomkombinationen gezählt. Die Zustände der Variablen sind dann so zu definieren, daß die große Vielzahl der als Elementarereignisse betrachteten möglichen Kombinationen von Atomtypen hinsichtlich struktureller Gemeinsamkeiten sowie der Häufigkeit ihres Auftretens sinnvoll zusammengefaßt werden. Die resultierende Gestaltung der betreffenden Variablen ist in Tabelle 8.4 und 8.5 dargestellt.

Außerdem wurde eine systematische Notation zur Beschreibung der Merkmale, welche zu den einzelnen Zuständen korrespondieren, entwickelt. Mithilfe derselben kann eine automatische Initialisierung der bedingten Wahrscheinlichkeiten der strukturbezogenen Variablen (vgl. Abbildung 8.6) durchgeführt werden.

Das Bayes-Netz ist nun bereit, um zur Klassifikation der einzelnen Positionen des Benzolringes eingesetzt zu werden. Bei seiner Evaluation sind zwei Ausrichtungen zu berücksichtigen: einerseits die korrekte Klassifikation der *ipso*-Position gegeben die Peakposition und die Summenformelinformation und andererseits die korrekte Klassifikation der *ortho*-Positionen, wenn die korrekte Klasse der *ipso*-Position bereits gegeben ist. Beides muß mit Blick auf die Hypothesengenerierung, welche sich im Verarbeitungsverlauf des Gesamtsystems an die Klassifikation anschließt, verläßlich sein.

Ein entsprechendes Hypothesengenerierungsmodul bleibt nun im letzten Schritt zur Vervollständigung des Gesamtsystems zu entwickeln. Der Strukturgenerator erhält die Ausgabe des Bayes-Netzes für jede der sechs Positionen des Benzolringes als Eingabe. Neben der *ipso*-Position werden auch unter der Annahme einer korrekten *ipso*-Klassifikation die *ortho*-Positionen klassifiziert, so daß insgesamt ein dreigliedriges Ringfragment entsteht. Diese Fragmente sollen überlappend zusammengefügt werden, so daß aufbauend auf der Klassifikation der einzelnen Substituenten auch das Substitutionsmuster ermittelt werden kann.

Neben den anstehenden Entwicklungen auf dem Weg zum Gesamtsystem SASCHA gibt es auch Ansatzpunkte für zukünftige Arbeiten zur grundsätzlichen Erweiterung der Funktionalitäten des Statistikmoduls, welche über den Rahmen der gegenwärtigen Arbeit hinausgehen.

Relative Häufigkeiten können etwa neben der Approximation der Wahrscheinlichkeiten prinzipiell dazu genutzt werden, Unabhängigkeiten oder Abhängigkeiten zwischen Ereignissen näher zu untersuchen. Entsprechende Routinen hierzu wären bei einer Weiterentwicklung des Kausalmodells hilfreich, wenn auch entferntere Atome betrachtet werden und sich die Frage stellt, in welcher Weise das Zusammenwirken ihrer Einflüsse zu modellieren ist. Ein weiterer Punkt ist die Frage, ob ein anderes Vorgehen als das derzeit praktizierte *Adding-One* im Umgang mit nicht beobachteten Ausprägungen von Ereignissen Vorteile bringt.

9 Hypothesengenerierung, Kontrollstrategie und Integration zum Gesamtsystem SASCHA

Nachdem ein Kausalmodell der Domäne entwickelt und mit Hilfe statistischer Verfahren als Bayes-Netz zur Klassifikation der Substituenten in den einzelnen Positionen des Benzolrings realisiert wurde, wird zum Abschluß der Verarbeitung noch ein Modul benötigt, welches die einzelnen Klassifikationsergebnisse zu dem gewünschten Substitutionsmuster zusammenfügt. Erst dieser Schritt der Hypothesengenerierung liefert als Resultat des Musteranalyseprozesses eine symbolische Beschreibung der Molekülstruktur.

Das betreffende Modul wird in Abschnitt 9.1 vorgestellt. Es muß die Ausgabe des Klassifikators verarbeiten, welche ein dreigliedriges Ringfragment für jede Position ergeben. Diese Fragmente sind überlappend zusammenzusetzen, um das gesamte Substitutionsmuster zu erhalten. Anschließend werden in Abschnitt 9.2 solche Fälle diskutiert, in welchen durch Nichtvereinbarkeit einzelner Bausteine keine oder keine befriedigende Hypothese generiert werden kann. Dies liefert einen Einblick in die Aufgabe des Entwickelns einer Kontrollstrategie, deren hohe Komplexität aufgrund des Umfangs der notwendigen Überlegungen und Abwägungen den Rahmen der gegenwärtigen Arbeit jedoch übersteigt. Es wird statt dessen auf eine möglichst informative Ergebnisausgabe wertgelegt.

Mit der Hypothesengenerierung und dem (wenngleich derzeit lediglich einen linearen Verarbeitungsverlauf umfassenden) Kontrollmodul sind sodann alle für die Integration zu einem Musteranalysesystem nötigen Module implementiert. Abschnitt 9.3 beschreibt ihre Verbindung zum Gesamtsystem SASCHA, und Abschnitt 9.4 schließlich resümiert in einem Überblick die abschließenden Entwicklungsschritte.

9.1 Hypothesengenerierung: Zusammensetzen des Benzolrings

Ziel der Hypothesengenerierung ist der Aufbau eines Benzolringes aus sechs Einzelfragmenten. Jedes davon ist das Resultat eines Klassifikationsschrittes und enthält somit Informationen über eines der Ringatome (*ipso*-Position) und seine wahrscheinlichsten Nachbarn (*ortho*-Positionen). Abbildung 9.1 zeigt ein solches Fragment.

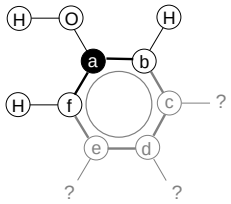


Abb. 9.1: *Ergebnis eines Klassifikationsschritts im Kontext des Benzolrings. Die Kohlenstoffatome des Rings sind schematisch mit a–f anstelle des Elementsymbols C bezeichnet.*

Im folgenden wird beschrieben, wie durch Überlagerung von Fragmenten der Benzolring aufgebaut wird. Die Grundidee wird unter der Annahme erläutert, daß alle sechs Fragmente sowohl bezüglich der *ipso*- wie auch der *ortho*-Positionen korrekt klassifiziert wurden. Abschnitt 9.1.2 beschreibt die für die Realisierung nötigen Funktionen und Datenstrukturen.

9.1.1 Grundidee

Bei der Hypothesengenerierung wird als Ausgangsinformation eine Liste von sechs dreigliedrigen Ringfragmenten verwendet, wie in Abbildung 9.2 dargestellt. Jedes davon beinhaltet den gemäß Klassifikation angenommenen Typ des *ipso*-Substituenten in der mittleren Position des Fragments sowie die unter dieser Voraussetzung wahrscheinlichsten *ortho*-Substituenten in den beiden äußeren Positionen. Die Liste ist gemäß der *a-posteriori*-Wahrscheinlichkeit der *ipso*-Klasse sortiert.

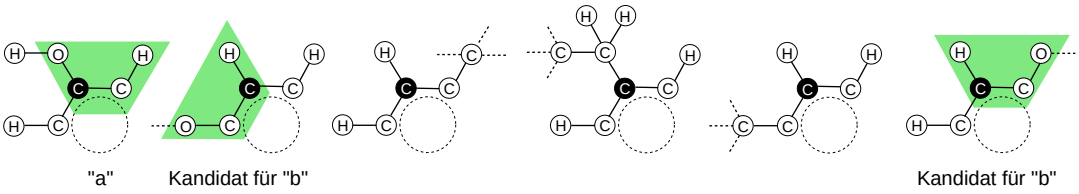


Abb. 9.2: *Eingabe für den Aufbau des Benzolringes. Die Fragmente sind nach der* a-posteriori-*Wahrscheinlichkeit der* ipso-*Klassen geordnet. Fragment "a" ist das Fragment mit der höchsten Wahrscheinlichkeit, die Kandidaten für "b" ergeben sich durch Überlappung.*

Das vorderste Listenelement mit der höchsten Wahrscheinlichkeit wird mit Fragment "a" benannt, das heißt, seine mittlere Position bezeichnet Position "a" des aufzubauenden Benzolringes. Kandidaten für die im Uhrzeigersinn folgende Position "b" werden anhand der Überlappung mit Fragment "a" ermittelt (in Abbildung 9.2 grün markiert): Die linke äußere Position des Kandidaten muß mit der mittleren von Fragment "a" sowie die mittlere Position des Kandidaten mit der rechten äußeren von "a" übereinstimmen. Bei mehreren Kandidaten wird der mit höchsten *a-posteriori*-Wahrscheinlichkeit in der *ipso*-Klassifikation bevorzugt.

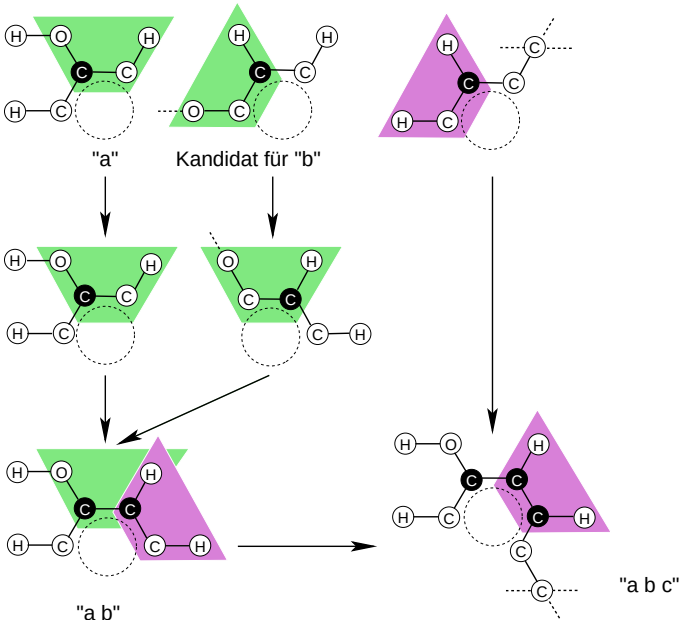


Abb. 9.3: *Schrittweiser Aufbau des Benzolrings anhand von Überlappungen. Die in jedem Aufbauschritt ausgenutzte Übereinstimmung ist jeweils farbig hervorgehoben.*

Der Aufbau von Fragment "ab" sowie der nachfolgende Aufbauschritt zu Fragment "abc"

sind in Abbildung 9.3 dargestellt: Nach der beschriebenen Ermittlung eines Kandidaten für Position “b” wird Fragment “a” durch Verschmelzung mit diesem zu Fragment “ab” mit insgesamt vier Positionen erweitert. Die vormals mittlere Position heißt nun “a”, die vormals rechte äußere Positionen, welche mit der mittleren des Kandidatenfragments übereinstimmt, heißt nun “b”, und die zuvor rechte äußere Position des Kandidaten ist nun die rechte äußere Position des neuen Fragments. Bei der Ermittlung von Kandidaten für die nächste Position “c” muß diese mit der mittleren des neuen Kandidatenfragments sowie dessen linke äußere Position mit Position “b” in Fragment “ab” übereinstimmen. Anschließend werden das Fragment “ab” und der Kandidat für “c” wie beim Aufbau von “ab” beschrieben zu einem neuen Fragment “abc” verschmolzen.

In derselben Weise wird ein Kandidat für die darauffolgende Position “d” ermittelt und das entsprechende Fragment an den im Aufbau befindlichen Ring angefügt. Die beiden verbleibenden Fragmente dienen dazu, den Ring zu schließen bzw. den Ringschuß zu verifizieren. Dies geschieht folgendermaßen, wie in Abbildung 9.4 veranschaulicht: Zunächst wird in der bekannten Weise ein Kandidat für Position “e” bestimmt und an Fragment “abcd” angefügt, so daß ein Fragment “abcde” entsteht. Der Ring kann geschlossen werden, falls rechte und linke äußere Position des Fragments übereinstimmen. Diese Position wird nun mit “f” bezeichnet. Der Ringschluß ist verifiziert, wenn Position “f” mit der mittleren Position und die Positionen “a” und “e” mit den beiden äußeren Positionen des letzten noch unverwendeten Fragments übereinstimmen.

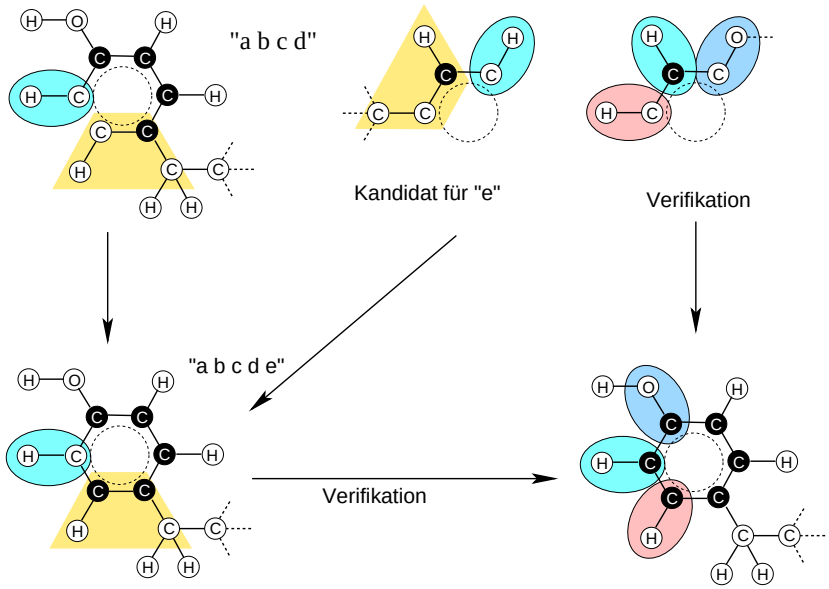


Abb. 9.4: *Der Benzolring kann mit dem fünften Fragment der Ausgangsliste erfolgreich aufgebaut werden (links). Durch Überlappung mit dem letzten verbleibenden Fragment wird der Ringschluß verifiziert (rechts).*

Bei der Umsetzung dieses Schemas sind jedoch grundsätzlich zwei Dinge zu beachten:

- Die *ortho*-Positionen sind vertauschbar.
- Die *s2ortho*-Zustände fassen meist mehrere Einzelausprägungen zusammen.

Der Vertauschbarkeit der *ortho*-Positionen jedes Einzelfragments wird in der Praxis dadurch Rechnung getragen, daß jeweils nicht ein Fragment, sondern zwei Fragmentalterna-

tiven, die sich wie Bild und Spiegelbild verhalten, betrachtet werden. Derselbe Ansatz der Berücksichtigung mehrerer Alternativen für jedes Fragment kann verwendet werden, um mehrere unterschiedliche Atomkombinationen, für welche ein und dieselbe *s2ortho*-Klasse steht, zu berücksichtigen. Die oben beschriebene Vorgehensweise wird dann jeweils für alle Alternativen iteriert. Ein Kandidat kann erfolgreich an ein Ausgangsfragment angefügt werden, wenn für mindestens eine der Alternativen des Kandidaten die beschriebene Übereinstimmung mit einer Alternative des Ausgangsfragments festgestellt wird.

Trotz der Berücksichtigung jeweils mehrerer möglicher Alternativen in den *ortho*-Positionen wird der Ring aufgrund der Eindeutigkeit der *ipso*-Positionen jedoch eindeutig aufgebaut. In Abbildung 9.3 und 9.4 sind jeweils diejenigen Ringatome schwarz hervorgehoben, deren Substituenten eindeutig feststehen. Zuletzt wird der Ring noch in eine Standardorientierung gebracht, um eine einheitliche und eindeutige Beschreibung des Substitutionsmusters zu ermöglichen.

9.1.2 Funktionen und Datenstrukturen

Das im vorangegangenen Abschnitt beschriebene Verfahren ist schematisch klar umrissen, so daß die notwendigen prinzipiellen Funktionalitäten nicht schwer auszumachen sind. Es werden außerdem jedoch Datenstrukturen benötigt, welche den Ansprüchen der Hypothesengenerierung gerecht werden. Wichtig ist dabei einmal mehr eine stark strukturbezogene Betrachtung organischer Moleküle, jedoch ist die in Abschnitt 7.2.2 eingeführte Darstellung in Gestalt einatomiger Strukturfragmente hier weniger geeignet. Vielmehr ist eine Realisierung anzustreben, welche ebenso wie der Strukturgenerator auf Benzolringe bzw. auf Ringfragmente spezialisiert ist.

Ein Benzolring zeichnet sich in der gegenwärtigen Sichtweise durch seine sechs Substituenten aus. Ist nur ein Ringfragment zu repräsentieren, werden die übrigen Ringpositionen in Analogie zu Abbildung 9.2 als unbekannt bezeichnet. Wichtig ist darüber hinaus die Anordnung der Positionen, wie sie in Abbildung 9.2 durch die Benennung mit a–f wiedergegeben wird. Im Rahmen des sukzessiven Ringaufbaus wird außerdem die Information der Wahrscheinlichkeitsbewertung jeder Position benötigt.

Ein Ring enthält demnach für jede der sechs Positionen das chemische Elementsymbol des *s1ipso*-Atoms des dort gebundenen Substituenten sowie die korrespondierende Klasse im *ipso*-Klassifikationsschritt und ihre *a-posteriori*-Wahrscheinlichkeit. Darüber hinaus ist, wiederum für jede Position, die wahrscheinlichste *ortho*-Klasse bei gegebener wahrscheinlichster *ipso*-Klasse bekannt und ebenfalls mit der zugehörigen *a-posteriori*-Wahrscheinlichkeit verbunden. Die entsprechende Datenstruktur ist in Abbildung 9.5 dargestellt.

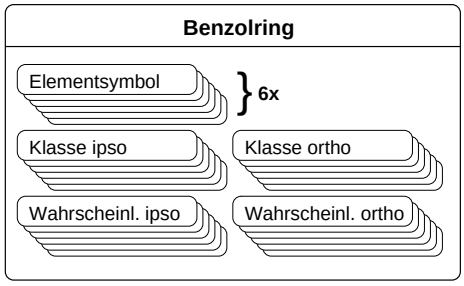


Abb. 9.5: *Datenstruktur zur Repräsentation von Benzolringen und Ringfragmenten.*

Handelt es sich nicht um einen vollständigen Benzolring, sondern um ein Ringfragment,

so sind einzelne der sechs Positionen undefiniert. Zur Repräsentation alternativer Belegungen einzelner Positionen wird eine Liste derartiger Objekte verwendet, die sich nur in den betreffenden Positionen unterscheiden und ansonsten identisch sind. Es sei in diesem Zusammenhang darauf hingewiesen, daß unterschiedliche Varianten der Belegung nur jeweils in den beiden Randpositionen eines Fragments möglich sind.

Beim Aufbau des Benzolringes, das heißt bei der Bestimmung des Substitutionsmusters anhand der Klassifikationsresultate, werden die folgenden Funktionen genutzt, um das in Abschnitt 9.1.1 beschriebene Verfahren umzusetzen:

- Überführen der Klassifikationsergebnisse in eine Repräsentation als Ringfragmente. Die Funktion liefert sechs Listen von Fragmentalternativen. Jede Liste steht für eine Position des Ringes. Die Alternativen beziehen sich auf die unterschiedlichen möglichen Belegungen der *ortho*-Positionen gemäß der *ortho*-Klassifikation.

- Anordnen der sechs Listen gemäß der *a-posteriori*-Wahrscheinlichkeit der *ipso*-Klassifikation.

- Ermitteln von Kandidaten für die nächste Position im Uhrzeigersinn. Die Funktion erhält ein Ringfragment und die unverwendeten Fragmente der Ausgangsliste (jeweils in Gestalt einer Liste von Alternativen) und liefert den Kandidaten mit der höchsten Wahrscheinlichkeitsbewertung der *ipso*-Klasse.

- Verschmelzen zweier Fragmente entsprechend ihrer Überschneidung. Die Funktion erhält den im Aufbau befindlichen Ring und ein Kandidatenfragment und fügt dieses an den Ring an. Dabei muß die Wahl der passenden Alternative in der *ortho*-Position beachtet werden.

- Schließen des Ringes als ein Spezialfall des Verschmelzens zweier Fragmente, bei dem eine zusätzliche Übereinstimmung des hinzuzufügenden Fragments mit dem gegenüberliegenden Rand des im Aufbau befindlichen Rings erforderlich ist.

- Verifizieren des Ringschlusses und damit eindeutige Bestimmung der letzten Ringposition. Die Funktion erhält den bis dahin aufgebauten Ring in Gestalt potentiell mehrerer Alternativen sowie das letzte unverwendete Fragment und liefert diejenige Ringalternative, die sich mit dem Fragment in Einklang bringen läßt.

- Drehung des Ringes in eine Normlage.

- Aufbau eines Substitutionsmusters. Die Funktion erhält die Klassifikationsergebnisse und nutzt obige Funktionen, um daraus einen Benzolring aufzubauen, an welchem das Substitutionsmuster abgelesen werden kann.

9.2 Konflikte bei der Hypothesengenerierung

Die Entwicklung eines Verfahrens für den Fehlerfall, das heißt daß der Benzolring nicht erfolgreich aufgebaut werden konnte, erfordert, wie sich im folgenden zeigen wird, umfangreiche Untersuchungen und Überlegungen betreffend eine günstige Strategie. Wenngleich dies somit den Rahmen der vorliegenden Arbeit übersteigt, da diese gegenwärtig ihren Schwerpunkt im Bereich der Modellentwicklung hat, sollen jedoch einige für den Aufbau des Gesamtsystems sowie auch für seine Weiterentwicklung in zukünftigen Arbeiten hilfreiche Vorüberlegungen dargelegt werden. Darüber hinaus wird das Grundgerüst eines Kontrollmoduls

entwickelt, welches zunächst zwar im Fehlerfall nur eine Ausgabe und Dokumentation des aufgetretenen Konflikts anstößt, gleichwohl aber sowohl als Grundlage zukünftiger Entwicklungen wie auch aufgrund der dokumentierten Ergebnisausgabe als Werkzeug für dieselben dienen kann.

Zu erwarten ist der Fehlerfall im besonderen bei falschen Klassifikationsergebnissen. Eine geeignete Kontrollstrategie muß dann systematisch und gezielt Revisionen einzelner Klassifikationsresultate veranlassen, um den Konflikt aufzulösen und so die Anfrage des Benutzers nach dem Substitutionsmuster der untersuchten Verbindung dennoch beantworten zu können. Die folgenden Ausführungen sollen dazu dienen, zum einen die Vielschichtigkeit und Komplexität der zur Entwicklung einer solchen Strategie nötigen Überlegungen zu verdeutlichen. Zum anderen sind sie auch für den gegenwärtigen Umgang mit Konflikten bei der Hypothesengenerierung interessant, um eine möglichst informative Ausgabe an den Benutzer zu ermöglichen: Diese soll einerseits den aufgetretenen Konflikt beschreiben und das bis zu seinem Auftreten generierte Zwischenresultat des Substitutionsmusters ausgeben, andererseits kann sie aber auch dazu dienen, Informationen zur Verfügung zu stellen, die für die zukünftige Entwicklung einer Kontroll-, Revisions- und Hypothesenverwaltungsstrategie genutzt werden können.

Hierbei ist zuerst einmal zu unterscheiden, wie gravierend die eingetretene Konfliktsituation ist. Konnte der Ring geschlossen, jedoch der Ringschluß nicht verifiziert werden? Scheiterte nach dem Anfügen des vorletzten Fragments bereits der Ringschluß? Traten sogar schon früher während des Aufbaus Konflikte auf? Je später im Verlauf der Strukturgenerierung das Problem auftritt und je weniger Positionen innerhalb des Benzolringes es betrifft, um so weniger schlimm wird der Konflikt bewertet.

Bevor jedoch ein Fehler als solcher identifiziert wird und Korrekturmaßnahmen eingeleitet werden, sollte sichergestellt sein, daß der Konflikt nicht darin seine Ursache hat, daß die Fragmente in der falschen Reihenfolge zusammengefügt wurden. Es sind daher alternative Aufbauwege zu überprüfen. Naheliegend wäre es, dabei das zuletzt angefügte Fragment als „konfliktverursachendes Fragment“ zu betrachten und nach einem alternativen Kandidaten für die betreffende Position zu suchen. Es ist jedoch möglich, daß der Fehlschlag beim Aufbau des Ringes bereits früher durch die Wahl eines falschen Fragments verursacht wurde. Daher erscheint es günstig, in jedem einzelnen Aufbauschritt festzuhalten, ob es andere Kandidaten für die betreffende Position gegeben hätte, und nicht erst im Falle eines Konflikts durch erneuten Versuch nach ihnen zu suchen. Sobald ein Konflikt auftritt, wird dann der betreffende alternative Aufbauweg aktiviert. Führt auch dies nicht zum Erfolg, muß jedoch der Konflikt behandelt werden.

9.2.1 Fehlschlag der Verifikation

Fehlschläge in der Verifikation des Ringschlusses sind die am wenigsten gravierenden Fehler, die im Rahmen der Hypothesengenerierung auftreten können. Um einen Benzolring erfolgreich aufzubauen genügen grundsätzlich fünf der sechs Fragmente, die sechste Position wird dann wie eingangs beschrieben aufgrund einer Übereinstimmung des rechten und linken Randes des im Aufbau befindlichen Ringes erschlossen. Für beide Ränder stehen jedoch aufgrund der Abbildung von jeweils mehreren Atomkombinationen auf ein und denselben `s2ortho`-Zustand mehrere Alternativen zur Verfügung. Somit ist die sechste Position potentiell noch nicht eindeutig bestimmt, sondern nur auf eine Auswahl von einigen Klassen eingeschränkt. Abbildung 9.6 verdeutlicht den Zustand des im Aufbau befindlichen Rings zu diesem Zeitpunkt.

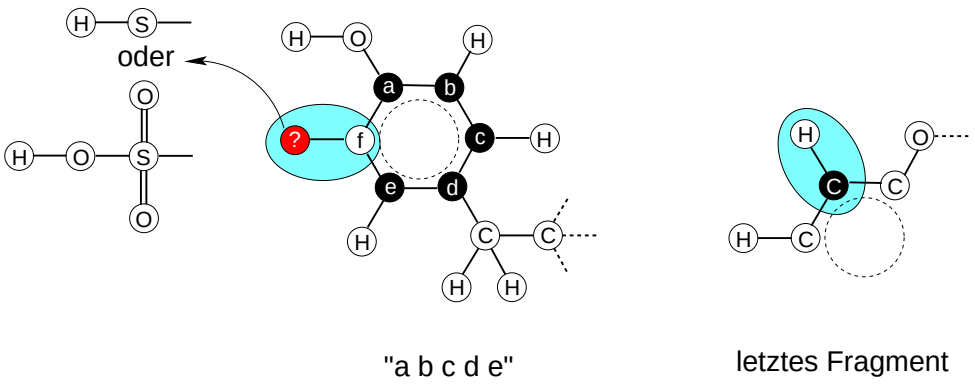


Abb. 9.6: *Im Aufbau befindlicher Benzolring. Die Kohlenstoffatome des Rings sind schematisch mit a – f bezeichnet. Als nächster Schritt ist der durchgeführte Ringschluß zu verifizieren. Das gezeigte Beispiel stellt einen Fehlschlag der Verifikation dar.*

Bei der Verifikation des Ringschlusses finden anschließend zwei Teilschritte statt: Zunächst wird überprüft, ob einer der in Frage kommenden Substituenten für Position “f” mit der *ipso*-Position des letzten verbleibenden Fragments übereinstimmt (*ipso*-Verifikationsschritt). Darauf folgt eine Prüfung der Übereinstimmung der Kombination von Position “a” und “e” mit den *ortho*-Positionen des Fragments.

Schlägt nur der zweite (*ortho*-)Verifikationsschritt fehl, wird aber die Position “f” durch die *ipso*-Klasse des letzten Fragments bestätigt, so kann aufgrund dieser zumindest teilweise erfolgten Bestätigung die Generierung des Substitutionsmusters als erfolgreich betrachtet und die entsprechende Hypothese zurückgeliefert werden. Auf jeden Fall sollte ihr jedoch ein bewertender Kommentar beigefügt werden, anhand dessen der Benutzer ihre Verläßlichkeit beurteilen kann. Hinsichtlich der Frage, welche der verfügbaren Informationen für eine derartige Bewertung geeignet sind, bietet sich die *a-posteriori*-Wahrscheinlichkeit derjenigen *ortho*-Klasse an, die gegeben die erfolgreich validierte *ipso*-Klasse zu einer vollständigen Validierung des Ringschlusses geführt hätte; für eine noch bessere Einschätzung kann sie zusammen mit dem Rang der betreffenden Klasse ausgegeben und außerdem in diesen beiden Werten der bestbewerteten Klasse gegenübergestellt werden.

Für den Fall, daß die *ipso*-Klasse des sechsten Fragments mit keiner der ermittelten Alternativen für Position “f” übereinstimmt, kann das ermittelte Substitutionsmuster dagegen nicht als erfolgreich generierte Hypothese angesehen werden. Bei einem solchen Fehlschlag der Verifikation wird die generierte Teilhypothese als fehlerhaft markiert. Analog zu obigem Vorgehen wird darüber hinaus für jede Substituentenklasse, die für Position “f” in Frage kommt, ihre *a-posteriori*-Wahrscheinlichkeit und ihr Rang ermittelt, jedoch anders als oben natürlich nach dem *ipso*-Klassifikationsschritt aus der Verteilung der Hypothesenvariable `C_arom`.

Auf eine Wiederholung des *ortho*-Klassifikationsschritts zur Ermittlung der *a-posteriori*-Wahrscheinlichkeiten für die *s2ortho*-Klassen wird zum gegenwärtigen Zeitpunkt verzichtet. Sie entspräche zwar einer Revision des Klassifikationsresultats, jedoch unter Verwendung von nur sehr wenig neuer, zusätzlicher Information. Somit ist zweifelhaft, ob daraus überhaupt im Regelfall ein Fragment resultierten würde, das zur erfolgreichen Validierung genutzt werden kann. Besonders klar wird dies an dem Fall, daß alle zum durchgeführten Ringschluß passenden *ipso*-Klassen schlechte Wahrscheinlichkeitsbewertungen besitzen: Hier wäre es nicht weniger plausibel, davon auszugehen, daß nicht das Validierungsfragment, sondern ei-

nes der bereits früher in den Ring eingefügten Fragmente fehlerhaft klassifiziert wurde. Es ist jedoch nicht offensichtlich, welches Fragment in diesem Fall auf welche Weise revidiert werden sollte. Hierin zeigt sich ein erstes Indiz dafür, daß die Entwicklung einer intelligenten Kontroll- und Revisionsstrategie einer intensiven Bearbeitung bedarf.

9.2.2 Konflikte während des Ringaufbaus

Läßt sich während des Aufbaus der Benzolring nicht schließen oder treten bereits früher Konflikte in der Form auf, daß sich keine Kandidaten mehr für die nächste Position finden lassen, und führt auch keiner der zu Beginn von Abschnitt 9.2 erwähnten alternativen Aufbauwege zum Erfolg, so ist damit das Ergebnis der Klassifikation in Frage gestellt. Offen bleibt jedoch, wie leicht und in welchem Umfang Revisionen angestrebt werden sollten.

Die vom Klassifikator gelieferte Ausgangsinformation sollte sicherlich um so zurückhaltender revidiert und statt dessen so lange und in so großem Umfang wie möglich an den ursprünglichen Resultaten festgehalten werden, je zuverlässiger der Klassifikator ist. Um dies jedoch festzustellen wird eine geeignete Kenngröße benötigt, bezüglich welcher der Klassifikationsschritt zu evaluieren ist. Dies ist Gegenstand von Kapitel 10.

Ein durchdachtes Vorgehen muß zudem dafür Sorge tragen, daß nicht wahllos der Raum möglicher Lösungen durchsucht wird – Ziel muß es vielmehr sein, die in Gestalt der Klassifikationsresultate, der zugehörigen Wahrscheinlichkeitsverteilungen und der Zwischenresultate der Hypothesengenerierung verfügbare Information auf intelligente Weise in den Revisionsprozeß einzubringen. Es gibt jedoch sehr viele Überlegungen, die bei der Entscheidung, wo eine Revision der Klassifikationsresultate ansetzen soll, eine Rolle spielen können. Einige werden exemplarisch im folgenden dargelegt.

1. Größtes Fragment:
   Im Sinne einer möglichst systematischen Erschließung des Raumes alternativer Lösungen bietet es sich an, stets vom größten zusammengesetzten Fragment auszugehen, das im Rahmen der Überprüfung alternativer Aufbauwege erreicht werden konnte. Dies muß nicht notwendigerweise diejenige Variante sein, die nach dem Standardverfahren unter Bevorzugung des Fragments mit der höchsten *a-posteriori*-Wahrscheinlichkeit aufgebaut wurde.

2. Ursache des Konflikts:
   Als „Konfliktfragment“ wird dasjenige Fragment bezeichnet, durch welches es unmöglich wurde, den Aufbau des Substitutionsmusters fortzusetzen, also das zuletzt an den im Aufbau befindlichen Ring angefügte Fragment. Man kann argumentieren, daß dieses Fragment zu revidieren ist, da es die Fortführung des Aufbaus blockiert. Dabei stellt sich die Frage, ob die Einschätzung der *ipso*- und damit auch der *ortho*-Klasse oder nur die letztere revidiert werden soll.

3. Plausible Klassen:
   Bereits bevor Informationen aus dem Verlauf der Hypothesengenerierung zur Verfügung stehen, können Überlegungen hinsichtlich für eine Revision infragekommender Fragmente angestellt werden. Es bietet sich an, bevorzugt solche Fragmente zu revidieren, bei welchen es (potentiell sogar mehrere) Klassen neben der wahrscheinlichsten gibt, die ebenfalls als plausibel anzusehen sind. Für diese Plausibilität wird ein Schwellwert benötigt – sinnvoll erscheint eine relative Festlegung mit Bezug zur Wahrscheinlichkeit der bestbewerteten Klasse, z.B. auf $\frac{1}{3}$: In diesem Fall würden all diejenigen Klassen als plausibel angesehen, deren Wahrscheinlichkeit mindestens $\frac{1}{3}$ der

Wahrscheinlichkeit der bestbewerteten Klasse beträgt. Daraus folgt auch, daß keine anderen Klassen plausibel sind, wenn die bestbewertete Klasse eine Wahrscheinlichkeit von mehr als 0,75 besitzt: $\frac{1}{3}$ dieses Wertes (0,25) schöpfen mit diesem zusammen die Gesamtwahrscheinlichkeitsmasse von 1,0 vollständig aus. Diese Überlegungen können sowohl für den *ipso*- wie auch für den *ortho*-Klassifikationsschritt angestellt werden. Außerdem stellt sich die Frage nach dem konkret zu wählenden Schwellwert: Das obige Beispiel von $\frac{1}{3}$ liefert sicherlich in mehr Fällen plausible Alternativen als etwa die Wahl von $\frac{4}{5}$ oder $\frac{9}{10}$. Letzteres würde bedeuten, daß nur dann alternative Resultate als plausibel angesehen werden, wenn sich keine der Alternativen durch eine deutlich höhere Wahrscheinlichkeit vor den anderen auszeichnet.

4. Unterstützung des *ortho*-Klassifikationsresultats:
   Anhand der *ipso*-Klassen können die angenommenen *ortho*-Klassen untermauert werden: Aus den *ipso*-Klassen werden Paare gebildet und die so erhaltenen Atomkombinationen auf *ortho*-Klassen abgebildet. Geht man davon aus, daß alle *ipso*-Klassen korrekt sind, so folgt daraus, daß keine anderen als die so gebildeten *ortho*-Klassen auftreten dürfen. Diese Betrachtung kann zusätzlich verfeinert werden, indem man berücksichtigt, daß jede *ipso*-Klasse nur zwei *ortho*-Klassen unterstützen kann.
   In analoger Weise können prinzipiell auch die vorgefundenen *ortho*-Klassen die *ipso*-Klassen bestätigen. Da jedoch teilweise mehrere Atomkombinationen auf ein und dieselbe *ortho*-Klasse abgebildet werden, sind hier aufgrund der Alternativen einerseits erheblich mehr Vergleiche nötig, andererseits sind aus demselben Grund weniger konkrete Informationen zu erwarten.

5. Unsichere Überschneidungen:
   Fragmente, deren *ortho*-Klassen eine relativ niedrige Wahrscheinlichkeit besitzen, können um so mehr als eine unsichere Informationsquelle für den Aufbau des Substitutionsmusters angesehen werden, je mehr alternative Atomkombinationen außerdem für die betreffende *ortho*-Klasse in Frage kommen. Wird beim Aufbau die *ortho*-Position eines solchen Fragments mit der *ipso*-Position eines anderen Fragments verschmolzen, die ebenfalls eine relativ niedrige Wahrscheinlichkeitsbewertung besitzt, so kann die betreffende Position als Schwachpunkt im Ringaufbau angesehen werden.

An diesen Überlegungen werden zunächst die drei grundsätzlichen Aspekte deutlich, die für die Entwicklung einer Revisionsstrategie, das heißt der Hauptfunktionalität des Kontrollmoduls, eine Rolle spielen: zum einen die Wahl des Ansatzpunktes, zum anderen die verfügbare zusätzliche Information aus dem Versuch, das Substitutionsmuster aufzubauen, und zum dritten die eigentliche Strategie, das heißt Gewichtung und Einbringung der Zusatzinformation sowie Bewertung der festgestellten Fakten.

Hinsichtlich der Wahl des Ansatzpunktes für Revisionen der Klassifikationsergebnisse bieten alle oben aufgeführten Punkte eine Grundlage. Fragmente, die beim Aufbau des Substitutionsmusters zu Schwachstellen führen (Punkt 5), kommen sicherlich ebenso in Frage wie konfliktverursachende Fragmente (Punkt 2) oder solche Fragmente, bei welchen es bereits ohne die Zusatzinformation aus dem Verlauf der Hypothesengenerierung alternative Klassen für die Substituenten gibt, die ebenfalls als plausibel eingeschätzt werden (Punkt 3).

Aber auch bei gegebenem Ansatzpunkt sind noch längst nicht alle Fragen beantwortet. Die in den nicht zu revidierenden Fragmenten enthaltene strukturelle Information soll nun genutzt werden, um das gewählte Fragment zu revidieren. Derartige Zusatzinformation liefert beispielsweise das in Punkt 4 beschriebene Prinzip der wechselseitigen Unterstützung

korrekter Klassen, aber auch sonstige Zusatzinformationen, die über die bislang berücksichtigte Summenformelinformation und den spektralen Befund hinausgehen, können eine Rolle spielen – etwa die (teilweise) Kenntnis der Ausgangssubstanzen oder des Syntheseweges bei der Herstellung der untersuchten Probe. In jedem Fall ist zu überlegen, auf welche Weise solche Informationen als zusätzliche Evidenzen in den Klassifikationsschritt einfließen können.

Auch stellt sich die Frage, wie im Falle erneuter Konflikte nach der Revision zu verfahren ist: Soll statt des gewählten Fragments ein anderes als Ansatzpunkt dienen, oder soll die in Abschnitt 9.1.1 beschriebene Strategie ausgehend von den neuen Ergebnissen wiederholt werden? Sollen dabei ältere Teilresultate mit berücksichtigt oder diese komplett verworfen werden? Diese und viele weitere Fragen und Überlegungen führen zum dritten und wohl auch anspruchsvollsten der oben genannten Punkte, der eigentlichen Strategie im Sinne einer Gewichtung und Einordnung aller zugänglichen Fakten.

Bei der Vielzahl möglicher Vorgehensweisen ist es nicht ohne weiteres möglich, eine einzelne, herausragende Strategie festzulegen, welche sich vor allen anderen Kombinationen von Kriterien und Überlegungen offensichtlich auszeichnet. Bereits die wenigen hier angedeuteten Punkte würden als Basis umfangreicher Erörterungen und Diskussionen ausreichen, um genügend Stoff für eine weitere Forschungsarbeit zu bieten. Die Entwicklung eines Kontrollmoduls, dessen Hauptfunktionalität die Steuerung derartiger Revisionen wäre, übersteigt damit, wie zu Beginn dieses Kapitels angedeutet, den Rahmen der gegenwärtigen Arbeit und ist in den Bereich zukünftiger Arbeiten zu verweisen.

Für die gegenwärtige Arbeit ist also zunächst nur eine möglichst informative Ausgabe an den Benutzer vorgesehen. Sie soll den aufgetretenen Konflikt beschreiben, das heißt wiedergeben, an welchem Punkt des Ringaufbaus er auftrat (Fragmentaufbau, Ringschluß oder Verifikation) und worin er bestand (Fehlen von Kandidaten für den nächsten Aufbauschritt sowie Fehlschlag des Ringschlusses oder der Verifikation als Sonderfälle). Zur Dokumentation der entstandenen Konfliktsituation wird neben dem bis dahin aufgebauten Ringfragment auch die Liste der Ausgangsfragmente zusammen mit ihren zugehörigen Wahrscheinlichkeitsbewertungen ausgegeben.

Darüber hinaus ist zu bedenken, daß sichergestellt werden muß, daß tatsächlich ein Konflikt vorliegt und nicht nur die vorausgegangenen Ringfragmente in der falschen Reihenfolge zusammengesetzt wurden. Gibt es in einem oder in mehreren Aufbauschritten alternative Möglichkeiten, den Aufbau fortzusetzen, so müssen auch diese verfolgt werden. Das Resultat ist das größte erreichbare Fragment, das heißt der größte auf allen alternativen Wegen erreichbare Fortschritt im Aufbau des Substitutionsmusters. Die folgenden Funktionen wurden ergänzt oder modifiziert, um das Beschriebene zu realisieren:

- Ermitteln von Kandidaten für das Anfügen des nächsten Fragments: Die Funktion liefert wie bisher den bestbewerteten Kandidaten zurück, jedoch werden auch die übrigen Kandidaten für die Verfolgung alternativer Aufbauwege gespeichert.

- Finden einer Teillösung: Das auf Seite 133 als Aufbauen eines Benzolringes skizzierte Vorgehen muß nicht notwendigerweise in einer vollständige Substitutionsmusterhypothese, sondern kann auch in einer Teillösung in Gestalt des größten erzielten Ringfragments resultieren.

- Rückgängigmachen von Aufbauschritten. Das letzte angefügte Fragment wird wieder vom im Aufbau befindlichen Ring entfernt, und es wird überprüft, ob es einen alternativen Kandidaten gibt. Ist dies nicht der Fall, werden so lange weitere Fragmente entfernt, bis eine Alternative für den betreffenden Schritt zur Verfügung steht. Beginnend mit dieser kann der Ringaufbau auf einem alternativen Weg fortgesetzt werden.

- Verfolgen alternativer Aufbauwege: Wird zunächst kein vollständiges Substitutionsmuster gefunden, so werden alle alternativen Aufbauwege untersucht. Die Zwischenspeicherung der einzelnen Schritte des Ringaufbaus ermöglicht dabei deren Rückgängigmachen, falls der Aufbau nicht fortgesetzt werden kann. Das größte insgesamt zu erzielende Ringfragment (vollständige Lösung oder beste Teillösung) wird zurückgeliefert.

- Kommentierte Ausgabe. Die Funktion erhält die beste insgesamt gefundene Teillösung und gibt sie aus. Zusätzliche Angaben in Gestalt von Wahrscheinlichkeitswerten oder der aus den Klassifikationsergebnissen resultierenden Ausgangsfragmente kommentieren den Erfolg oder Fehlschlag des Ringaufbaus.

9.3 Integration der Einzelmodule

Wie in Kapitel 5 beschrieben sollen alle bis hierher entwickelten Module ineinandergreifen und so gemeinsam ein Musteranalysesystem bilden. Dies wurde bereits bei der Entwicklung berücksichtigt, so daß sich die jeweiligen Funktionalitäten unproblematisch miteinander zu dem Gesamtsystem SASCHA vereinen lassen.

Als ausführbares Programm („`SASCHA`“) dient dabei das Kontrollmodul. Es wird von der Kommandozeile aus gestartet, wobei der gewünschte Anwendungsfall (Parameteradaption, Spektrenauswertung oder Evaluation, vgl. Kapitel 5, insbesondere Abbildung 5.1) und gegebenenfalls weitere Parameter angegeben werden. Das Kontrollmodul aktiviert dann entsprechende Routinen und Abläufe innerhalb der jeweils benötigten Module sowie zwischen denselben. Die folgenden Abschnitte beschreiben die auswählbaren Funktionen und deren Parametrisierung unterschieden nach den drei Anwendungsfällen der Spektrenauswertung, Parameteradaption und Evaluierung.

9.3.1 Durchführung der Parameteradaption

Bevor SASCHA zur Spektrenauswertung eingesetzt werden kann, müssen, wie in Kapitel 8 beschrieben, die Wahrscheinlichkeiten des Bayes-Netzes festgelegt werden. Neben den entsprechenden Funktionalitäten ist auch die Möglichkeit, eine gegebene Stichprobe auf bestimmte strukturelle Merkmale hin zu untersuchen Teil des Statistikmoduls. Sie liefert Informationen, auf deren Grundlage das Modell, falls nötig, angepaßt werden kann und die auch zur Entwicklung des gegenwärtigen Modells benutzt wurden.

Innerhalb des Gesamtsystems erhält man durch die Auswahl von `Statistik` Zugang zu den entsprechenden Funktionen, die wie folgt parametrisiert sind:

- `Histogramm <Position> <Abstand>`
  Es wird eine Datei generiert, die die in der gegebenen Position (z.B. `IPSO`) und im gegebenen Abstand (z.B. `2` für Atome der 2. Sphäre) vorkommenden Atome und die absolute Häufigkeit ihres Vorkommens aufgelistet sind. Die Namen der zu untersuchenden Stichprobendateien werden von STDIN gelesen.

- `Dateiliste <Position> <Abstand> <Typen>`
  Es wird eine Datei generiert, die diejenigen Dateinamen auflistet, die Strukturen enthalten, welche in der gegebenen Position und im gegebenen Abstand die als `<Typen>` angegebene Besetzung aufweisen. Die Stichprobe wird wiederum von STDIN gelesen.

- `Ortho-Inkremente <Sphäre1> <S2ortho> <S2ipso> <Liste>`
  Die Funktion führt die zur Bestimmung der Wahrscheinlichkeit $P(\texttt{i2ortho}|\texttt{s2ortho})$ nötigen Untersuchungen durch und speichert die Ergebnisse. Die anzugebenden Dateien enthalten dazu benötigtes Literatur- bzw. Expertenwissen: Die Datei `<Sphäre1>` enthält die in der ersten Sphäre anzutreffenden Atomtypen und ihre korrespondierenden Inkremente, die Datei `<S2ortho>` enthält die für die *ortho*-Positionen möglichen Klassen, und in der Datei `<Liste>` ist eine Liste von Spektrendateien gespeichert. `<S2ipso>` gibt das gemeinsame Präfix der Namen derjenigen Dateien an, die die zu den möglichen *s2ipso*-Klassen korrespondierenden Inkremente enthalten.

- `Init <Auswahl>`
  Über diese Angabe können die Wahrscheinlichkeiten des Bayes-Netzes initialisiert werden. `<Auswahl>` besteht aus einer oder mehreren der folgenden Optionen:

  - `-p <Liste>` zur Berechnung von *a-priori*-Wahrscheinlichkeiten basierend auf den in <Liste> angegebenen Stichprobendateien.
  - `-i <Liste>` zur Berechnung der bedingten Wahrscheinlichkeiten der strukturbezogenen („inneren“) Variablen (`C_arom`, `s2ipso` und `s2ortho`) basierend auf den in <Liste> angegebenen Stichprobendateien. Alternativ führt die Auswahl von `-I` ohne weitere Parameter zu einer gleichverteilten Initialisierung.
  - `-1 <Datei>` zur Berechnung von $P(\texttt{Sphäre1}|\texttt{C\_arom})$ mithilfe des in `<Datei>` angegebenen Wissens.
  - `-2i <Datei>` zur Berechnung von $P(\texttt{i2ipso}|\texttt{s2ipso})$ mithilfe des in `<Datei>` angegebenen Wissens. Alternativ wählt `-2I <Datei>` die in Abschnitt 8.3.1 beschriebene weiche Initialisierungsvariante.
  - `-2o <Datei>` zur Berechnung von $P(\texttt{i2ortho}|\texttt{s2ortho})$ mithilfe des in `<Datei>` angegebenen Wissens. `-2O <Datei>` steht wiederum für eine weiche Initialisierung.
  - `-s <Zahl>` zur Berechnung der mithilfe der „weichen Summenbildung“ (vgl. Abschnitt 8.3.3) zu bestimmenden bedingten Wahrscheinlichkeiten $P(\texttt{Sphäre2}|\texttt{i2ipso},\texttt{i2ortho})$ und $P(\texttt{peakpos}|\texttt{Sphäre1},\texttt{Sphäre2})$. `<Zahl>` ist eine Gleitkommazahl, die die „Weichheit“ der Summenbildung parametrisiert.

9.3.2 Auswertung von Spektren

Die Auswertung von Spektren erfolgt in zwei getrennten Schritten für Klassifikation der einzelnen Positionen und Hypothesengenerierung zum Aufbau des Substitutionsmusters. Die beiden Module werden von einander getrennt gehalten; ihre Verzahnung kann sich mit der Entwicklung einer Revisionsstrategie in zukünftigen Arbeiten noch verändern. Insbesondere erlaubt dieses Vorgehen aber auch ein isoliertes Betrachten der Klassifikation als Einzelschritt, was hinsichtlich der Evaluation des Klassifikators vor dem Hintergrund der Wissensrepräsentation als Schwerpunkt der gegenwärtigen Arbeit vorteilhaft ist.

Durch die Angabe `Klassifikator` erhält man innerhalb des Gesamtsystems Zugang zu den folgenden Funktionalitäten im Zusammenhang der Spektrenauswertung:

- `Ipso-Klasse <BNIF> <JCAMP>`
  Diese Funktion veranlaßt die Klassifikation der sechs Peaks der Benzolringatome hinsichtlich der *ipso*-Positionen. Als Eingabe dient dabei die als `<JCAMP>` angegebene Datei, welcher die benötigte Summenformel- und Peakinformation entnommen wird.

Entsprechende Evidenzen werden in das Bayes-Netz, das in Gestalt der Datei <BNIF> gegeben ist, eingetragen. Die wahrscheinlichste Klasse wird zusammen mit ihrer *a-posteriori*-Wahrscheinlichkeit gespeichert.

- `Ortho-Klasse <BNIF> <JCAMP>`
  Die Funktion leistet dasselbe wie vorgenannte, jedoch findet die Klassifikation hinsichtlich der *ortho*-Positionen statt. Außerdem wird die korrekte *ipso*-Klasse als zusätzliche Evidenz der JCAMP-Datei entnommen.

- `Ipso-Ortho-Klasse <BNIF> <JCAMP>`
  Die Funktion leistet dasselbe wie vorgenannte, jedoch wird die als zusätzliche Evidenz einzutragende *ipso*-Klasse nicht der JCAMP-Datei entnommen. Statt dessen findet zuvor eine Klassifikation der *ipso*-Position statt, deren Ergebnis für den Eintrag der zusätzlichen Evidenz verwendet wird. Beide wahrscheinlichsten Klassen werden zusammen mit ihren *a-posteriori*-Wahrscheinlichkeiten gespeichert.

- `Ipso-Ortho-Verteilung <BNIF> <JCAMP>`
  Die Funktion leistet prinzipiell dasselbe wie vorgenannte, jedoch liefert sie nicht nur die wahrscheinlichste Klasse, sondern zwei gesamte Verteilungen: zum einen die Verteilung bei der Anfrage nach der *ipso*-Position und zum anderen die Verteilung bei der Anfrage nach der *ortho*-Position, bei welcher die wahrscheinlichste *ipso*-Klasse als zusätzliche Evidenz verwendet wurde.

Über die Angabe `Strukturgenerator` stehen anschließend die Funktionalitäten der Hypothesengenerierung zur Verfügung. Bislang ist hier nur eine Funktionalität implementiert, die den Versuch des Aufbaus eines Substitutionsmusters betrifft und den Erfolg oder Mißerfolg des Versuches kommentiert. Weitere Funktionalitäten werden mit der Entwicklung einer Revisionsstrategie folgen, die es ermöglichen soll, auf einen anfänglichen Fehlschlag des Ringaufbaus zu reagieren – dies fällt jedoch in den Bereich zukünftiger Arbeiten.

- `Kommentiert <BNIF> <ortho> <Verteilungen>`
  Die Funktion versucht den Aufbau eines Substitutionsmusters aus den Klassifikationsresultaten der Peaks der sechs Benzolringatome. Als Eingabe dienen dabei die Verteilungen von `C_arom` und `s2ortho` nach dem *ipso*- bzw. *ortho*-Klassifikationsschritt, die der als Parameter angegebenen Datei `<Verteilungen>` entnommen werden. Diese referenziert die Klassen jeweils als Index des betreffenden Zustandes, dessen Bezeichner dem als `<BNIF>` angegebenen Bayes-Netz zu entnehmen ist. Die Datei `<ortho>` dient dazu, die jeweils auf einen Zustand der Variablen `s2ortho` abgebildeten Substituentenkombinationen aufzuschlüsseln, welche während des Aufbauens des Substitutionsmusters zu beachten sind. Sofern der Benzolring und damit das Substitutionsmuster vollständig aufgebaut werden kann, wird dieses Ergebnis ausgegeben, anderenfalls das beste erreichbare Teilergebnis (das heißt das größte erreichbare Ringfragment) zusammen mit einem Kommentar, wodurch die Fortführung des Ringaufbaus scheiterte.

9.3.3 Evaluierung des Klassifikators

Durch die Trennung des Klassifikations- und des Hypothesengenerierungsmoduls ist, wie bereits im vorigen Abschnitt erwähnt, eine Evaluierung des Klassifikators für sich genommen besonders unproblematisch möglich. Die Evaluierung selbst wird ausführlich in Kapitel 10 behandelt. An dieser Stelle sollen lediglich die zu diesem Zweck zur Verfügung stehenden

Funktionen beschrieben werden. Sie sind über die Wahl von `Klassifikator` zugänglich und wie folgt parametrisiert:

- `Ipso-Eval <BNIF>` Diese Funktion veranlaßt die Klassifikation der sechs Peaks der Benzolringatome hinsichtlich der *ipso*-Positionen. Das in Gestalt der Datei `BNIF` gegebene Bayes-Netz dient dabei als Grundlage. Nacheinander wird eine über STDIN gegebene Liste von JCAMP-Dateien verarbeitet. Ihnen wird die benötigte Summenformel- und Peakinformation sowie aus der enthaltenen Strukturinformation zu Vergleichszwecken auch die korrekte Klasse entnommen. Für jeden Peak wird das Klassifikationsergebnis, die *a-posteriori*-Wahrscheinlichkeit der entsprechenden Klasse und die korrekte Klasse gespeichert.

- `Ortho-Eval <BNIF>` Die Funktion leistet dasselbe wie obige, jedoch findet die Klassifikation hinsichtlich der *ortho*-Positionen statt. Außerdem wird die jeweils korrekte *ipso*-Klasse als zusätzliche Evidenz der JCAMP-Datei entnommen.

9.4 Ergebnisse

Mit dem in diesem Kapitel Beschriebenen stehen alle für das Musteranalysesystem SASCHA nötigen Module zur Verfügung. Es wurde ein Konzept vorgestellt, wie aus dreigliedrigen Ringfragmenten, welche der Klassifikationsschritt des Systems liefert, Hypothesen für das Substitutionsmuster generiert werden können. Dabei wird die Überlappung der Fragmente ausgenutzt, um sukzessive den Benzolring aufzubauen, an welchem sodann das Substitutionsmuster abgelesen werden kann. Entsprechendes wurde in einem Modul zur Hypothesengenerierung realisiert.

Im Fall von Konflikten während des Ringaufbaus wird eine geeignete Strategie zur systematischen, gezielten Revision von Klassifikationsergebnissen benötigt. Sie fällt in den Bereich des Kontrollmoduls, welches in dieser Form zu realisieren jedoch den Rahmen der gegenwärtigen Arbeit aufgrund der Vielschichtigkeit der dazu notwendigen Überlegungen weit übersteigt. Vielmehr wird am Schwerpunkt der Modellentwicklung, das heißt der Repräsentation von Wissen, festgehalten. Bei der Hypothesengenerierung auftretende Konflikte werden gemeinsam mit den in ihrem Zusammenhang interessanten Wahrscheinlichkeitswerten und Hypothesenrängen ausgegeben. Auf diese Weise wird das derzeit recht elementare Kontrollmodul zu einem nützlichen Werkzeug für die Entwicklung einer intelligenten Revisionsstrategie im Zuge zukünftiger Arbeiten.

Weiterhin wurde beschrieben, wie das entwickelte System praktisch eingesetzt wird. Seine Funktionalitäten, die dem in den vorangegangenen Kapiteln konzipierten entsprechen, und wie sie angesprochen werden, wurden zusammengefaßt. Die einzelnen Module greifen im Rahmen des Gesamtsystems SASCHA wie Zahnräder ineinander und realisieren erst durch ihre Integration das angestrebte Ziel. Insbesondere bleiben Klassifikation und Hypothesengenerierung dabei der Definition der einzelnen Module entsprechend funktionell von einander getrennt. Auf diese Weise soll zum einen die in zukünftigen Arbeiten zu realisierende Revisionsstrategie für die Strukturgenerierung reibungslos integrierbar sein, und zum anderen kann so der Klassifikator, in dessen Leistung sich unmittelbar die Qualität der Wissensrepräsentation widerspiegelt, für sich genommen evaluiert werden. Dies ist der nächste durchzuführende Schritt, welcher unmittelbar zu den Ergebnissen der gegenwärtigen Arbeit führt.

10 Evaluierung

Für die Erkennung von Substitutionsmustern an Benzolderivaten wurde, wie in den vorangegangenen Abschnitten beschrieben, ein Bayes-Netz entwickelt, welches die Kausalzusammenhänge der Auswertung von ^{13}C-NMR-Spektren modelliert. Es dient dazu, die sechs Positionen des Benzolringes im einzelnen zu klassifizieren. Im darauffolgenden Schritt wird aus den Einzelresultaten der Benzolring zusammengesetzt, so daß das gesuchte Substitutionsmuster abgelesen werden kann. Treten dabei aufgrund von Klassifikationsfehlern Schwierigkeiten auf, so kann nur ein Teilergebnis ausgegeben werden.

Die Leistung des Klassifikationsschrittes ist für das Gesamtsystem von zentralem Interesse: Auf der Seite der Entwicklung ist die Wissensrepräsentation der gesetzte Schwerpunkt, und da das dem System zugrundeliegende Wissen in Gestalt des zur Klassifikation genutzten Bayes-Netzes niedergelegt ist, bezieht sich die Evaluierung des Klassifikationsschrittes unmittelbar auf diesen Schwerpunkt. Mit Blick auf den praktischen Einsatz zur Spektrenauswertung ist außerdem anzuführen, daß die Klassifikationsresultate als Eingabe für die Hypothesengenerierung dienen und ihre Qualität somit auch einen starken Einfluß auf das Endergebnis des gesamten Musteranalyseprozesses hat.

Im folgenden wird daher das entwickelte Bayes-Netz in seiner Eigenschaft als Klassifikator evaluiert. Da die Klassifikation in zwei Teilschritten stattfindet, wobei zuerst die wahrscheinlichste strukturelle Klasse der *ipso*-Position ermittelt wird und dann gegeben diese Klasse die benachbarten *ortho*-Positionen klassifiziert werden, ist auch die Evaluierung unter diesen beiden Aspekten durchzuführen. Nach einigen allgemeinen Bemerkungen zur Durchführung der Evaluation in Abschnitt 10.1 sind diese Auswertungen Gegenstand des Abschnitts 10.2. Mit Blick auf eine individuelle Bewertung der einzelnen Modellierungsentscheidungen sowie die Verbesserung des Kausalmodells werden außerdem in Abschnitt 10.3 verschiedene Varianten des Modells in Betracht gezogen und unter den gleichen Aspekten evaluiert. Abschnitt 10.4 resümiert die so gewonnenen Erkenntnisse vor dem Hintergrund des Schwerpunkts der Wissensrepräsentation.

10.1 Allgemeine Bemerkungen

Die Evaluierung des entwickelten Bayes-Netzes als Klassifikator hat zweierlei zum Ziel: Einerseits ist es natürlich interessant, welche Leistungsfähigkeit mit der gewählten Modellierung erreicht werden kann, andererseits ist es aber auch möglich, durch wechselweisen Vergleich unterschiedlicher Modellierungen dieselben zu bewerten, und zwar bei entsprechendem Vorgehen bis auf die Ebene grundsätzlicher Modellierungsentscheidungen. Auf diese Weise werden wertvolle Erfahrungen konkretisiert und zusammengeführt, die für die Weiterentwicklung des Modells wie auch des Gesamtsystems genutzt werden können.

Ehe sich die folgenden Abschnitte mit diesen beiden Aspekten, Klassifikationsleistung und Bewertung von Modellierungsentscheidungen, beschäftigen, sollen an dieser Stelle einige prinzipielle Punkte angesprochen werden. Dabei steht die verwendete Stichprobe im Fokus des Interesses, da ihre Beschaffenheit, aber auch ihre Nutzung, eine Rolle bei der Bewertung der Resultate spielt.

Grundsätzlich stehen verschiedene Kenngrößen zur Verfügung, um die Leistung eines Klassifikators zu bewerten. Betrachtet man für ein Zwei-Klassen-Problem die Zahl der innerhalb einer Stichprobe korrekt positiv (p), falsch positiv (fp), korrekt negativ (n) und falsch negativ (fn) klassifizierten Beispiele, so lassen sich die folgenden Kenngrößen berechnen:

$$\text{Sensitivität:} \quad \frac{p}{p+fp} \tag{10.1}$$

$$\text{Spezifität:} \quad \frac{n}{n+fn} \tag{10.2}$$

$$\text{Relevanz:} \quad \frac{p}{p+fn} \tag{10.3}$$

$$\text{Segreganz:} \quad \frac{n}{n+fp} \tag{10.4}$$

$$\text{Effizienz:} \quad \frac{n+p}{p+fp+n+fn} \tag{10.5}$$

$$\text{Fehlerrate:} \quad \frac{fp+fn}{p+fp+n+fn} \tag{10.6}$$

Sensitivität und Spezifität geben den Anteil der korrekt klassifizierten unter allen positiv bzw. negativ klassifizierten Beispielen wieder. Relevanz (positiver prädiktiver Wert) und Segreganz (negativer prädiktiver Wert) beschreiben ebenfalls den Anteil korrekt klassifizierter Beispiele, jedoch bezogen auf alle tatsächlich positiven bzw. tatsächlich negativen Beispiele. Die Effizienz (Korrektklassifikationsrate) gibt den Anteil korrekter Resultate bezogen auf die Gesamtmenge der Beispiele an, die Fehlerrate den Anteil falscher Resultate in der Gesamtmenge.

Die gegebene Klassifikationsaufgabe als ein 14-Klassen-Problem läßt sich prinzipiell auf vierzehn Zwei-Klassen-Probleme zurückführen, so daß sich jede der beschriebenen Kenngrößen berechnen ließe, es stellt sich jedoch die Frage, welche davon für die Evaluation zu wählen ist. Sie sollte idealerweise die Leistungsfähigkeit des Klassifikationsschrittes so bewerten, daß sie auch im Kontext des Gesamtsystems, z.B. mit Blick auf die Entwicklung einer Revisionsstrategie im Rahmen der Hypothesengenerierung, als Gütemaß für den Klassifikator dient. Vor dem Hintergrund, daß die Klassifikationsergebnisse im Anschluß der Hypothesengenerierung übergeben werden, ist hier sicherlich die Frage der Zuverlässigkeit am interessantesten. Eine Kenngröße, die dies wiedergibt, ist die Effizienz (Gleichung (10.5)), aber auch die Fehlerrate (Gleichung (10.6)), welche die die Tendenz des Klassifikators wiedergibt, Fehler zu machen, und sich mit der Effizienz zu 1 ergänzt. Je niedriger diese Tendenz, desto „selbstsicherer“ kann das System gestaltet werden, das heißt um so zurückhaltender sollten im Falle von Konflikten in der Hypothesengenerierung Revisionen der Klassifikationsergebnisse vorgenommen werden. Dies ist eine wichtige Information für die Entwicklung einer Revisions- und Kontrollstrategie, wie sie in zukünftigen Arbeiten geschehen kann. Eine niedrige Fehlerrate ist zudem zum aktuellen Zeitpunkt um so bedeutsamer, da Fehlklassifikationen derzeit (ohne eine Revisionsstrategie) noch in einem Fehlschlag der Hypothesengenerierung resultieren.

Die Stichprobe, welche zur Evaluation herangezogen wird, wurde freundlicherweise von der BASF AG zur Verfügung gestellt. Es handelt sich dabei um reale, im Labor aufgenommene Daten. Zu jeder untersuchten Substanz sind Spektrum und Molekülstruktur sowie eine Zuordnung der Peaks zu den verursachenden Kohlenstoffatomen gegeben. Bei allen Proben handelt es sich um Verbindungen, die einen Benzolring enthalten, jedoch wird gemäß der folgenden beiden Kriterien eine Auswahl verwendbarer Datensätze vorgenommen: Zum ersten

werden nur solche Moleküle betrachtet, bei denen das aromatische System nicht über den Benzolring hinausgeht. Anderenfalls könnte der Benzolring bezogen auf die Betrachtung der Elektronenverteilung im Molekül schwerlich vom übrigen System getrennt betrachtet werden. Zweitens dürfen keine Zyklen von einer Position des Benzolringes zu einer anderen vorkommen. Bislang gibt es keine Strategie zum Umgang mit einer solchen Konstellation bei der Betrachtung der Einflüsse der betreffenden Atome, angefangen bei der Frage, welcher Position die an einem solchen Ringschluß beteiligten Atome zuzuordnen wären.

Insgesamt stehen somit 903 Verbindungen mit insgesamt 5418 zu klassifizierenden Kohlenstoffatomen zur Verfügung. Über ihre Funktion als Testmenge hinaus dient diese Stichprobe jedoch auch der Gewinnung derjenigen bedingten Wahrscheinlichkeiten des Bayes-Netzes, welche empirisch initialisiert werden sollen, so daß die Gesamtmenge in eine Initialisierungs- und eine Testmenge zu unterteilen ist. Bei den üblicherweise gewählten Größenverhältnissen beider Mengen von 3:1 kommt man zu 678 Verbindungen (4068 Benzolringatomen) für die empirische Initialisierung und 225 Verbindungen (1350 Benzolringatomen) für die Testmenge.

Setzt man die Zahl der Trainingsbeispiele in Bezug zu den benötigten bedingten Wahrscheinlichkeiten, so wird klar, daß ihre Zahl sehr knapp bemessen ist: für die Variablen `C_arom` und `s2ortho` werden jeweils 128 Wahrscheinlichkeitsverteilungen benötigt – für das Schätzen jeder Verteilung stehen also im Schnitt 31,8 Trainingsbeispiele zur Verfügung, während die Verteilungen aber jeweils 14 Zustände für `C_arom` bzw. 22 Zustände für `s2ortho` abdecken.

Um so bedeutsamer erscheint vor diesem Hintergrund die getroffene Modellierungsentscheidung einer binären Repräsentation der Summenformelinformation, bei der nur Anwesenheit oder Abwesenheit der betrachteten chemischen Elemente wiedergegeben werden (vgl. Abschnitt 6.2.1). Würde nur für eine einzige Variable eine Realisierung mit 3 Zuständen gewählt, so wären 192 anstelle von 128 Verteilungen zu schätzen, und die durchschnittliche Zahl von Beispielen pro Verteilung fiele auf etwa 21,2: Damit stünde bereits nicht einmal mehr ein Beispiel je infragekommendem Zustand zur Verfügung, wenn jeder der 22 Zustände von `s2ortho` aufgrund der Summenformel theoretisch möglich ist.

Betrachtet man die Variable `s2ipso`, so werden aufgrund von deren zusätzlicher Abhängigkeit von `C_arom` neben den Summenformelvariablen insgesamt 1792 Verteilungen benötigt. Im Schnitt stehen dabei je Verteilung also nur etwa 2,3 Beispiele zur Verfügung. Dies ist mit Sicherheit ein nicht unerheblicher Unsicherheitsfaktor für die Klassifikation und wird die Ergebnisse negativ beeinflussen. Im Zuge der Entwicklung von Modellvarianten oder generell einer Weiterentwicklung des Modells sollte dieser Punkt nicht unberücksichtigt bleiben.

Aufgrund der knappen Datenlage wird zudem ein 10fach-Kreuzvalidierungsverfahren für die Evaluierung gewählt: Die Stichprobe wird dazu zunächst in zehn Teilmengen untergliedert, neun davon bilden die Trainingsmenge, die verbleibende dient als Testmenge. Die Evaluierung wird insgesamt zehn Mal durchgeführt, wobei jeweils eine andere der zehn Teilmengen die Rolle der Testmenge übernimmt. Vor jedem Durchlauf müssen dementsprechend auch die empirischen Parameter des Bayes-Netzes angepaßt werden.

Dieses Verfahren ist bei einer knappen Datenlage besonders günstig: Insbesondere in solchen Fällen besteht die Gefahr, durch eine unvorteilhafte Auswahl von Datensätzen für die Testmenge schlechte Resultate zu erhalten, die nicht die tatsächliche Leistungsfähigkeit des getesteten Systems widerspiegeln. Dies vermeiden *k-fach-Kreuzvalidierungsverfahren*[1], da systematisch unterschiedliche Testmengen für die Evaluierung herangezogen werden und aus den jeweiligen Ergebnissen der Durchschnitt gebildet wird.

[1] Der gewählte Wert von 10 für k ist ein typischerweise benutzter Wert.

Hinsichtlich der Erwartungen an die zu erzielende Leistungsfähigkeit sind dabei drei Punkte zu bemerken: Zum ersten macht die knappe Datenlage es wahrscheinlich, daß die Generalisierungsfähigkeit des Modells eingeschränkt ist, was in einer höheren Fehlerrate resultiert. Zweitens sei darauf hingewiesen, daß das Ziel der Untersuchung verschiedener Modellvarianten nicht hauptsächlich im Erreichen höchster Korrektklassifikationsraten besteht, sondern daß vielmehr deren Änderung bei der Änderung bestimmter Modellierungsentscheidungen von Interesse ist, um effektive und weniger effektive Maßnahmen von einander abzugrenzen.

In diesen Zusammenhang reiht sich unmittelbar auch der dritte Punkt: Die Modellierungsgenauigkeit beschränkt sich bei der Strukturrepräsentation auf eine Umgebung von zwei Sphären um das Atom, dessen Peak betrachtet wird. Andere Systeme, die eine Kombination aus Strukturgenerator und Spektrenvorhersage verwenden (z.B. MOLGEN in [MMW00], [MW01]) benutzen dagegen fünf Sphären zuzüglich einer weiteren für alle entfernteren Atome. Insofern sind die Erwartungen an die absoluten Werte, die bei der Evaluation erreicht werden, sicherlich zu dämpfen. Es sei jedoch noch einmal darauf hingewiesen, daß bewußt zunächst ein eingeschränkteres, dadurch jedoch überschaubareres Modell gewählt wurde, um Erfahrungen zu sammeln und die zugrundegelegten Prinzipien der Modellierung beurteilen zu können. Diese Erkenntnisse sind von großem Wert für alle zukünftigen Verfeinerungen.

Im folgenden Abschnitt wird nun sowohl der *ipso*- als auch der *ortho*-Klassifikationsschritt basierend auf dem in Kapitel 6 entwickelten Modell evaluiert. Dabei wird jeweils der Anteil falsch klassifizierter Einzelpositionen für einige elementare Initialisierungsvarianten ermittelt, und die Ergebnisse werden einander gegenübergestellt. Im darauffolgenden Abschnitt schließt sich die Entwicklung und Evaluierung verschiedener Modellvarianten sowie die Interpretation der dabei gewonnenen Ergebnisse vor dem Hintergrund der Bewertung einzelner Modellierungsschritte und -entscheidungen an.

10.2 Klassifikation der *ipso*-Position und der *ortho*-Positionen

Zur Ermittlung der Klassifikationsleistung für das entwickelte Modell wird ein 10fach-Kreuzvalidierungsverfahren verwendet. Die Stichprobe wird dabei in zehn Teile unterteilt. Das Bayes-Netz wird zunächst initialisiert, wobei die Bestimmung empirisch zu ermittelnder Wahrscheinlichkeiten auf neun Teilen der Stichprobe erfolgt. Klassifikation sowohl der *ipso*-Position als auch der *ortho*-Positionen werden dann auf dem verbleibenden zehnten Teil untersucht. Die Ergebnisse beider Klassifikationsschritte sind interessant, da erst beide zusammen zu dreigliedrigen Ringfragmenten führen, welche der Ausgangspunkt für das Zusammensetzen des Gesamtsubstitutionsmusters sind.

Als Eingabe für die Klassifikation der *ipso*-Position dienen die Position des Peaks des entsprechenden Ringatoms sowie das Vorhandensein oder Fehlen der chemischen Elemente Sauerstoff, Schwefel, Stickstoff, Fluor, Chlor, Brom und Iod. Für die *ortho*-Positionen geht neben der Position des Peaks des *ipso*-Kohlenstoffs und der Information über die chemischen Elemente die *ipso*-Klasse als Evidenz in die Klassifikation ein. Während im Fall der regulären Spektrenauswertung die zuvor ermittelte wahrscheinlichste *ipso*-Klasse hierfür genutzt wird, ist bei der Evaluierung ein anderes Vorgehen sinnvoll. Um die Klassifikationsleistung unabhängig von möglichen vorhergehenden Fehlern zu untersuchen, wird die korrekte *ipso*-Klasse in diesem Fall den dem Spektrum beigefügten Strukturdaten entnommen.

Tabelle 10.1 zeigt die Klassifikationsleistung hinsichtlich der *ipso*-Position und der *ortho*-Positionen. Dabei werden unterschiedliche Initialisierungen des Bayes-Netzes einander gegenübergestellt: Zum ersten können die bedingten Wahrscheinlichkeiten der strukturbezogenen Variablen gleichverteilt angenommen oder empirisch ermittelt werden. Der Vergleich

beider Varianten gibt Aufschluß darüber, welchen Einfluß Faktoren, die vom Modell nicht explizit erfaßt werden, die sich jedoch quasi „versteckt“ in statistischen Zusammenhängen widerspiegeln, auf das Klassifikationsergebnis haben. In Tabelle 10.1 (und allen folgenden Tabellen) werden diese beiden Fälle anhand der Begriffe „empirisch“ und „gleichverteilt“ unterschieden.

Zweitens kann ein „harter“ oder „weicher“ Übergang zwischen den strukturbezogenen Variablen und den zugehörigen (Teil-)Inkrementen angenommen werden. „Hart“ bedeutet, daß nur jeweils genau ein möglicher Wert gegeben ein bestimmtes Strukturmerkmal eine Wahrscheinlichkeit von $1,0$ erhält, alle anderen erhalten die Wahrscheinlichkeit $0,0$. Somit ist jedem Strukturmerkmal ein zugehöriger Inkrementwert fest zugeordnet. Bei einer weichen Initialisierung erhalten auch andere Werte Wahrscheinlichkeiten $> 0,0$. Diese Wahrscheinlichkeiten sind jedoch niedrig gewählt, da die betreffenden Ausprägungen des abhängigen Ereignisses zwar nicht ausgeschlossen werden, aber verglichen mit dem gemäß Literatur- oder empirischem Wissen korrekten korrespondierenden Inkrement unwahrscheinlicher sein sollen.

Schließlich kann zum dritten noch die Parametrisierung der weichen Summenbildung (vgl. Abschnitt 8.3.3) beim Zusammenführen der Inkremente und Teilinkremente variiert werden: Eine „Weichheit“ von $0,0$ bedeutet, daß lediglich die Rundungsfehler der zusammengeführten Inkremente oder Teilinkremente ausgeglichen werden, während Werte $> 0,0$ ein Toleranzintervall angeben, innerhalb dessen vom exakten Additionsergebnis abweichende Werte als „richtig“ angesehen werden. Die Begriffe „exakt“ und „tolerant“ kennzeichnen die beiden Fälle in den Tabellen. Bei der „toleranten“ Initialisierung wurde eine Intervallgröße von 10,0 gewählt.

Strukturvariablen	Inkremente	weiche Summe	Fehlerrate *ipso*	Fehlerrate *ortho*
empirisch	hart	exakt	26,91%	83,90%
empirisch	hart	tolerant	26,89%	85,69%
empirisch	weich	tolerant	27,04%	86,75%
empirisch	weich	exakt	26,72%	84,84%
gleichverteilt	weich	exakt	82,08%	90,01%
gleichverteilt	weich	tolerant	81,31%	92,41%
gleichverteilt	hart	tolerant	80,97%	90,86%
gleichverteilt	hart	exakt	81,36%	88,03%

Tab. 10.1: *Klassifikationsleistung bei unterschiedlicher Initialisierung des Bayes-Netzes*

Die Fehlerrate in Tabelle 10.1 wie auch in allen folgenden Tabellen ergibt sich als Durchschnittswert der gemäß Gleichung 10.6 berechneten einzelnen Fehlerraten der zehn Kreuzvalidierungsdurchgänge. Sie ist auf der Ebene einzelner Benzolringatome, nicht auf der Ebene verschiedener Moleküle, zu verstehen. Bei ihrer Bewertung ist zu beachten, daß das hier untersuchte, initiale Modell als Ausgangspunkt weitergehender Untersuchungen zu verstehen ist. Bereits in Kapitel 6 wurden während seiner Entwicklung Aspekte aufgezeigt, die vielversprechende alternative Modellierungsentscheidungen zulassen. Es wurde jedoch zunächst eine möglichst einfache Repräsentation angestrebt, um den Einfluß jedes weiteren Entwicklungsschrittes beurteilen zu können.

Gleichwohl steht fest, daß die ermittelten Fehlerraten zu großen Teilen nicht akzeptabel sind. Eine Erkennung der *ortho*-Positionen kann mit dem gegenwärtigen Modell quasi als unmöglich bezeichnet werden. Trotz dieser offensichtlichen Unzulänglichkeit innerhalb des Modells gelingt jedoch die Klassifikation der *ipso*-Position bei entsprechender Parametrisie-

rung des Netzes, nämlich sofern der Zusammenhang zwischen Struktur- und Summenformelinformation empirisch initialisiert wird. Auch die Fehlerrate betreffend die *ortho*-Klassen ist in diesen Fällen 5 bis 6% (absolut) niedriger.

Es liegt nicht fern zu vermuten, daß bei dieser Konstellation die *a-priori*-Wahrscheinlichkeiten der *ipso*-Klassen einen zu dominanten Einfluß auf das Klassifikationsergebnis haben. Dies würde bedeuten, daß zumeist auf ein und dieselbe Klasse geschlossen würde. Betrachtet man die einzelnen falsch klassifizierten Beispiele, so ist es jedoch nicht etwa so, daß bei allen auf ein und dieselbe Klasse geschlossen wurde, sondern es ist ein klarer Zusammenhang zwischen infragekommenden Klassen und der Lage des Peaks zu erkennen. Dies bedeutet, daß der Zusammenhang zwischen den Klassen und bestimmten für sie typischen Intervallen durchaus erkannt wird; läßt die chemische Verschiebung jedoch mehrere mögliche Schlüsse zu, so kommt es leicht zu Verwechslungen zugunsten der häufigeren Klasse.

Werden nun die Klassifikationsergebnisse der Testbeispiele weiter im einzelnen betrachtet, so ist zu bemerken, daß die resultierende Wahrscheinlichkeit der bestbewerteten Klasse häufig relativ niedrig ist: In etwa einem Siebtel der Fälle liegt sie, teilweise sogar sehr deutlich, unter 50%. Umgekehrt gibt es aber einige auffallend hohe Wahrscheinlichkeiten bei falschen Klassifikationsergebnissen. Dies spricht deutlich für das Wirken von Einflüssen, die vom Modell nicht erfaßt werden. Vor allem der Einfluß entfernterer Atome als jenen der zweiten Sphäre kommt hier in Betracht, aber auch eine nicht realitätsgetreue Erfassung des Zusammenwirkens der bereits im Modell erfaßten Einflüsse.

Als letzter Punkt bleibt noch zu bemerken, daß die Erkennungsraten von einem Kreuzvalidierungsdurchgang zum anderen merklich schwanken. Beispielsweise erreichen Durchgang 5 und 6 bei „empirischer“ Initialisierung eine Fehlerrate um 55% in der *ortho*-Klassifikation, die somit 30% (absolut) unter dem Durchschnittswert aller Durchgänge liegt. Zwischen den übrigen Durchgängen gibt es jedoch nur leichte Schwankungen, wie Tabelle 10.2 exemplarisch für die „empirische, weiche, exakte“ Initialisierung zu entnehmen ist.

Durchgang	0	1	2	3	4	5	6	7	8	9
Fehlerzahl *ipso*	133	145	146	155	148	146	143	143	140	149
Fehlerzahl *ortho*	487	505	496	504	500	**305**	**310**	501	491	499
Anzahl Beispiele	540	546	546	546	540	540	540	540	540	540

Tab. 10.2: *Zahl der Fehlklassifikationen in den einzelnen Kreuzvalidierungsdurchgängen für die Initialisierungsvariante „empirisch, weich, exakt“ des Bayes-Netzes*

Ein analoges Verhalten der *ipso*-Klassifikationsergebnisse ist jedoch nicht zu beobachten, was ebenfalls für die oben bereits angedeutete Vermutung spricht: Die Zusammenhänge innerhalb des Modells sind nicht nur hinsichtlich der auf zwei Sphären beschränkten strukturellen Umgebung zu ungenau, sondern auch insgesamt zu vage wiedergegeben. Zwar werden trotzdem im Falle einer empirischen Initialisierung die Tatsachen besser widergespiegelt als durch die Annahme einer Gleichverteilung, jedoch ist die Erkennungsleistung insgesamt nicht sehr zuverlässig.

Eine ungünstige Datenlage in Gestalt einer zu geringen Datenmenge ist hier mit Sicherheit eine der möglichen Ursachen. Unter dem Aspekt der eingeschränkten Modellierungsgenauigkeit durch die auf zwei Sphären beschränkte Strukturrepräsentation bietet sich jedoch auch eine alternative Evaluierung des Modells an, und zwar unter Verwendung einer Auswahl der Stichprobe, die besser mit der Modellierungsgenauigkeit übereinstimmt.

Hier kommen natürlich solche Verbindungen in Frage, die möglichst wenige Einflüsse aufweisen, die das Modell bislang nicht abdeckt, also solche, die nicht oder nur wenig über die im Modell erfaßten Atome der zweiten Sphäre hinausgehen. Vorrangig bieten sich hier monosubstituierte Benzolderivate wie in Abbildung 10.1 an, da sie die meisten Beispiele beisteuern, in welchen die Einflüsse etwaiger *meta*- und *para*-Substituenten ausgeschlossen sind (Klassifikation der Positionen a, b und f) oder in denen möglichst wenige derartige Einflüsse auftreten (jeweils nur ein nicht erfaßter Einfluß in den übrigen Fällen). In der Stichprobe gibt es 88 monosubstituierte Verbindungen (entsprechend 528 Einzelpeaks).

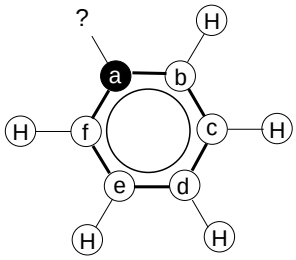


Abb. 10.1: *Schematische Darstellung eines monosubstituierten Benzolderivats. Die Ringpositionen sind mit a–f statt mit C für Kohlenstoff, der Substituent ist mit einem Fragezeichen bezeichnet.*

Eine andere Möglichkeit besteht in der Auswahl von Verbindungen, die nur kurze Substituenten tragen, die also innerhalb des einzelnen Substituenten nicht oder nur wenig über die Modellierungsgenauigkeit hinausgehen. Eine Untersuchung der Stichprobe ergab nur 13 Verbindungen, die Substituenten der maximalen Länge 2 enthielten, jedoch 138 Verbindungen mit Substituenten der maximalen Länge 3. Das letztere entspricht 828 Einzelpeaks und wird im folgenden als Stichprobe mit „kurzen Substituenten“ bezeichnet. Die Einflüsse von *meta*- und *para*-Substituenten werden hier gleichwohl in derselben Weise vernachlässigt wie auf der Gesamtstichprobe. Etwaige Unterschiede könnten dennoch aufschlußreich sein.

In beiden Fällen, bei der Auswahl von Verbindungen mit kurzen Substituenten wie auch bei der Auswahl von monosubstituierten Verbindungen, ist aufgrund der selbst im Fall der Gesamtstichprobe schon geringen Datenlage nicht unbedingt mit einer immensen Verbesserung zu richten, doch könnten sich sowohl hier als auch im Vergleich mit späteren Varianten des Modells interessante Tendenzen zeigen.

Betrachtet man hier die Ergebnisse der Auswahl ausschließlich monosubstituierter Verbindungen, so wird für den Fall der „empirischen“ Initialisierung deutlich, daß bei übereinstimmender Modellierungsgenauigkeit in der Tat erheblich bessere Erkennungsergebnisse erzielt werden können. Dies ist der Fall, obwohl, wie sich in den stark schwankenden Fehlerzahlen beim *ortho*-Klassifikationsschritt widerspiegelt, die Stichprobe deutlich zu klein für die empirische Gewinnung der hohen Anzahl benötigter empirischer Verteilungen ist.

Diese Verbesserung läßt sich jedoch nicht auf Verbindungen mit kurzen Substituenten übertragen. Dies ist damit zu begründen, daß nicht alle der nicht explizit modellierten Einflüsse einen gleich starken Einfluß auf die chemische Verschiebung der klassifizierten Position haben: Gegenüber dem Einfluß entfernterer Atome innerhalb des *ipso*-Substituenten haben die Klassen der übrigen Substituenten aufgrund ihrer direkten Verbindung zum aromatischen System des Benzolrings einen größeren Einfluß. Die Tatsache, daß sogar eine Verschlechterung der Klassifikationsleistung verglichen mit der Gesamtstichprobe zu beobachten ist, begründet sich in dem wiederum zu geringen Umfang der hier evaluierten Auswahl.

Initialisierung	Monosubstituierte	kurze Substituenten
empirisch,hart,exakt	ipso: 8,96% ortho: 45,44%	ipso: 30,65% ortho: 94,30%
empirisch,hart,tolerant	ipso: 8,03% ortho: 69,70%	ipso: 30,82% ortho: 96,47%
empirisch,weich,tolerant	ipso: 8,03% ortho: 66,11%	ipso: 30,93% ortho: 97,55%
empirisch,weich,exakt	ipso: 8,93% ortho: 51,13%	ipso: 30,67% ortho: 97,31%
gleichverteilt,weich,exakt	ipso: 94,63% ortho: 52,48%	ipso: 66,59% ortho: 97,43%
gleichverteilt,weich,tolerant	ipso: 93,47% ortho: 66,13%	ipso: 65,39% ortho: 97,78%
gleichverteilt,hart,tolerant	ipso: 93,47% ortho: 69,33%	ipso: 65,63% ortho: 96,95%
gleichverteilt,hart,exakt	ipso: 94,44% ortho: 46,02%	ipso: 66,36% ortho: 95,16%

Tab. 10.3: *Evaluationsergebnisse unter Verwendung einer Auswahl von Datensätzen der Stichprobe*

Zusammenfassend läßt sich sagen, daß das Modell in seiner jetzigen Form den Ansprüchen eines zuverlässigen Musteranalysesystems nicht gerecht wird und daß somit die einzelnen Modellierungsentscheidungen zu überdenken sind. Um so aufschlußreicher wird der im folgenden Abschnitt angestrebte Vergleich mit diesbezüglichen Modellvarianten sein. Festgehalten werden sollte jedoch auch, daß trotz der bisher unbefriedigenden Fehlerraten, die auf einzelne ungünstige Modellierungsentscheidungen zurückgeführt werden, die grundsätzliche Eignung des Ansatzes nicht widerlegt ist. Es wird vielmehr erwartet, daß die Entwicklung fortgeschrittener Modellvarianten zu einer deutlichen Verbesserung führen wird.

10.3 Varianten

Im Zuge der Entwicklung des Modells in Kapitel 6 wurden bereits Ideen angedeutet, wie einzelne Modellierungsentscheidungen anders getroffen werden könnten. Verschiedene derartige Varianten sollen nun evaluiert und einander gegenübergestellt werden, um den Einfluß der jeweils variierten Modellierungsschritte zu untersuchen. Änderungen des Modells sind insbesondere auch im Kontext der größtenteils unbefriedigenden bisherigen Fehlerraten empfehlenswert. Zur besseren Nachvollziehbarkeit ist in Abbildung 10.2 noch einmal das ursprüngliche Modell dargestellt. Die ihrer Bedeutung nach unterschiedenen Arten von Variablen sind dabei farbig hervorgehoben.

10.3.1 Wiedergabe der Summenformelinformation

Ein Gedanke, der bereits bei den Überlegungen zur Modellentwicklung eine Rolle spielte ist die Frage, ob es ausreicht, die Summenformelinformation binär zu repräsentieren, oder ob das Modell von einer Darstellung profitieren würde, die auch die Zahl der Atome jedes betrachteten chemischen Elements berücksichtigt. Eine erste Variante des Modells dient der Untersuchung dieser Frage.

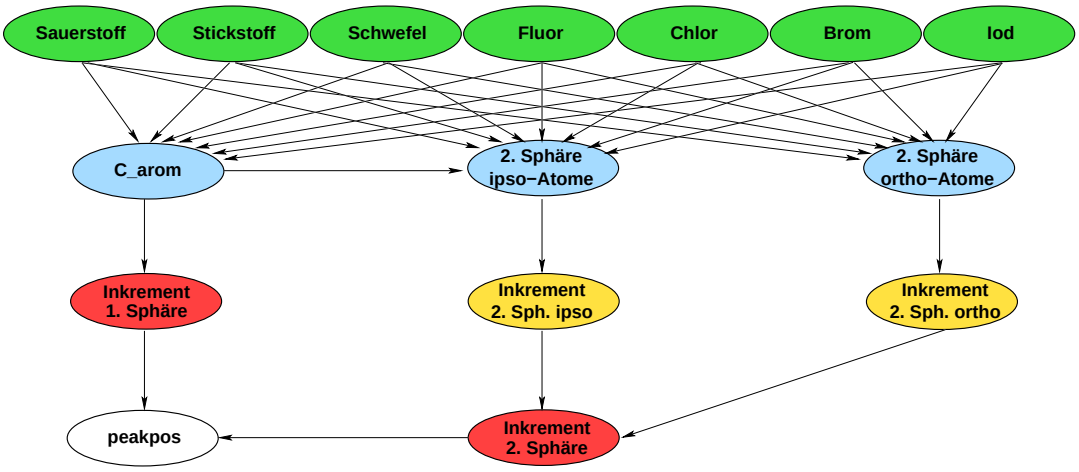


Abb. 10.2: *Modell des Zusammenhangs zwischen Spektrum und Struktur, wie es in Kapitel 6 erarbeitet wurde. Die Summenformelvariablen sind grün, Strukturvariablen blau, Variablen für die Wiedergabe der sphärenweisen Inkremente rot und für die Wiedergabe von Teilinkrementen gelb hervorgehoben.*

Wird jedoch die Zahl der Zustände bei den Summenformelvariablen erhöht, so führt dies zu einer mit der gegebenen Stichprobe nicht mehr handhabbaren Anzahl benötigter Verteilungen für die Strukturvariablen, wie bereits in Abschnitt 10.1 ausgeführt wurde. Eine Änderung in der Sichtweise des Kausalzusammenhangs erlaubt es jedoch, die gewünschte Änderung dennoch einzuführen: Die bisherige Sicht faßte die Möglichkeit des Auftretens einer Gruppe als Konsequenz des Vorhandenseins von Atomen eines bestimmten chemischen Elements auf. Beispielsweise wurde das Auftreten einer OH-Gruppe für möglich angesehen, *weil* Sauerstoff im Molekül anwesend war. Diese Sicht läßt sich jedoch auch gerade umkehren, und zwar dahingehend, daß das Vorkommen bestimmter Gruppen in den durch die Strukturvariablen wiedergegebenen Positionen die Beobachtung des Vorhandenseins von Atomen bestimmter chemischer Elemente nach sich zieht. In diesem Fall wäre dann beispielsweise eine bestimmte Anzahl von Sauerstoffatomen im Molekül beobachtbar, *weil* bestimmte sauerstoffhaltige Gruppen in den modellierten Ringpositionen vorliegen. Die entsprechende Änderung des Modells ist in Abbildung 10.3 exemplarisch für Sauerstoff dargestellt.

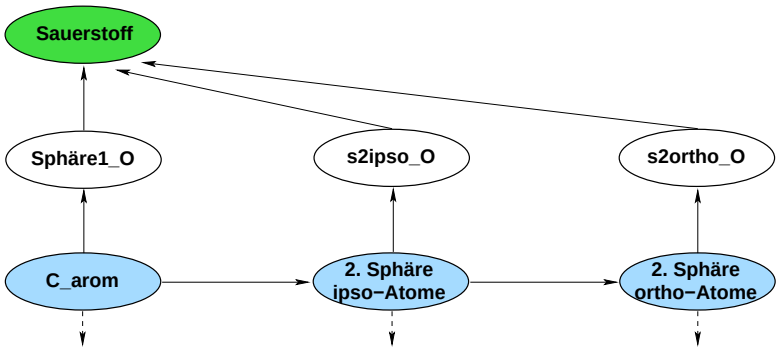


Abb. 10.3: *Änderung des Modells, die die Summenformelinformation als Konsequenz der Molekülstruktur betrachtet. Die zusätzlichen Zwischenvariablen dienen der Vereinfachung.*

Neben dem oben beschriebenen wurde zur Vereinfachung für jedes chemische Element und jede betrachtete Position eine Zwischenvariable, hier `Sphäre1_0`, `s2ipso_0` und `s2ortho_0`, eingefügt. Diese Variablen modellieren als Eigenschaft der betreffenden (Teil-) Sphäre die Anzahl der dort vorhandenen Atome des jeweiligen chemischen Elements. Auf diese Weise wird nur der jeweils relevante Anteil der Information der Strukturvariablen zu den Validierungsvariablen transportiert: Die Vorkommen von Stickstoff, Schwefel usw. spielen keine Rolle für die Zahl der Vorkommen von Sauerstoff im Molekül. Die Ergebnisse der Evaluierung dieses Modells sind nachfolgend Tabelle 10.4 zu entnehmen.

Initialisierung	Gesamtmenge	Monosubstituierte	kurze Substituenten
empirisch,hart,exakt	ipso: 27,48% ortho: 61,00%	ipso: 7,62% ortho: 30,32%	ipso: 28,69% ortho: 62,99%
empirisch,hart,tolerant	ipso: 27,57% ortho: 59,82%	ipso: 8,01% ortho: 30,69%	ipso: 28,82% ortho: 65,15%
empirisch,weich,tolerant	ipso: 27,29% ortho: 60,39%	ipso: 8,01% ortho: 30,69%	ipso: 28,75% ortho: 66,66%
empirisch,weich,exakt	ipso: 27,15% ortho: 61,07%	ipso: 7,80% ortho: 29,95%	ipso: 28,44% ortho: 64,04%
gleichverteilt,weich,exakt	ipso: 58,66% ortho: 82,18%	ipso: 70,16% ortho: 84,38%	ipso: 48,75% ortho: 82,72%
gleichverteilt,weich,tolerant	ipso: 58,77% ortho: 81,62%	ipso: 71,27% ortho: 88,03%	ipso: 47,53% ortho: 84,28%
gleichverteilt,hart,tolerant	ipso: 58,27% ortho: 79,68%	ipso: 71,64% ortho: 85,65%	ipso: 48,75% ortho: 80,61%
gleichverteilt,hart,exakt	ipso: 59,31% ortho: 80,11%	ipso: 71,32% ortho: 78,91%	ipso: 49,12% ortho: 78,59%

Tab. 10.4: *Fehlerrate bei der Evaluierung des abgewandelten Modells mit Umkehr der Kausalrichtung zwischen Struktur- und Summenformelvariablen sowie Berücksichtigung der Zahl der Vorkommen der Atome der betrachteten chemischen Elemente.*

Hier zeigt sich zunächst auf der Gesamtstichprobe bei der *ipso*-Klassifikation für „empirisch“ initialisierte Netze keine große Veränderung; die Fehlerraten liegen nach wie vor im Bereich von 27%. Sowohl für monosubstituierte Verbindungen als auch für Verbindungen mit kurzen Substituenten ist jedoch eine leichte Verbesserung festzustellen. Eine deutliche Verbesserung gibt es bei allen Stichproben für „gleichverteilt“ initialisierte Netze, und zwar jeweils in der Größenordnung von 20% (absolut).

Deutlich niedriger werden die Fehlerraten ebenfalls bei allen drei Untersuchungen beim *ortho*-Klassifikationsschritt, mit Ausnahme der „gleichverteilten“ Initialisierung für monosubstituierte Verbindungen. Hierfür gibt es jedoch eine Erklärung, und zwar die Initialisierung des Übergangs von der Struktur- zur Summenformelinformation: Da diese wissensbasiert, ohne Bezug zur untersuchten Stichprobe, geschieht, berücksichtigt sie nicht, daß es sich um monosubstituierte Verbindungen handelt. Daher kann der Klassifikator in seinen Folgerungen nicht berücksichtigen, daß nur ein einziger Substituent entweder in der *ipso*- oder in einer der *ortho*-Positionen vorhanden sein darf. Es wäre zwar möglich, die Initialisierung entsprechend zu ändern, doch wäre dies nicht von Vorteil, da das Modell auf diese Weise stärker spezialisiert würde als ursprünglich beabsichtigt. Statt dessen wird eine andere Variante der Summenformelmodellierung angestrebt. Zuvor soll jedoch das Resumee der gegenwärtigen Evaluierung abgeschlossen werden.

Läßt man die gleichverteilten Initialisierungsvarianten im Zusammenhang mit monosubstituierten Verbindungen außen vor, so ist zu bemerken, daß die Zuverlässigkeit des Klassifikationsresultats weniger stark davon abhängig ist, ob die „gleichverteilte“ oder die „empirische“ Initialisierungsvariante gewählt wurde. Dies bedeutet, daß ein Teil der zuvor in statistischen Zusammenhängen „versteckten“ Faktoren, die sich nur in der „empirischen“ Initialisierungsvariante widerspiegeln, nun explizit geworden ist. Gleichwohl gibt es, wie die zwar geringeren, aber immer noch sehr deutlichen Unterschiede zeigen, noch immer derartige Faktoren, die es in zukünftigen Weiterentwicklungen des Modells zu erfassen gilt.

Weiterhin fällt insbesondere bei empirischer Initialisierung für die kleineren Teilmengen der Verbindungen mit kurzen Substituenten bzw. der monosubstituierten Verbindungen ins Gewicht, daß durch die Umkehr der Kausalrichtung und Einbringung der Zwischenvariablen erheblich weniger Verteilungen zu schätzen sind: Es stehen nur etwa 480 Beispiele aus monosubstituierten Verbindungen bzw. etwa 750 Beispiele aus Verbindungen mit kurzen Substituenten zur Verfügung, nun sind jedoch nur noch 24 Verteilungen für jede der Summenformel-Variablen statt zuvor mindestens 128 für die Strukturvariablen zu schätzen. In jedem Fall bleibt festzuhalten, daß die Anzahl der Atome pro Element innerhalb der Summenformelinformation eine wichtige Rolle spielt und daher auch in die Modellierung einzubringen ist.

Insbesondere vor dem Hintergrund der oben gegebenen Erklärung für die schlechten Fehlerraten im *ortho*-Klassifikationsschritt für monosubstituierte Verbindungen sind jedoch weitere Überlegungen hinsichtlich einer alternativen Realisierung angebracht. Einen guten Ansatzpunkt liefert dabei die Tatsache, daß hinsichtlich des Zusammenhangs zwischen Summenformel und Strukturformel beide beschriebenen Sichtweisen des Kausalzusammenhangs sinnvoll sind: Dies bedeutet, daß es gültige und ungültige Konfigurationen gibt, das heißt solche, bei denen Summenformel- und Strukturinformation übereinstimmen und solche, bei denen dies nicht der Fall ist. Es handelt sich also um eine ungerichtete Abhängigkeit, wie bereits in Abschnitt 6.2.1 in Erwägung gezogen. Wenngleich es in manchen Fällen ausreichend oder sogar vorteilhaft sein kann, sich für eine der möglichen Sichtweisen zu entscheiden, liegt hier jedoch die Vermutung nahe, daß die ungerichtete Abhängigkeit auch als solche repräsentiert werden sollte.

Wie in Abschnitt 4.3.2 beschrieben werden ungerichtete Abhängigkeiten in die gerichteten Kausalzusammenhänge eines Bayes-Netzes eingebracht, indem eine Zwischenvariable als gemeinsames Kind der betreffenden Variablen (hier also der Summenformel- und der Strukturvariablen) eingeführt wird. Diese Zwischenvariable wird im folgenden als Validierungsvariable bezeichnet und erhält die Zustände `gültig` und `ungültig`. Abbildung 10.4 zeigt diese Änderung des Modells exemplarisch für Sauerstoff.

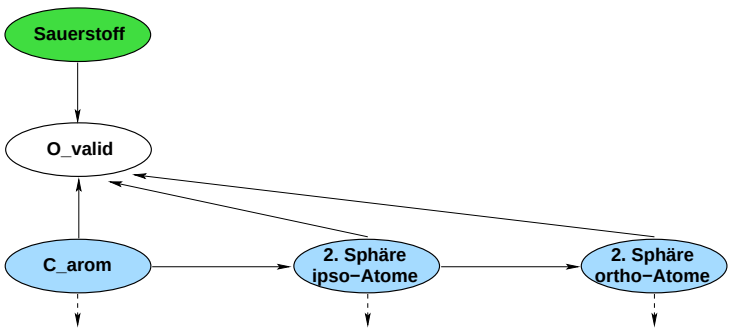


Abb. 10.4: *Modellierung der ungerichteten Abhängigkeit zwischen Struktur- und Summenformelinformation am Beispiel von Sauerstoff. Struktur- und Summenformelvariablen (hier* `Sauerstoff`*) erhalten eine Validierungsvariable (hier* `O_valid`*) als gemeinsames Kind.*

Die Wahrscheinlichkeit P(O_valid=ungültig) beträgt 1 für alle ungültigen Kombinationen der Elternvariablen, also diejenigen Kombinationen, die für mehr Vorkommen von Sauerstoff in der Struktur stehen, als die Summenformel zuläßt. Umgekehrt kann jedoch die Summenformel mehr Sauerstoffatome enthalten als in den Strukturvariablen aufgeführt, da nicht die gesamte Molekülstruktur explizit repräsentiert ist, sondern nur die unmittelbare Umgebung von zwei Sphären. Die Relation, die die Abhängigkeit der Elternvariablen beschreibt (vgl. Gleichung (4.15), S. 67) ist also eine Größergleich-Relation.

Bei der Anfrage an das Bayes-Netz wird für jedes betrachtete chemische Element gültig als Zustand der Validierungsvariablen als zusätzliche Evidenz angegeben, um den gewünschten Effekt zu erzielen. Tabelle 10.5 zeigt die Ergebnisse der Evaluation dieser Modellvariante.

Initialisierung	Gesamtmenge	Monosubstituierte	kurze Substituenten
empirisch,hart,exakt	ipso: 26,81% ortho: 59,76%	ipso: 7,25% ortho: 33,52%	ipso: 27,86% ortho: 63,76%
empirisch,hart,tolerant	ipso: 27,40% ortho: 59,25%	ipso: 7,82% ortho: 33,33%	ipso: 27,98% ortho: 66,99%
empirisch,weich,tolerant	ipso: 27,17% ortho: 59,13%	ipso: 7,64% ortho: 33,33%	ipso: 27,88% ortho: 67,76%
empirisch,weich,exakt	ipso: 26,72% ortho: 59,08%	ipso: **3,36%** ortho: 33,33%	ipso: 29,20% ortho: 65,07%
gleichverteilt,weich,exakt	ipso: 41,90% ortho: 82,23%	ipso: 39,05% ortho: 42,73%	ipso: 46,50% ortho: 87,18%
gleichverteilt,weich,tolerant	ipso: 41,38% ortho: 82,63%	ipso: 47,66% ortho: 56,57%	ipso: 40,67% ortho: 88,17%
gleichverteilt,hart,tolerant	ipso: 40,94% ortho: 79,76%	ipso: 47,85% ortho: 60,56%	ipso: 41,76% ortho: 85,65%
gleichverteilt,hart,exakt	ipso: 42,36% ortho: 78,85%	ipso: 42,52% ortho: 37,08%	ipso: 46,36% ortho: 81,29%

Tab. 10.5: *Fehlerraten bei der Evaluation des abgewandelten Modells mit einer ungerichteten Abhängigkeit zwischen Summenformel- und Strukturinformation.*

Am auffälligsten ist bei dieser Evaluierung sicherlich die Verbesserung im *ipso*-Klassifikationsschritt bei „gleichverteilter“ Initialisierung: Sowohl bezogen auf die Gesamtstichprobe wie auch auf die Auswahl monosubstituierter Verbindungen wird die Fehlerrate verglichen mit dem Originalmodell in etwa halbiert, und auch im Vergleich zur vorherigen Modellvariante, welche zwar die Multiplizitäten der Heteroatome, nicht aber die ungerichtete Abhängigkeit zwischen Struktur- und Summenformelinformation berücksichtigte (Tabelle 10.4), ist eine erhebliche Verbesserung zu beobachten. Lediglich für Verbindungen mit kurzen Substituenten ist der Zugewinn verglichen mit der vorherigen Variante mit etwa 2–7% (absolut) nicht in jedem Fall so deutlich, bezogen auf das Originalmodell sinkt die Fehlerrate jedoch auf etwa zwei Drittel des ursprünglichen Wertes, und die Verbesserung ist insgesamt somit ebenfalls groß.

Gleichzeitig ist bei „empirischer“ Initialisierung kaum eine Verbesserung festzustellen. Daraus kann man folgern, daß es nach wie vor Faktoren gibt, die nicht explizit durch das Modell erfaßt werden, sondern sich nur implizit durch statistische Eigenschaften bei der „empirischen“ Initialisierung widerspiegeln. Ein Fall ist jedoch besonders hervorzuheben, und zwar die „empirisch-weich-exakte“ Initialisierung bei der Untersuchung monosubstitu-

ierter Verbindungen: Hier beträgt die Fehlerrate nur 3,36%. Dabei spielt natürlich zum ersten die Übereinstimmung der Modellierungsgenauigkeit eine Rolle, die wie bereits erwähnt im Falle monosubstituierter Verbindungen am treffendsten ist (vgl. S. 149), und bei der außerdem offensichtlich nicht nur die Multiplizitätsinformation aus der Summenformel, sondern auch der ungerichtete Zusammenhang zur Strukturinformation relevant ist.

Im Kontext der übrigen „empirischen“ Initialisierungsvarianten zeigt sich darüber hinaus, daß eine zu starke Aufweichung der Kausalzusammenhänge im Bereich der Inkremente („tolerante“ Initialisierung) ebenso zu einer Verschlechterung der Klassifikationsleistung führt wie ein zu striktes Festhalten an „harten Lehrbuchwerten“. Hinsichtlich der „toleranten“ weichen Summenbildung gilt dies ebenso bei den „gleichverteilt“ initialisierten Netzen. Dort ist jedoch die im Bereich des Struktur-Inkrement-Überganges „hart“ initialisierte Variante der „weichen“ überlegen. Ist empirisches Wissen („Erfahrung“) vorhanden, so bringt die „weiche“ Initialisierung also Vorteile, während ohne dasselbe ein Festhalten an den exakten Inkrementwerten der bessere Weg ist.

Im *ortho*-Klassifikationsschritt ist dagegen bei „gleichverteilter“ Initialisierung die „harte“, „exakte“ Variante die mit der niedrigsten Fehlerrate für monosubstituierte Verbindungen. Sie erreicht sogar annähernd dieselbe Größenordnung wie bei empirischer Initialisierung, während insbesondere die „toleranten“ Varianten eine um 20% (absolut) oder mehr höhere Fehlerrate aufweisen. Eine allzu starke Aufweichung der modellierten Kausalzusammenhänge wirkt sich also auch hier negativ aus.

Wie im Originalmodell und in der Modellvariante ohne ungerichtete Abhängigkeit zwischen Summenformel- und Strukturinformation ist auch bei der neuen Modellvariante die Leistung auf der Teilmenge der monosubstituierten Verbindungen am besten. Insgesamt bleibt sie jedoch im gleichen Bereich wie bei der vorhergehenden Variante, die die Multiplizitäten, aber nicht die ungerichtete Abhängigkeit zur Molekülstruktur bei der Summenformelinformation beachtet. Durch die Wiedergabe der ungerichteten Abhängigkeit gibt es ausschließlich im Fall der monosubstituierten Verbindungen bei gleichverteilter Initialisierung eine Verbesserung. Hier hatte die vorherige Variante initialisierungsbedingt zu einer merklichen Verschlechterung geführt.

Bevor sich der folgende Abschnitt weiteren Variationsmöglichkeiten der Modellierung widmet, sollen nun die bis hierher gewonnenen Erkenntnisse zusammengefaßt werden:

- Die Leistung des *ipso*-Klassifikationsschrittes wird durch eine „empirische“ Initialisierung extrem positiv beeinflußt. Der starke Einfluß statistischer Faktoren spricht für das Wirken nicht vom Modell erfaßter Gesetzmäßigkeiten. Dies bedeutet auch, daß bei der Beurteilung von Änderungen des Modells den „gleichverteilten“ Initialisierungsvarianten ein größeres Gewicht zukommt.

- Durch die Hinzunahme der Multiplizitäten aus der Summenformelinformation wird eine merkliche Verbesserung erreicht.

- Wird der Zusammenhang zwischen Summenformel und Molekülstruktur in der Sichtweise als ungerichtete Abhängigkeit repräsentiert, ist eine weitere Verbesserung zu erzielen.

- Ein Modell mit „empirisch-weich-exakter“ Initialisierung, welches die ungerichtete Abhängigkeit zur Molekülstruktur und die Multiplizitäten aus der Summenformelinformation berücksichtigt, erzielte mit einer Fehlerrate von 3,36% im *ipso*- und 33,33% im *ortho*-Klassifikationsschritt die bislang besten Resultate.

10.3.2 Zusammenwirken der Inkremente und Teilinkremente

Ein weiterer Punkt von Interesse ist das Zusammenwirken der einzelnen Einflüsse, die den Atomen in den betrachteten Positionen zugeordnet sind, da die Elektroneneinflüsse innerhalb des Moleküls nicht lokal betrachtet werden können. Die ursprüngliche Modellierung ordnet der *s1ipso*-, *s2ipso*- und *s2ortho*-Position jeweils einen Inkrementwert zu, der in Abhängigkeit von den Atomen in der betreffenden Position zur chemischen Verschiebung des Fokusatoms beiträgt. Wie in Abschnitt 6.1.3 erwähnt, steht die im resultierenden Modell verwendete Wahrscheinlichkeit $P(\texttt{Sphäre2}|\texttt{i2ipso},\texttt{i2ortho})$ in einer ebensolchen Beziehung zu den (im Modell nicht benötigten) Wahrscheinlichkeiten $P(\texttt{Sphäre2}|\texttt{i2ipso})$ und $P(\texttt{Sphäre2}|\texttt{i2ortho})$ wie Kreuzterme zu den Einzelinkrementen im Inkrementansatz der Spektrenvorhersage (vgl. Abschnitt 3.1.2), von dem das Vorgehen inspiriert wurde.

Bei der Zusammenführung der in `i2ipso` und `i2ortho` repräsentierten Teilinkremente zum Inkrement der zweiten Sphäre findet zur Zeit nur eine „weiche Summenbildung“ statt, die entweder nur die Rundung auf volle ppm berücksichtigt („exakte“ Initialisierung) oder gewissermaßen Additionsfehler innerhalb eines gewissen Toleranzintervalls akzeptiert („tolerante“ Initialisierung). Dem Gedanken der Verwendung von Kreuztermen wird dies nur bedingt gerecht, da im Prinzip die Annahme getroffen wird, daß beim Zusammenwirken der Einflüsse die Additivität eine dominante Rolle spielt und nur wenig von der exakten Summenbildung abgewichen werden muß.

Insbesondere ist eine differenzierte Behandlung einzelner Substituentenkombinationen, wie sie durch individuelle Kreuzterme realisiert würde, nur eingeschränkt vorhanden. Dieser Aspekt fließt statt dessen indirekt in die Bestimmung der *s2ortho*-Inkremente ein, die wie in Abschnitt 8.3.1 beschrieben empirisch bestimmt werden, wobei die Inkremente der übrigen Positionen bekannt sein müssen. Über diese bekannten Inkremente wird jedoch bereits eine Abhängigkeit von den entsprechenden Atomen hergestellt. So wird streng genommen in erster Linie das *s2ortho*-Inkrement abhängig von ihnen bestimmt, anstatt beim Zusammenführen der Teilinkremente explizit das Zusammenwirken unterschiedlicher Einflüsse zu berücksichtigen. In beiden Fällen resultiert das Vorgehen zwar in mehreren unterschiedlichen Inkrementswerten, die für ein und dieselbe Belegung der *s2ortho*-Position in Frage kommen, jedoch liegt der Unterschied darin, daß bei der gegenwärtigen Realisierung keine Zuordnung der unterschiedlichen Beträge zu bestimmten Substitutionsmustern erreicht werden kann. Vor diesem Hintergrund bietet sich eine alternative Gestaltung des Modells im Bereich der Struktur-Inkrement-Zusammenhänge an.

Eine erste Möglichkeit besteht dabei im Verzicht auf die Variablen der Teilinkremente: Es wird nur noch seitens der Struktur zwischen *ipso*- und *ortho*-Anteil unterschieden, die zugehörigen Teilinkremente entfallen jedoch, und das resultierende Gesamtinkrement der zweiten Sphäre wird direkt in Abhängigkeit von den Strukturvariablen der zweiten Sphäre empirisch ermittelt. Die sich so ergebende bedingte Wahrscheinlichkeit $P(\texttt{Sphäre2}|\texttt{s2ipso},\texttt{s2ortho})$ erfordert 504 Wahrscheinlichkeitsverteilungen. Bei dem gegebenen Größe der Trainingsmenge für die Gesamtstichprobe von knapp 4900 Einzelpeaks scheint dieser Ansatz zumindest vertretbar. Auch für die monosubstituierten Verbindungen kommt er in Frage, da hier aufgrund der Tatsache, daß nur genau ein Substituent vorhanden ist, viele Zustandskombinationen von `s2ipso` und `s2ortho` von vornherein ausgeschlossen sind, was die Datenlage für die übrigen Kombinationen verbessert. Die Verbindungen mit kurzen Substituenten, bei denen den erwähnten 504 Verteilungen nur etwa 750 Trainingsbeispiele gegenüberstehen, werden dagegen nur unter Vorbehalt untersucht.

Initialisierung	Gesamtmenge	Monosubstituierte	kurze Substituenten
empirisch,hart,exakt	ipso: 27,17% ortho: 60,39%	ipso: 4,17% ortho: 30,12%	ipso: 29,43% ortho: 68,06%
empirisch,hart,tolerant	ipso: 27,13% ortho: 60,46%	ipso: 4,54% ortho: 29,91%	ipso: 27,84% ortho: 66,68%
empirisch,weich,tolerant	ipso: 27,22% ortho: 60,46%	ipso: 4,54% ortho: 30,46%	ipso: 28,21% ortho: 66,67%
empirisch,weich,exakt	ipso: 27,26% ortho: 60,31%	ipso: 4,17% ortho: 30,07%	ipso: 29,79% ortho: 67,57%
gleichverteilt,weich,exakt	ipso: 42,73% ortho: 69,23%	ipso: 39,93% ortho: 60,19%	ipso: 37,68% ortho: 70,38%
gleichverteilt,weich,tolerant	ipso: 43,80% ortho: 70,56%	ipso: 48,29% ortho: 64,35%	ipso: 37,30% ortho: 69,25%
gleichverteilt,hart,tolerant	ipso: 43,78% ortho: 70,12%	ipso: 44,77% ortho: 63,06%	ipso: 36,34% ortho: 68,75%
gleichverteilt,hart,exakt	ipso: 41,38% ortho: 68,68%	ipso: 39,93% ortho: 51,32%	ipso: 36,82% ortho: 70,13%

Tab. 10.6: *Fehlerrate bei der Evaluation der Modellvariante ohne die einzelnen Teilinkremente der zweiten Sphäre sowie mit Umkehr der Kausalrichtung bei der Summenformelmodellierung.*

Initialisierung	Gesamtmenge	Monosubstituierte	kurze Substituenten
empirisch,hart,exakt	ipso: 26,50% ortho: 57,31%	ipso: 3,77% ortho: **22,08%**	ipso: 27,63% ortho: 67,16%
empirisch,hart,tolerant	ipso: 26,46% ortho: 57,81%	ipso: 3,96% ortho: 24,81%	ipso: 27,39% ortho: 67,76%
empirisch,weich,tolerant	ipso: 26,10% ortho: 57,86%	ipso: 3,96% ortho: 24,81%	ipso: 27,51% ortho: 67,63%
empirisch,weich,exakt	ipso: 26,63% ortho: 57,33%	ipso: 3,77% ortho: 22,64%	ipso: 27,75% ortho: 67,16%
gleichverteilt,weich,exakt	ipso: 30,73% ortho: 69,66%	ipso: 41,39% ortho: 57,92%	ipso: 31,10% ortho: 67,14%
gleichverteilt,weich,tolerant	ipso: 28,66% ortho: 67,40%	ipso: 46,55% ortho: 62,45%	ipso: 29,92% ortho: 70,49%
gleichverteilt,hart,tolerant	ipso: 28,63% ortho: 67,03%	ipso: 41,92% ortho: 61,16%	ipso: 29,68% ortho: 70,49%
gleichverteilt,hart,exakt	ipso: 29,34% ortho: 69,23%	ipso: 41,02% ortho: 49,05%	ipso: 31,10% ortho: 67,13%

Tab. 10.7: *Fehlerrate bei der Evaluation der Modellvariante ohne die einzelnen Teilinkremente der zweiten Sphäre sowie mit Wiedergabe der ungerichteten Abhängigkeit zwischen Summenformel- und Strukturinformation*

Die neue Modellvariation, auf die Teilinkremente der zweiten Sphäre zu verzichten, wird in beide im vorigen Abschnitt entwickelten Modelle, die die Multiplizitäten der einzelnen Atomspezies aus der Summenformelinformation mit berücksichtigen, integriert. Obwohl prinzipiell festgestellt wurde, daß die Wiedergabe des Zusammenhangs zwischen Molekülstruktur und Summenformel als ungerichtete Abhängigkeit vorteilhaft ist, wird auch die Variante ohne diese Modellierungsentscheidung evaluiert, um einen Einblick in die Relation beider Aspekte, Summenformelrepräsentation und Aufschlüsselung des Inkrements der zweiten Sphäre, zu erhalten. Tabelle 10.6 zeigt die Ergebnisse der Evaluierung der Variante mit der Abbildung 10.3 entsprechenden Summenformelmodellierung, Tabelle 10.7 zeigt die Ergebnisse mit der Summenformelmodellierung gemäß Abbildung 10.4 unter Berücksichtigung der ungerichteten Abhängigkeit zwischen Struktur und Summenformel.

Wie bereits bei den vorigen Untersuchungen ist auch hier in beiden Fällen im *ipso*-Klassifikationsschritt bei „empirischer“ Initialisierung kaum eine Veränderung festzustellen; ähnliches gilt für die Auswahl der Verbindungen mit kurzen Substituenten. Für monosubstituierte Verbindungen werden Ergebnisse erzielt, die an das bislang beste Resultat der „empirisch-weich-exakten“ Initialisierung der Modellvariante mit Repräsentation des Summenformel-Struktur-Zusammenhangs als ungerichteter Abhängigkeit heranreichen. Zusätzlich wurde eine größere Robustheit gegen Aufweichungen der strikten Zusammenhänge durch eine „weiche“ Initialisierung oder „tolerante“ weiche Summenbildung erreicht. Dies ist positiv zu bewerten, da diese Aufweichung dazu dient, das Wirken nicht explizit modellierter Einflüsse auszugleichen und so auch bei einer weniger guten Übereinstimmung zwischen Modell und Stichprobe, als dies bei monosubstituierten Verbindungen der Fall ist, die richtigen Schlüsse ziehen zu können.

Ebenfalls positiv zu bewerten ist die Leistung bei Untersuchung der Gesamtstichprobe und „gleichverteilter“ Initialisierung: Die Kombination aus ungerichteter Abhängigkeit im Bereich der Summenformelrepräsentation und Entfernen der Variablen für die Teilinkremente der zweiten Sphäre führt hier zu einer kaum mehr höheren Fehlerrate als bei empirischer Initialisierung. Dies ist ein deutliches Signal dafür, daß das Zusammenwirken der einzelnen Einflüsse bei der Modellierung zusätzlicher Aufmerksamkeit bedarf.

Hinsichtlich des *ortho*-Klassifikationsschrittes lassen sich die meisten der obigen Gedanken übertragen. Sowohl für die Gesamtstichprobe als auch für die Verbindungen mit kurzen Substituenten bleiben die Werte bei „empirischer“ Initialisierung annähernd gleich. Bei „gleichverteilter“ Initialisierung ist dagegen eine weitere Verbesserung zu beobachten, die jeweils in der Größenordnung von 10% (absolut) oder mehr liegt. Interessant ist jedoch, daß im *ortho*-Klassifikationsschritt teilweise auch bei den monosubstituierten Verbindungen Verbesserungen zu beobachten sind, und zwar bei „empirischer“ Initialisierung und unter Einbeziehung der ungerichteten Abhängigkeit sogar um bis zu einem Drittel, von um 33% auf 22-25%. Im besonderen wurde das bislang beste Klassifikationsergebnis für den *ortho*-Schritt bei der Untersuchung der monosubstituierten Verbindungen mit dem zuletzt entwickelten Modell und bei „empirisch-hart-exakter“ Initialisierung erreicht. Es scheinen also in um so stärkerem Maße solche Faktoren an Gewicht zu gewinnen, die sich in erster Linie über statistische Zusammenhänge ausdrücken, die sich in der „empirischen“ Initialisierung widerspiegeln, und die bislang nicht explizit im Modell erfaßt werden.

Bei „gleichverteilter“ Initialisierung liegen die erreichten Werte unabhängig von der Summenformelrepräsentation innerhalb jeder der drei betrachteten Stichproben jeweils in etwa in ein und demselben Bereich, woraus man folgern kann, daß für die *ortho*-Klassifikation eine Präzisierung des Zusammenwirkens der unterschiedlichen Einflüsse bedeutsamer ist als die Art der Summenformelrepräsentation, so lange jedenfalls die Information der Atomanzahlen wiedergegeben wird.

Verfolgt man jedoch die Überlegungen weiter, die zum Verzicht auf die Aufschlüsselung des Inkrements der zweiten Sphäre in einen *ipso*- und einen *ortho*-Anteil führten, so läßt sich dieselbe Argumentation auch für die Zusammenführung der Inkremente der ersten und zweiten Sphäre anwenden: Das Zusammenwirken der Einflüsse findet implizit bereits in der empirischen Bestimmung des Inkrements der zweiten Sphäre Eingang, und nicht, wie ursprünglich angestrebt, während der Zusammenführung der beiden Inkremente in Abhängigkeit vom Substitutionsmuster. Ein analoges Vorgehen wie im Fall der in `i2ipso` und `i2ortho` repräsentierten Teilinkremente würde nun jedoch zum Entfernen der Inkrementvariablen `Sphäre1` und `Sphäre2` führen, so daß die Peakposition ohne erkennbare Systematik direkt von den Strukturmerkmalen abhinge. Statt der beabsichtigten differenzierte Modellierung bliebe damit nur noch ein statistischer Zusammenhang zwischen Spektrum und Struktur übrig.

Wenngleich die Tatsache, daß im Rahmen der Initialisierungsvarianten statistische Einflüsse in Gestalt der „empirischen“ Initialisierung einen positiven Einfluß auf die Erkennungsleistung haben, vordergründig für eine solche Entscheidung spricht, so liefe sie zugleich jedoch dem gewählten Grundansatz einer expliziten Wiedergabe von Fachwissen entgegen. Daher soll an dieser Stelle eine Möglichkeit gesucht werden, dem explizit repräsentierten Wissen mehr Gewicht zu verleihen, und zwar so, daß im Idealfall das Zusammenwirken der einzelnen Beiträge in Abhängigkeit vom Substitutionsmuster gezielt moduliert wird. Hierbei spielen zwei Aspekte eine Rolle:

1. Die Beiträge der einzelnen Strukturbestandteile zur Lage des untersuchten Peaks sollen explizit vorhanden bleiben und quasi die Argumentationsgrundlage bilden, eine beobachtete chemische Verschiebung auf eine bestimmte strukturelle Umgebung zurückzuführen.

2. Das Zusammenwirken der einzelnen Inkremente und Teilinkremente soll vom Prinzip her für jede Kombination von Strukturmerkmalen individuell gestaltet werden können.

Diese beiden Ideen werden nun jeweils einzeln in das Modell integriert ist. Als erstes wird dem systematischen Zusammenhang zwischen einem Strukturmerkmal und seinem Einfluß auf die chemische Verschiebung des fokussierten Benzolringatoms dadurch Rechnung getragen, auf den der erste Punkt Bezug nimmt. Dies geschieht, indem keines der zugehörigen Inkremente mehr empirisch bestimmt wird. In analoger Weise wie bereits in Abschnitt 8.3.1 beschrieben wird dabei nun auch für die auf die *s2ortho*-Atome zurückzuführenden Teilinkremente der zweiten Sphäre verfahren: Die in [Ewi79] veröffentlichten Auswertungen von Spektren monosubstituierter Benzolderivate enthalten Angaben über den Einfluß der jeweiligen Substituenten auf die *ipso*-, aber auch auf die *ortho*-Position. Nun wird zunächst eine Unabhängigkeit aller an der beobachteten Verschiebung eines Kohlenstoffatoms beteiligten Anteile angenommen. Dadurch können die Teilinkremente der beiden *s2ortho*-Atome als rein additiv sowie als invariant gegeben unterschiedliche Substitutionsmuster angenommen werden. Die gegebenen Einflüsse eines einzelnen Substituenten auf die *ortho*-Position können dann verwendet werden, um durch paarweise Summation den *ortho*-Anteil der zweiten Sphäre zu berechnen. Dasselbe Vorgehen, das bereits zur Feststellung des Wertebereichs von `i2ortho` diente, wird nun also auch zur Bestimmung der bedingten Wahrscheinlichkeit $P(\texttt{i2ortho}|\texttt{s2ortho})$ verwendet.

Um nun das in Wahrheit nicht streng additive Zusammenwirken und somit die wechselseitige Abhängigkeit der Einflüsse im Modell explizit zu machen, wird eine neue Variable `Einflüsse` eingeführt. Sie soll die Diskrepanz zwischen der bis hierher sehr theoretischen Überlegung (summiert in einer ebenfalls neu eingeführten Variable `Erwartung`)

und der tatsächlich beobachteten chemischen Verschiebung ausgleichen. Ihr Wertebereich wird anhand der Stichprobe bestimmt, und auch die zugehörige Wahrscheinlichkeit $P(\texttt{Einflüsse}|\texttt{C\_arom},\texttt{s2ortho})$ wird empirisch ermittelt. Die zweite neu benötigte Wahrscheinlichkeit $P(\texttt{peakpos}|\texttt{Einflüsse},\texttt{Erwartung})$ ergibt sich durch Summenbildung. Abbildung 10.5 verdeutlicht die am Modell vorgenommenen Änderungen.

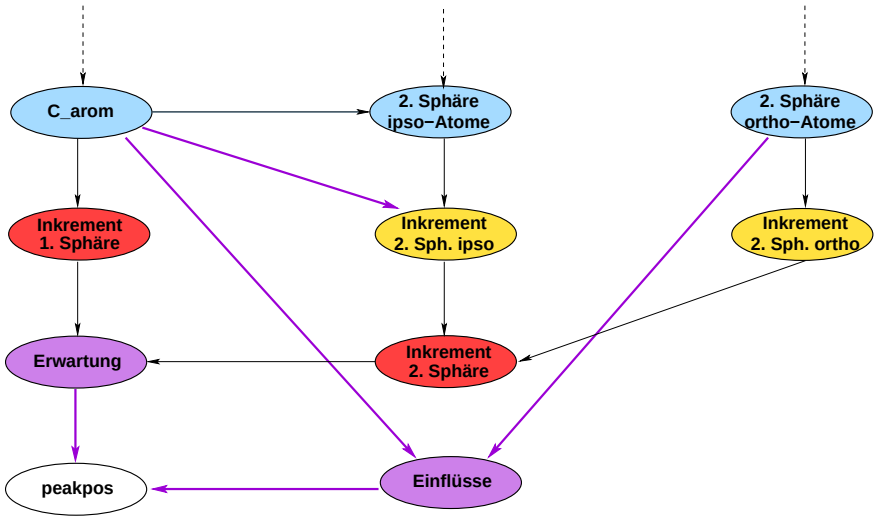


Abb. 10.5: *Änderung des Modells mit dem Ziel einer besseren Wiedergabe des Zusammenspiels der mit den unterschiedlichen Positionen assoziierten Einflüsse. Die dazu neu eingeführten Variablen und Änderungen der Kausalstruktur sind lila hervorgehoben.*

Die neue Variable `Einflüsse` gibt also die notwendigen Korrekturen hinsichtlich der Lage des Peaks abhängig vom vorliegenden Substitutionsmuster wieder. Sie ist dabei aber nicht abhängig von der *s2ipso*-Position, sondern deren Zusammenwirken mit den übrigen Einflüssen wird auf andere Weise, wie im folgenden beschrieben, in das Modell eingebracht. Ein derartiges Vorgehen reduziert die Zahl der benötigten Verteilungen für die Variable `Einflüsse` ganz erheblich: Derzeit werden $14 \cdot 26 = 364$ Verteilungen benötigt, anderenfalls wären es $364 \cdot 24 = 8736$, was den Umfang des mit der gegebenen Stichprobe handhabbaren bei weitem übersteigen würde.

Die Abhängigkeit des effektiven `i2ipso`-Einflusses von den umgebenden Atomen wird im Modell nun dadurch explizit gemacht, daß `i2ipso` eine zusätzliche Abhängigkeit von `C_arom` erhält. Dies läßt sich folgendermaßen motivieren: Da die möglichen Belegungen der *s2ipso*-Position über deren Valenz von der in `C_arom` repräsentierten *s1ipso*-Position abhängen, ist es nicht möglich, beides unabhängig von einander zu betrachten. Es kann nicht, in Analogie zur Definition der Standardgruppen in Abschnitt 6.1.3 (S. 90), eine Ausprägung der ersten Sphäre herangezogen werden, die eine beliebige Belegung der *s2ipso*-Position erlaubt und deren zugehöriges Inkrement als Null angenommen werden könnte. Daher ist es nur konsequent, die Abhängigkeit zwischen erster und zweiter Sphäre der *ipso*-Position in der beschriebenen Weise auch bis auf die Ebene der Inkremente weiterzuführen.

Vor diesem Hintergrund mag es irritierend erscheinen, daß dennoch *ipso*- und *ortho*-Anteil der zweiten Sphäre in `Sphäre2` zu einem Gesamtinkrement zusammengeführt werden, obwohl die wechselweise Einflußnahme der Atome in den unterschiedlichen Positionen auf einander nun anderweitig erfaßt wird. In der Tat hat die Variable `Sphäre2` nun nicht mehr ihren ursprünglichen Sinn, jedoch erfüllt sie beim *Divorcing* von `Sphäre1` auf der einen und `i2ipso` und `i2ortho` auf der anderen Seite nach wie vor ihren Zweck (vgl. Abschnitt 6.2.2).

Bei der gegenwärtigen Fokussierung von Inkrementen und Teilinkrementen ist jedoch eine andere Anpassung sinnvoll, und zwar eine Überarbeitung der von `i2ipso` und `i2ortho` jeweils abgedeckten Intervalle (vgl. Abschnitt 6.2.2). In beiden Fällen werden die repräsentierten Bereiche erheblich erweitert, um auch in den Randbereichen eine nicht weniger genaue Wiedergabe des jeweiligen Beitrages zu erlauben. Die Evaluierungsergebnisse des bis hierher entwickelten Modells sind in Tabelle 10.8 dargestellt.

Es wurde dabei darauf verzichtet, eine Evaluierung auf der recht knapp bemessenen Teilmenge der Verbindungen mit kurzen Substituenten vorzunehmen, da die bedingte Wahrscheinlichkeit $P(\texttt{Einflüsse}|\texttt{C\_arom},\texttt{s2ortho})$ sehr umfangreich ist. Für die hier benötigten 364 Verteilungen stehen nur etwa 750 Beispiele zur Verfügung. Zwar ist das Verhältnis theoretisch für die monosubstituierten Verbindungen noch ungünstiger, hier gibt es jedoch viele Zustandskombinationen der Variablen `C_arom` und `s2ortho`, die nicht auftreten können, so daß die übrigen Verteilungen besser geschätzt werden können.

Initialisierung	Gesamtmenge	Monosubstituierte
empirisch,hart,exakt	ipso: 37,76% ortho: 67,53%	ipso: 20,30% ortho: 38,15%
empirisch,hart,tolerant	ipso: 46,80% ortho: 67,40%	ipso: 19,12% ortho: 37,78%
empirisch,weich,tolerant	ipso: 46,78% ortho: 67,65%	ipso: 19,12% ortho: 37,78%
empirisch,weich,exakt	ipso: 37,33% ortho: 67,11%	ipso: 20,30% ortho: 37,78%
gleichverteilt,weich,exakt	ipso: 37,44% ortho: 67,11%	ipso: 20,30% ortho: 37,78%
gleichverteilt,weich,tolerant	ipso: 46,74% ortho: 67,65%	ipso: 19,12% ortho: 37,78%
gleichverteilt,hart,tolerant	ipso: 46,74% ortho: 67,40%	ipso: 19,12% ortho: 37,78%
gleichverteilt,hart,exakt	ipso: 37,88% ortho: 67,57%	ipso: 20,30% ortho: 38,15%

Tab. 10.8: *Evaluation der Variante des Modells, die durch Trennung wissensbasierter und empirischer Anteile in den Inkrementen eine bessere Wiedergabe des Zusammenspiels der einzelnen Einflüsse anstrebt.*

Auch hier ist, wie bei der vorherigen Modellvariante, zu bemerken, daß die Fehlerraten unabhängig von einer „empirischen“ oder „gleichverteilten“ Initialisierung sind. Dies gilt nun auch für die monosubstituierten Verbindungen, was zuvor nicht der Fall war, und zwar trotz des erwähnten ungünstigen Verhältnisses zwischen der Zahl benötigter Verteilungen und der Stichprobengröße. Die Fehlerraten variieren hier insgesamt sogar nur um 1,18% von einer Initialisierungsvariante zur anderen.

Andererseits liegen die Fehlerraten auf der Gesamtstichprobe betrachtet jedoch deutlich über den zuvor durch Weglassen der Teilinkremente erreichten Werten und bei „toleranter“ weicher Summenbildung erheblich höher als für die „exakten“ Initialisierungsvarianten. Die vorherige Alternative der Modellgestaltung erfaßte also, gleichwohl in Gestalt statistischer Zusammenhänge, mehr relevante Faktoren als die jetzige Realisierung. Es bietet sich jedoch auch ein anderer Vergleich der Evaluationszahlen an, und zwar mit derjenigen Modellva-

riante, die die Summenformelinformation einschließlich der Multiplizitäten der einzelnen Atomspezies berücksichtigt und eine ungerichtete Abhängigkeit zwischen Summenformel und Molekülstruktur beinhaltet (vgl. Tabelle 10.5). Ihr gegenüber ist hier nur im Bereich der Inkremente und Teilinkremente eine Veränderung vorgenommen worden, und zwar eine klarere Trennung von Literaturwissen und statistischen Anteilen.

Während für den *ipso*-Klassifikationsschritt im Falle einer „empirischen“ Initialisierung für beide untersuchten Mengen eine deutliche Verschlechterung der Erkennungsleistung zu verzeichnen ist, ist diese für den *ortho*-Klassifikationsschritt geringer. Bei „gleichverteilter“ Initialisierung jedoch ist neben der bereits erwähnten Tatsache, daß sich die Fehlerraten kaum von denen bei „empirischer“ Initialisierung unterscheiden, eine deutlich verbesserte Fehlerrate auf beiden Stichproben zu beobachten.

Dies kann dahingehend interpretiert werden, daß durch die Trennung von Literaturwissen und statistischen Anteilen („Erfahrung“) die bislang recht starke Initialisierungsabhängigkeit erheblich verringert wurde. Dies bedeutet zwar, daß nun statistische Faktoren, die durch eine „empirische“ Initialisierung ins Modell eingebracht werden, die Erkennung nicht mehr so stark positiv beeinflussen können, es bedeutet aber zugleich, daß statt dessen offensichtlich die im engeren Sinne auf Fachwissen basierenden Modellanteile ein größeres Gewicht erhalten, was sehr im Sinne des ursprünglichen Ansatzes ist.

Die alternative Modellierung hat darüber hinaus den Vorteil, daß man in sehr analoger Vorgehensweise zu den bereits repräsentierten weitere Sphären hinzufügen kann, um eine präzisere Wiedergabe der Molekülstruktur zu erreichen. Soll bei gegebener chemischer Verschiebung ein bestimmtes Inkrement empirisch ermittelt werden, so müssen alle übrigen Inkremente feststehen, da man sonst eine Gleichung mit mehreren Unbekannten zu lösen hätte. Das ursprüngliche Modell und seine bisherigen Varianten folgten dabei im Prinzip der Idee, stets die Inkremente der unteren Sphären aus Literaturwissen zu bestimmen und nur in der höchsten explizit repräsentierten Sphäre einen Anteil empirisch zu gewinnen. Dieser erfaßte damit implizit auch die Einflüsse der entfernteren Atome in nicht explizit repräsentieren Sphären. Wollte man nun die Modellierungsgenauigkeit um eine Sphäre erweitern, so müßte deren Einfluß aus der summenhaften Betrachtung der zuvor höchsten Sphäre herausgezogen und die Modellierung somit erst modifiziert werden, bevor sie erweitert werden und der empirische Anteil auf die neue, nun höchste Sphäre übergehen kann.

In der neu entwickelten Realisierung dagegen sind die Inkrementbeiträge aller explizit repräsentierten Positionen von vornherein auf Expertenwissen bezogen. Die Variable `Einflüsse` übernimmt jegliche empirischen Anteile, und zwar betreffend sowohl nicht explizit repräsentierte Strukturanteile wie auch Korrekturen, die betreffend des Zusammenwirkens der repräsentierten Sphären nötig sind. Änderungen des Bestehenden sind somit nicht notwendig, wenn in Analogie zu den bereits im Modell enthaltenen Sphären weitere zur Erhöhung der Modellierungsgenauigkeit integriert werden sollen.

Die gegenüber den bisherigen Bestwerten im *ipso*-Klassifikationsschritt deutlich höheren Fehlerraten und die stark variierenden Fehlerzahlen der einzelnen Kreuzvalidierungsdurchgänge sind unabhängig von der untersuchten Stichprobe auf die knappe Datenlage zurückzuführen: $P(\texttt{Einflüsse}|\texttt{C\_arom},\texttt{s2ortho})$ benötigt nicht nur, wie bereits erwähnt, 364 Verteilungen, sondern jede Verteilung deckt auch nicht weniger als 83 Werte ab. Es gibt also 30212 mögliche Konstellationen der drei beteiligten Variablen, und folglich wäre ein Vielfaches dieser Zahl an Beispielen erforderlich, um die entsprechenden Verteilungen verläßlich schätzen zu können.

In der Konsequenz ergeben sich für viele der entsprechenden Verteilungen Gleichverteilungen, weil die entsprechenden Zustandskombinationen der Elternvariablen `C_arom` und

s2ortho überhaupt nicht beobachtet werden. Dies ist zum einen ein Resultat des bei der empirischen Schätzung bedingter Wahrscheinlichkeiten angewandten *Adding-One* Verfahrens (vgl. Abschnitt 8.3.2 sowie [Fin03], S. 105 ff.), das gerade bei einer knappen Datenlage zu einer Übergewichtung der nicht beobachteten Ereignisse führt. Diese Überbewertung wirkt sich insbesondere für die Evaluation der monosubstituierten Verbindungen nachteilig aus, da hier einige Zustandskombinationen gar nicht auftreten können.

Somit könnte ein Ansatz zur weiteren Verbesserung in einer Neugestaltung der Zustände der Variablen s2ortho liegen, wie sie im folgenden Abschnitt beschrieben wird. Außerdem besteht die Möglichkeit, daß sich die Zahl der Zustände der Variablen Einflüsse reduzieren läßt, wenn in zukünftigen Weiterentwicklungen die Modellierungsgenauigkeit weiter erhöht wird und somit mehr relevante Faktoren explizit erfaßt werden. Der gleichermaßen interessanteste wie auch facettenreichste (und damit anspruchsvollste) Ansatz besteht darüber hinaus in einem anderen Umgang mit nicht beobachteten Ereignissen. In diesem Punkt reichen die Möglichkeiten von Untersuchungen betreffend geeignete (Standard-)Verteilungen bis hin zur Entwicklung eines ganz neuen, eigenen Initialisierungskonzepts für $P(\texttt{Einflüsse}|\texttt{C\_arom},\texttt{s2ortho})$, was einen deutlichen Schritt in Richtung der ursprünglichen Idee von Kreuztermen zur Wiedergabe des Zusammenspiels der Einflüsse bedeuten würde. Diesbezügliche Entwicklungen haben jedoch viel Spielraum und sollten wohl überlegt durchgeführt werden, so daß ihnen in zukünftigen Arbeiten gesondert Aufmerksamkeit gewidmet werden muß.

Zum Abschluß der augenblicklichen Überlegungen betreffend die Repräsentation der Inkremente und Teilinkremente sollen die dabei gewonnenen Erkenntnisse nun noch einmal zusammengefaßt werden.

- Die Verbesserung der Erkennungsleistung bei Entfernen der Variablen für die Teilinkremente der zweiten Sphäre ist ein Hinweis darauf, daß die tatsächlichen Gegebenheiten des Zusammenwirkens der einzelnen Beiträge noch nicht adäquat repräsentiert sind.

- Diese Änderung führt jedoch zu einer Verschiebung weg von der wissensorientierten Modellierung hin zu einer stärker auf statistischen Zusammenhängen basierenden Sichtweise. Dies entspricht nicht dem ursprünglich verfolgten Ansatz.

- Eine klarere Trennung von Literaturwissen und empirischen Anteilen führt dagegen zu einem größeren Gewicht des modellierten Wissens und schlägt sich in geringeren Unterschieden zwischen den einzelnen Initialisierungsvarianten nieder.

- Durch diese Änderung wird die Integration zusätzlicher Variablen für eine auf mehr als zwei Sphären genaue Strukturbeschreibung erleichtert.

- Verluste in der Erkennungsleistung bei „empirischer“ Initialisierung können auf eine unzureichende Datenlage und eine zu geringe Modellierungsgenauigkeit im Sinne des vorgenannten Punktes zurückgeführt werden.

10.3.3 Variationen der Zustandsgestaltung

Die Zustände von s2ipso und s2ortho wurden basierend auf einer statistischen Analyse der Stichprobe festgelegt. Jede einzelne theoretisch in Frage kommende Belegung der betreffenden Position wurde dabei als ein Elementarereignis angesehen, und die statistische Analyse diente dazu, dieselben zusammenzufassen und so auf eine handhabbare Zahl von Zuständen

abzubilden. In diesem Punkt ist jedoch sicherlich Raum für alternative, nicht minder sinnvolle Zusammenfassungen.

Betrachtet man die Zustände von `s2ortho`, so ist festzuhalten, daß sich die Elementarereignisse der aktuellen Zustände strukturell jeweils klar voneinander abgrenzen lassen. Zieht man jedoch die zugehörigen Inkrementbeiträge hinzu, so zeigen sich zum einen für ein und denselben Zustand mehrere verschiedene Intervalle, in welchen der Betrag des Inkrements liegen kann. Zum anderen kommt hinzu, daß ein und dasselbe Intervall im Zusammenhang mit unterschiedlichen `s2ortho`-Zuständen auftritt. Wenngleich es zweifellos sinnvoll ist, bei der Repräsentation *struktureller* Moleküleigenschaften die Zusammenfassung der Elementarereignisse zu Zuständen nach *strukturellen* Gesichtspunkten vorzunehmen, so liegt angesichts der erwähnten Beobachtung zugleich eine Alternative nahe: Es wäre ebenso denkbar, den Betrag des zugehörigen Inkrements als Charakteristikum bei der Zusammenfassung mit einzubeziehen.

Zu diesem Zweck werden die einzelnen betrachteten Substituentenklassen nun nach Stärke und Vorzeichen ihres zugehörigen *ortho*-Inkrements zu sieben Gruppen (von stark positiv (Iod) über neutral (Wasserstoff) bis stark negativ) zusammengefaßt. Die entsprechende Überarbeitung von `s2ortho` und die Einordnung der einzelnen Substituentenklassen sind in Abbildung 10.6 dargestellt. Tabelle 10.9 gibt die Evaluationsergebnisse des Modells wieder, das neben der Trennung von Literaturwissen und empirischen Anteilen durch Einbringung der Variablen `Einflüsse` (vgl. Tabelle 10.8) auch die beschriebene Neugestaltung von `s2ortho` beinhaltet. Auch hier wurde aufgrund des Umfangs von $P(\texttt{Einflüsse}|\texttt{C\_arom}, \texttt{s2ortho})$ auf eine getrennte Evaluation von Verbindungen mit kurzen Substituenten verzichtet.

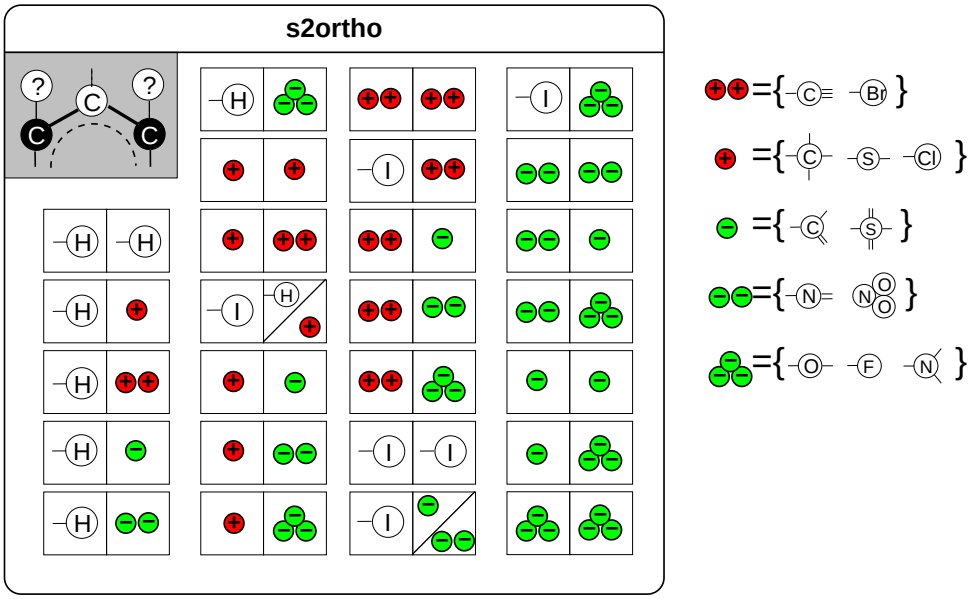


Abb. 10.6: *Änderung der Gestaltung der Variablen* `s2ortho`*. Die in den beiden* ortho*-Positionen vorkommenden Substituenten werden nun nach Stärke und Vorzeichen ihres zugehörigen Inkrements kategorisiert.*

Die im Bereich der monosubstituierten Verbindungen im *ipso*-Klassifikationsschritt festzustellende minimale Verschlechterung im Vergleich zur vorhergehenden Modellvariante, die ebenfalls Literaturwissen und empirische Anteile durch Einführung der Variable `Einflüsse` trennte, sollte angesichts der von einem Kreuzvalidierungslauf zum anderen stark schwan-

Initialisierung	Gesamtmenge	Monosubstituierte
empirisch,hart,exakt	ipso: 36,85% ortho: 61,81%	ipso: 21,30% ortho: 26,50%
empirisch,hart,tolerant	ipso: 36,48% ortho: 64,91%	ipso: 21,30% ortho: 30,00%
empirisch,weich,tolerant	ipso: 36,93% ortho: 65,46%	ipso: 21,30% ortho: 33,52%
empirisch,weich,exakt	ipso: 37,67% ortho: 62,01%	ipso: 21,30% ortho: 27,78%
gleichverteilt,weich,exakt	ipso: 37,56% ortho: 71,57%	ipso: 20,14% ortho: 29,12%
gleichverteilt,weich,tolerant	ipso: 31,72% ortho: 75,19%	ipso: 20,69% ortho: 29,88%
gleichverteilt,hart,tolerant	ipso: 32,07% ortho: 75,89%	ipso: 20,69% ortho: 30,44%
gleichverteilt,hart,exakt	ipso: 39,40% ortho: 72,00%	ipso: 22,26% ortho: 29,65%

Tab. 10.9: *Evaluation des Modells nach Neugestaltung der Variablen* `s2ortho`

kenden Fehlerzahlen nicht überbewertet werden. Viel interessanter ist, daß hier erstmals „gleichverteilte“ Initialisierungsvarianten in drei der vier Fälle besser abschneiden als die entsprechenden „empirischen“ Varianten. Auch dabei gilt gleichwohl das oben gesagte, und es wäre mit Sicherheit aufschlußreich, die Evaluation mit einer deutlich größeren Stichprobe monosubstituierter Verbindungen zu wiederholen.

Auf der umfangreicheren Gesamtstichprobe zeigt sich im *ipso*-Klassifikationsschritt sowohl bei „gleichverteilter“ wie auch bei „empirischer“ Initialisierung in den Varianten mit „toleranter“ weicher Summenbildung eine deutliche Verbesserung der Erkennungsleistung. In den übrigen Fällen blieben die erzielten Werte in etwa konstant, die Erkennungsraten liegen durch die beschriebene Verbesserung nun jedoch insgesamt dichter zusammen. Interessant ist, daß die beste Erkennungsleistung bei diesem Modell bei „gleichverteilt-weich-toleranter“ Initialisierung erreicht wird. Die Aufweichung der Struktur-Inkrement-Zuordnung sowie das Erlauben von Abweichungen bei der Summation der einzelnen Beiträge leisten also gerade bei der nun gegebenen stärkeren Gewichtung explizit repräsentierten Literaturwissens ihren Beitrag dazu, eine beobachtete Peakposition trotz geringfügiger Abweichungen vom strikten „Lehrbuchwert“ auf die richtigen Strukturmerkmale zurückzuführen.

Darüber hinaus werden bei „gleichverteilter“ Initialisierung relativ zu dem Modell mit Wiedergabe der ungerichteten Abhängigkeit im Bereich der Summenformel, jedoch ohne Änderung des Zusammenspiels der Inkremente (Tabelle 10.5) vergleichbare oder bessere Resultate erzielt. Die Ergebnisse des Modells, das auf die Unterteilung der Inkremente der zweiten Sphäre verzichtete (Tabelle 10.7), konnten jedoch nicht wieder erreicht werden. Da in letzterem Modell statistische Zusammenhänge stärker betont werden, kann dies so interpretiert werden, daß die in ihnen verborgenen Gesetzmäßigkeiten noch immer nicht treffend erfaßt werden, daß die gemachten Änderungen jedoch dennoch ein erfolgversprechender Ansatz sein könnten. Die Idee der Trennung von Literaturwissen und empirischen Anteilen ist also weiterzuverfolgen und eingehender zu untersuchen.

Die Untersuchung des *ortho*-Klassifikationsschritts wirft jedoch neue Fragen auf. Für mo-

nosubstituierte Verbindungen ist gegenüber dem Vergleichsmodell mit der ursprünglichen s2ortho-Gestaltung eine Verbesserung der Fehlerrate festzustellen, die je nach Initialisierungsvariante bis zu 11,5% (absolut) beträgt. Daraus könnte man ableiten, daß eine kategorienartige Erkennung der *ortho*-Klasse gegenwärtig besser möglich ist als eine Unterscheidung einzelner Kombinationen von Substituenten in den *ortho*-Positionen.

Für die Gesamtstichprobe ist dagegen nur bei „empirischer“ Initialisierung eine leichte Verbesserung um 2–5% (absolut) festzustellen; bei „gleichverteilter“ Initialisierung ist dagegen eine Verschlechterung um 4–9% (absolut) zu verzeichnen. Eine mögliche Ursache liegt hier in den Einflüssen der Substituenten in den von Modell bislang nicht explizit erfaßten Ringpositionen: Wie im vorigen Abschnitt beschrieben sind diese aufgrund der Vielzahl möglicher Effekte, die sich im Modell vor allem im kaum handhabbaren Umfang der Wahrscheinlichkeit $P(\texttt{Einflüsse}|\texttt{C\_arom},\texttt{s2ortho})$ zeigt, aus dem Blickwinkel des Klassifikators kaum einzuschätzen. Diese Einflüsse nicht direkt benachbarter Substituenten machen selbst eine kategorisierte Betrachtung der *ortho*-Positionen schwierig.

In nächsten Schritt ist nun die Frage interessant, ob sich die verbesserte Erkennung bei kategorienhafter Betrachtung für monosubstituierte Verbindungen auf ein Modell übertragen läßt, das die ursprüngliche Gestaltung des Struktur-Inkrement-Zusammenhangs (vor Einbringung der Variablen Einflüsse) beibehält, und wie sich ein solches Modell bezogen auf die Gesamtstichprobe verhält. Tabelle 10.10 zeigt die Evaluierung einer Modellvariante, die strukturell derjenigen vor Einführung der Variablen Einflüsse entspricht (vgl. Tabelle 10.5), jedoch die oben beschriebene Neugestaltung von s2ortho enthält.

Initialisierung	Gesamtmenge	Monosubstituierte
empirisch,hart,exakt	ipso: 29,94% ortho: 65,42%	ipso: 15,16% ortho: 34,28%
empirisch,hart,tolerant	ipso: 29,27% ortho: 67,96%	ipso: 16,30% ortho: 35,02%
empirisch,weich,tolerant	ipso: 29,38% ortho: 67,61%	ipso: 17,04% ortho: 35,93%
empirisch,weich,exakt	ipso: 29,49% ortho: 65,83%	ipso: 16,67% ortho: 35,74%
gleichverteilt,weich,exakt	ipso: 38,39% ortho: 69,32%	ipso: 28,38% ortho: 38,17%
gleichverteilt,weich,tolerant	ipso: 41,76% ortho: 69,34%	ipso: 43,19% ortho: 43,63%
gleichverteilt,hart,tolerant	ipso: 40,33% ortho: 69,23%	ipso: 43,19% ortho: 46,04%
gleichverteilt,hart,exakt	ipso: 38,77% ortho: 69,08%	ipso: 30,60% ortho: 34,65%

Tab. 10.10: *Evaluation der Modellvariante ohne Einführung der Variablen* Einflüsse*, jedoch mit Neugestaltung der Zustände der Variablen* s2ortho

Die geringfügige Steigerung der Fehlerrate im *ipso*-Klassifikationsschritt der Gesamtstichprobe läßt sich mit der Kategorisierung der *ortho*-Substituenten erklären: Dadurch entsteht eine größere Unsicherheit, welche der in der Summenformel aufgezählten Atome noch zur Verfügung stehen, als dies bei der Betrachtung individueller Substituentenkombinationen in den *ortho*-Positionen der Fall ist. Bei den monosubstituierten Verbindungen ist dagegen bei

„gleichverteilter“ Initialisierung eine leichte Verbesserung zu beobachten, die Verluste bei „empirischer“ Initialisierung sind dagegen um so gravierender. Offenbar wirkt sich hier die zusätzliche Unschärfe durch die Kategorisierung der *ortho*-Substituenten besonders negativ aus.

Vergleicht man jedoch die Änderungen miteinander, die sich jeweils zwischen denjenigen Modellen mit sowie ohne explizite Trennung von Wissen und empirischen Anteilen ergeben, die sich nur um die Gestaltung der Zustände von `s2ortho` unterscheiden (Tabelle 10.8 gegenüber 10.9 sowie Tabelle 10.5 gegenüber 10.10), so ist die Verbesserung für „gleichverteilt-tolerant“ initialisierte Netze, die sich bei den Modellen mit expliziter Trennung auf der Gesamtstichprobe zeigte, im anderen Fall nicht vorhanden. Jedoch liegen auch hier die erzielten Fehlerraten nun insgesamt enger zusammen. Auf den monosubstituierten Verbindungen veränderten sich die Ergebnisse in den Modellen mit expliziter Trennung nur wenig. Im anderen Fall zeigt sich bei „empirischer“ Initialisierung eine deutliche Verschlechterung und bei „gleichverteilter“ Initialisierung eine Verbesserung.

Für die Interpretation der Ergebnisse kann zunächst einmal mehr festgehalten werden, daß in der Gesamtstichprobe zu viele vom Modell nicht erfaßte Faktoren eine zu erhebliche Rolle spielen, insbesondere die Einflüsse von Substituenten in vom Modell noch nicht erfaßten Positionen des Benzolringes. Betreffend die monosubstituierten Verbindungen, in welchen diese Einflüsse keine so starke Rolle spielen, kann man sagen, daß auch die Evaluation nach Neugestaltung der Variablen `s2ortho` die stärkere Gewichtung des repräsentierten Literaturwissens bei dessen klarerer Abtrennung von empirischen Anteilen bestätigt. Darüber hinaus gibt es Hinweise, daß dadurch auch ein besserer Umgang mit systematischen Unschärfen (konkret der Kategorisierung von *ortho*-Substituenten sowie eine „tolerante“ weiche Summenbildung) erreicht wird, während dies in der ursprünglichen Repräsentation der Inkremente und Teilinkremente nicht der Fall ist, sondern diese Modellvariante statt dessen einem stärkeren Einfluß statistischer Faktoren, die etwa in Form einer „empirischen“ Initialisierung eingebracht werden, unterworfen ist. Weitere Untersuchungen wären jedoch angezeigt, um dies zu belegen sowie auch die Faktoren, die in der Gesamtstichprobe neben den im Modell erfaßten eine Rolle spielen, näher zu betrachten.

Aktuell sind jedoch die Erkennungsraten im *ortho*-Klassifikationsschritt noch von Interesse, da der Gedanke im Raum steht, daß eine kategorisierte Betrachtung der *ortho*-Positionen vom derzeitigen Modell besser realisiert wird als die exakte Erkennung einzelner Substituentenkombinationen. In der Tat ist bei der zuletzt untersuchten Modellvariante bei „gleichverteilter“ Initialisierung nun auch auf der Gesamtstichprobe eine deutliche Verbesserung zum Referenzmodell um 10% oder mehr (absolut) festzustellen, was zuvor in erster Linie auf Basis monosubstituierter Derivate beobachtet wurde. Bei diesen schwankt nun die Verbesserung zwischen 2,5% und 24,5%. Bei den verschiedenen „gleichverteilten“ Initialisierungen gehen in jedem Fall die Unterschiede der Fehlerraten untereinander zurück. Bei „empirischer“ Initialisierung ist durch die unschärfere Zustandsgestaltung auf beiden Stichproben eine leichte Verschlechterung zu bemerken.

Vergleicht man nun wiederum jeweils die Änderungen anstatt der konkreten Fehlerzahlen, so zeigt sich auf beiden Stichproben bei „empirischer“ Initialisierung im einen Fall eine leichte Verbesserung und im anderen eine leichte Verschlechterung. Dies steht im Einklang mit obiger Theorie, jedoch muß nach wie vor der Bedarf weiterer Untersuchungen, insbesondere nach einer Erhöhung der Modellierungsgenauigkeit unter Berücksichtigung der *meta*- und *para*-Einflüsse, betont werden.

Darüber hinaus wird nun auf der Gesamtstichprobe bei „gleichverteilter“ Initialisierung eine bessere Erkennungsrate für Kategorien von *ortho*-Substituenten erreicht als bei der ur-

sprünglichen Zustandsgestaltung. Dies läßt sich damit erklären, daß die empirischen Faktoren, die sich hier positiv auswirken, in der anderen Modellierung mit klarer Abgrenzung von Literaturwissen und empirischen Anteilen abgewichtet werden. Darin wird statt dessen, wie bereits erwähnt, der Wissensanteil stärker gewichtet. Eine Steigerung des Detailsgrads in der Modellierung sollte somit um so nachdrücklicher der nächste anzustrebende Schritt sein.

Bis dahin sollte jedoch weder die eine noch die andere Modellvariante ausgeschlossen werden. Während die ursprüngliche Gestaltung die Möglichkeit mit sich bringt, daß sich empirische Faktoren positiv auswirken, hat die Neugestaltung den Vorteil einer leichteren Erweiterbarkeit, wenn die Strukturrepräsentation weiter ausgearbeitet werden soll, was zugleich eine detailliertere Wiedergabe der damit verbundenen Einflüsse bedeutet.

Betreffend die Erkennungsleistung im *ortho*-Klassifikationsschritt läßt sich anhand der Beobachtungen für die „gleichverteilten“ Initialisierungsvarianten festhalten, daß der tendenzielle Einfluß der Substituenten, wie er durch eine kategorienartige Zustandsgestaltung dargestellt wird, bei der gegenwärtigen Modellierungsgenauigkeit besser erfaßt wird als die exakte Substituentenkombination.

Es ist jedoch fraglich, insbesondere bei der Betrachtung der monosubstituierten Verbindungen, ob eine unschärfere Modellierung, wie sie die Neugestaltung der Zustände von `s2ortho` bedeutet, angestrebt werden sollte. Im Grunde liefe dies der anzustrebenden Erhöhung der Modellierungsgenauigkeit, insbesondere im Bereich der Molekülrepräsentation, entgegen.

In ähnlicher Weise wie für `s2ortho` beschrieben könnte man auch eine Neugestaltung von `s2ipso` anstreben, sei es zur Überprüfung eines ähnlichen Verhaltens wie im Falle der *ortho*-Position oder vor dem Hintergrund einer möglichen Verbesserung der Erkennungsleistung. Aufgrund der Abhängigkeit der möglichen Klassen von der Valenz der in `C_arom` wiedergegebenen ersten Sphäre ist hier jedoch eine Unterteilung anhand der Atome in den einzelnen Plätzen der *s2ipso*-Position nach der Stärke ihres jeweiligen Einflusses nicht ohne weiteres möglich. Insofern sind an dieser Stelle weitere Überlegungen nötig, um eine Präzisierung des Modells zu erreichen.

Ein Fazit der Untersuchung der Neugestaltung von `s2ortho` ist in Form einer klaren Aussage oder Richtlinie für zukünftige Weiterentwicklungen des Modells schwer zu geben. Vielmehr zeigen die gefundenen Indizien und darauf basierenden Vermutungen einen deutlichen Bedarf weiterer Untersuchungen sowie einer Weiterentwicklung des Modells, um Einflüsse explizit zu erfassen, die bislang hauptsächlich quasi „versteckt“ in statistischen Faktoren, vor allem bei den „empirischen“ Initialisierungsvarianten zutagetraten.

10.4 Zusammenfassung der Ergebnisse

Das in Kapitel 6 entwickelte Kausalmodell wurde als Bayes-Netz auf unterschiedliche Weise initialisiert und die Zuverlässigkeit der Klassifikation bei der Betrachtung der *ipso*-Position und der *ortho*-Positionen untersucht. Die Evaluation fand auf der Gesamtstichprobe, einer Auswahl ausschließlich monosubstituierter Verbindungen und einer Auswahl von Verbindungen, deren Substituenten maximal drei Bindungen lang waren, jeweils in Form eines 10fach-Kreuzvalidierungsverfahrens statt.

Insgesamt hat sich der auf expliziter Wissensrepräsentation basierende Ansatz, realisiert in Gestalt eines Bayes-Netzes, als grundsätzlich geeignet erwiesen. Ausgehend von einem initialen Modell wurde dasselbe in mehreren Schritten verbessert, und es wurden dabei Aspekte aufgezeigt, die bei seiner Weiterentwicklung eine Rolle spielen können oder sollten.

Bei den ersten Untersuchungen betreffend die Repräsentation der Summenformelinformation zeigte sich klar, daß es in diesem Punkt von Bedeutung ist, die gesamte Information zu nutzen und neben der Präsenz der einzelnen chemischen Elemente auch die Multiplizitäten der jeweiligen Atome zu berücksichtigen. Außerdem wird durch die Repräsentation des Zusammenhangs zwischen Summenformel und Molekülstruktur als ungerichtete Abhängigkeit eine weitere Verbesserung erzielt, verglichen mit der Beschränkung auf die Sichtweise der Summenformel als Konsequenz der Molekülstruktur.

Die besten Erkennungsraten im *ipso*-Klassifikationsschritt wurden für monosubstituierte Verbindungen mit derartigen Modellen erreicht. Für das zuletzt beschriebene betrug die Korrektklassifikationsrate für die empirische Initialisierungsvariante mit einer weichen Strukturmerkmal-Inkrement-Zuordnung und ohne Toleranzintervall bei der weichen Summenbildung 96,64% im *ipso*- und 66,67% im *ortho*-Klassifikationsschritt.

Ein Modell, das darüber hinaus auf die Untergliederung des Beitrags der zweiten Sphäre zur chemischen Verschiebung in einen *ipso*- und einen *ortho*-Anteil verzichtete, erreichte mit derselben Initialisierungsvariante, jedoch harter Strukturmerkmal-Inkrement-Zuordnung die beste Korrektklassifikationsrate im *ortho*-Klassifikationsschritt von 87,92% und zugleich mit 96,23% im *ipso*-Klassifikationsschritt ein kaum unter obigem Bestwert liegendes Ergebnis.

Wenngleich die weiteren Verbesserungen, die diese Modellvariante erreichte, ein deutliches Signal dafür ist, daß das Zusammenwirken der Beiträge der einzelnen Strukturbestandteile zur chemischen Verschiebung noch nicht adäquat wiedergegeben wird, so ist dieser Ansatz jedoch quasi ein Schritt in die falsche Richtung: Im Prinzip führt er von der wissensorientierten Modellierung weg und zu einer stärker auf statistischen Zusammenhängen basierenden Sichtweise und entspricht somit nicht dem ursprünglich verfolgten Ansatz.

Zudem relativiert die Tatsache, daß bis hierher stets die empirischen Initialisierungsvarianten zu merklich besseren Ergebnissen führten als dasselbe Modell mit einer gleichverteilten Initialisierung, obige Werte ein wenig. Die Tatsache spricht für das Wirken nicht vom Modell erfaßter Einflüsse. Hier kommen insbesondere die Einflüsse von Substituenten in den *meta*-Positionen und der *para*-Position in Frage, die wegen der auf zwei Sphären begrenzten Strukturbetrachtung bislang nicht explizit repräsentiert sind. Untersuchungen und Weiterentwicklungen in diesem Bereich sollten der erste Schritt zukünftiger Arbeiten sein.

Das beste Resultat bei einer gleichverteilten Initialisierungsvariante wurde, wiederum für monosubstituierte Verbindungen, mit einer Korrektklassifikationsrate von etwa 80% im *ipso*-Klassifikationsschritt für folgendes Modell erzielt: Als Alternative für die Repräsentation des Zusammenspiels der einzelnen Inkrementbeiträge wurde eine klare Aufteilung derselben in Literaturwissen und empirische Anteile untersucht. Auf diese Weise wurde das Gewicht des modellierten Wissens erhöht, was sich in geringeren Unterschieden zwischen den einzelnen Initialisierungsvarianten zeigte. Insbesondere führten empirische und gleichverteilte Initialisierung des Summenformel-Struktur-Zusammenhangs nun zu sehr ähnlichen Werten. Ein weiterer Vorteil dieser Variante ist die Möglichkeit der unproblematischen Integration einer detaillierteren Wiedergabe der Molekülstruktur sowie der korrespondierenden Einflüsse.

Dies ist besonders von Interesse, da mit der gegenwärtigen, mit der Beschränkung auf zwei Sphären bzw. hinsichtlich der Ringpositionen auf die *ortho*-Positionen nicht sehr genauen Modellierung die unterschiedlichen strukturellen Konstellationen nicht ausreichend gut voneinander abgegrenzt werden können. Es wirken neben den vom Modell abgedeckten weitere Einflüsse, deren Integration konsequenterweise im nächsten Schritt anzustreben ist. Dabei können die bisherigen Ergebnisse und die bis hierher gesammelten Erfahrungen für ein zielgerichtetes Vorgehen ausgenutzt werden.

11 Zusammenfassung und Ausblick

Unter Strukturaufklärung in der organischen Chemie versteht man das Aufdecken struktureller Eigenschaften organischer Moleküle. Die NMR-Spektroskopie zählt dabei zu den Methoden mit der größten Bedeutung. Dank moderner Methoden kann der gesamte Prozeß heutzutage vom technischen Standpunkt her als Routineaufgabe angesehen werden; mit den technischen Neuerungen hat sich jedoch der Engpaß innerhalb der Arbeitsabläufe von der Untersuchung der Probe in den Bereich der Auswertung der dadurch gewonnenen immer größeren Datenmengen verlagert.

Daher ist die Entwicklung von Computerprogrammen in diesem Bereich ein wichtiges Ziel in der modernen organischen Chemie. Der klassische Ansatz folgt dabei einem Grundprinzip, das bereits in den Anfängen der Strukturaufklärung Gültigkeit besaß, und geht in zwei Schritten vor: Zuerst werden basierend auf der Summenformel der unbekannten Substanz alle infragekommenden Molekülstrukturen aufgelistet. Dann werden die NMR-Spektren der hypothetischen Strukturkandidaten mit dem experimentellen Spektrum der unbekannten Substanz verglichen, um die Hypothesen zu bewerten.

NMR-Spektren enthalten jedoch aufgrund des Zusammenhangs zwischen spektralen und strukturellen Eigenschaften eines Moleküls eine Vielzahl von Informationen, welchen die alleinige Verwendung zu Vergleichszwecken kaum gerecht wird. Dementsprechend wäre es für einen Menschen, der mit einem Strukturaufklärungsproblem befaßt ist, viel naheliegender, sich durch die Auswertung des Spektrums die darin codierte Strukturinformation zu erschließen, anstatt wahllos alle mit der gegebenen Summenformel übereinstimmenden Strukturen zu notieren und erst im Nachhinein eine nach der anderen zu validieren. Auch in aktuellen Systemen wird der Ansatz verfolgt, Generierung und Validierung von Strukturhypothesen stärker mit einander zu verflechten, da es wenig sinnvoll erscheint, ohne jeglichen Vorannahmen den gesamten Raum möglicher Molekülstrukturen vollständig zu durchsuchen.

Während dabei in der Regel das Konzept eines bestehenden Systems verändert wird, um es durch die Idee der Integration der beiden Schritte der Hypothesengenerierung und -validierung zu bereichern, wurde in der vorliegenden Arbeit mit dem System SASCHA ein ganz neues System konzipiert, das sich am menschlichen Vorgehen orientieren sollte. Das Kürzel SASCHA bedeutet „**s**achlich **a**rgumentierendes **S**ystem für die **ch**emische **A**nalyse“ und nimmt Bezug auf das hinter der Entwicklungsidee stehende Ideal: ein System, welches auf der Basis eines analogen Vorgehens mit dem Menschen in eine Art sachlichen Dialog eintreten und ihm nicht nur seine Einschätzung darlegen kann, sondern auch in der Lage ist zu belegen, welche der beobachteten Fakten diese im Detail untermauern. Sicherlich liegt ein derartiges Fachgespräch zwischen Mensch und Computersystem heutzutage noch im Bereich der Utopie, doch die Idee eines solchen „virtuellen Laborassistenten“ ist zweifellos ein guter (wenngleich ehrgeiziger) Leitfaden.

Es sollte also ein Musteranalysesystem entstehen, welches aus dem zu analysierenden Spektrum (Muster) das Substitutionsmuster eines Benzolringes (symbolische Beschreibung) gewinnt. Methoden und Systemaufbau konnten dabei so gewählt werden, daß sie dem beschriebenen Ansatz entsprechen, ohne auf die Gegebenheiten eines bereits bestehenden Systems Rücksicht nehmen zu müssen. Der Schwerpunkt wurde dabei auf das explizite Einbringen von Expertenwissen gelegt, was im Sinne des Nachahmens einer menschlichen Vorgehensweise naheliegend ist.

Durch die Einschränkung auf Substitutionsmuster von Benzolderivaten wurde die Musteranalyseaufgabe auf eine überschaubare, zugleich aber nicht triviale Teilmenge der immensen Vielfalt organischer Verbindungen eingeschränkt, um einzelne Aspekte insbesondere der Wissensrepräsentation als dem Schwerpunkt des Systems im Detail untersuchen und evaluieren zu können. Darüber hinaus war das Ziel der Arbeit nicht, ein Strukturaufklärungssystem zu schaffen, das vom ersten Tag an die Leistung aktueller Systeme erreicht, welche auf jahrelange Weiterentwicklungen, Verbesserungen und Erfahrungen zurückblicken. Vielmehr sollte mithilfe des entwickelten Forschungssystems gezeigt werden, daß der an der Herangehensweise des Menschen orientierte Ansatz und die in diesem Zusammenhang zur Wissenrepräsentation genutzte Methodik von Bayes-Netzen grundsätzlich für ein Strukturaufklärungssystem geeignet sind. Im Bereich der Wissensrepräsentation sind dabei einzelne Modellierungsschritte im Detail untersucht und mehrere alternative Entscheidungen evaluiert worden. Unabhängig von der absolut zu erreichenden Leistung sollten dadurch wertvolle Erkenntnisse gewonnen werden, die für zukünftige Arbeiten genutzt werden können.

Zunächst waren jedoch einige Grundvoraussetzungen zu schaffen, um überhaupt ein einsatzfähiges Spektrenauswertungssystem zu erhalten. Das System SASCHA wurde modular konzipiert, wobei jedes Modul einen in sich abgeschlossenen Aufgabenbereich hat; zugleich greifen sie aber, verbunden über ein steuerndes Kontrollelement, wie Zahnräder ineinander, um die gestellte Musteranalyseaufgabe zu lösen (vgl. Kapitel 5). Die einzelnen Teilbereiche sind dabei Datenvorverarbeitung, Wissensrepräsentation, statistische Funktionen, Hypothesengenerierung sowie das erwähnte Kontrollmodul.

Idealerweise wäre darüber hinaus ein Dialogmodul zu realisieren, welches für die Kommunikation mit dem Benutzer zuständig ist. Für das gerade erst neu entstehende Forschungssystem schien es jedoch angemessen, sich zunächst mit dem Fall eines idealen Verarbeitungsverlaufs zu befassen. In diesem Fall ist keine Intervention oder sonstige Einflußnahme des Benutzers über die Dateneingabe hinaus erforderlich, so daß die Interaktion mit dem System minimal ist und eine Ausgabe des Verarbeitungsergebnisses ausreicht.

Der weite Bereich der Mensch-Maschine-Kommunikation wurde somit vorerst ausgespart, er kann und sollte jedoch Gegenstand zukünftiger Arbeiten sein, wenn das heutige Forschungssystem SASCHA dem Ideal seiner Namensgebung folgend zu einem wirklichen intelligenten Werkzeug ausgebaut und praktisch eingesetzt werden soll. Bayes-Netze bieten hierfür einen denkbar günstigen Ausgangspunkt, da sie durch die explizite Wissensmodellierung ein hohes Maß an Nachvollziehbarkeit der Verarbeitungsresultate für den Benutzer ermöglichen. Erst ein nachhaltig realisiertes Dialogmodul könnte dieses Potential jedoch wirklich ausschöpfen, indem es dem Menschen die entsprechende Information zugänglich macht.

Die Betrachtung eines fehlerfreien, reibungslosen Verarbeitungsverlaufs führte weiterhin dazu, daß die Aufgabe des Kontrollmoduls sich derzeit auf das Anstoßen einer linearen Abfolge von Verarbeitungsschritten beschränkt. Ein nächster Entwicklungsschritt könnte darin bestehen, Kriterien und Maßstäbe festzulegen, welche die Ansprüche an die Zwischenergebnisse jedes Verarbeitungsschrittes definieren, und das Einhalten derselben zu kontrollieren. Genügt ein Zwischenergebnis ihnen nicht, wäre dies an das betreffende Modul zurückzumelden. Für jedes Modul wären dann wiederum geeignete Maßnahmen zu entwerfen, die dann ergriffen werden können.

Das Vorverarbeitungsmodul dagegen gehörte naturgemäß zu den elementarsten zu realisierenden Systembestandteilen. Seine Aufgabe ist das Einlesen von Spektren- und Strukturdaten im JCAMP-Format (vgl. Kapitel 7), die Extraktion der für den weiteren Prozeß benötigten Information und die Weitergabe derselben. Diese Aufgabe wurde in zwei Ebenen realisiert, und zwar zum einen einer datennahen, stark auf das JCAMP-Format bezogenen, und zum

anderen einer stärker informationsorientierten, die einen engeren Bezug zur Repräsentation von Molekülstrukturen hat. Diese zweite Ebene ist auf die gewählte Modellierung struktureller Aspekte spezialisiert, während die erstere so allgemein gehalten ist, daß sie sich für viele denkbare Aufgaben im Zusammenhang mit JCAMP-Daten wieder- und weiterverwenden läßt.

Zukünftige Arbeiten im Bereich des Vorverarbeitungsmoduls könnten z.B. die Qualität der verarbeiteten Daten betreffen, indem etwa das Einhalten bestimmter Bedingungen beim NMR-Experiment kontrolliert wird. Im weiteren Kontext kommt auch die automatische Auswahl der zu den Benzolringatomen korrespondierenden Peaks in Frage, jedoch wäre dies zweifellos eine Aufgabe, die nicht auf das Vorverarbeitungsmodul beschränkt ist.

Ebenfalls elementar sind die Aufgaben des Statistikmoduls, das zur Parametrisierung des Bayes-Netzes dient, in welchem das zugrundegelegte Wissen repräsentiert ist (vgl. Kapitel 8). Die *a-priori*- und bedingten Wahrscheinlichkeiten, die darin die Kausalzusammenhänge zwischen spektralen und strukturellen Eigenschaften quantifizieren, können entweder explizit vorgegeben, gemäß bestimmter Gesetzmäßigkeiten berechnet oder empirisch ermittelt werden. Jede dieser drei Möglichkeiten kommt gegenwärtig an gegebener Stelle zum Einsatz.

Das Statistikmodul beinhaltet ebenfalls die Funktionen zur Untersuchung einer Stichprobe, deren Ergebnisse als Grundlage empirisch geschätzter Wahrscheinlichkeiten herangezogen werden. Die Ergebnisse können aber auch genutzt werden, um Entscheidungen bei der Modellentwicklung zu treffen. Im Zuge zukünftiger Weiterentwicklungen im Bereich des Modells ist es insofern zu erwarten, daß sich auch die Notwendigkeit entsprechender Erweiterungen der Fähigkeiten des Statistikmoduls ergibt. Ein nicht unwichtiger Schritt wäre auch eine Verfeinerung des Umgangs mit nicht beobachteten Ereignissen.

Die genannten empirischen Erhebungen können außerdem prinzipiell als das Lernen der Charakteristik einer durch die untersuchte Stichprobe repräsentierten Einsatzumgebung angesehen werden, wenngleich dies eine sehr einfache Form des Lernens ist. Darüber hinaus fand bislang jedoch keine Beschäftigung mit den Möglichkeiten des Lernens statt, die im Zusammenhang mit Bayes-Netzen existieren. Auch dies ist ein möglicher Aspekt zukünftiger Arbeiten.

Die Entwicklung des nun schon mehrfach erwähnten Bayes-Netzes, in welchem das dem System zugrundegelegte Wissen repräsentiert wird, stellt den Schwerpunkt der vorliegenden Arbeit dar, auf welchen im Anschluß detaillierter eingegangen wird. Zuvor soll jedoch die Betrachtung der Zusammenhänge des Gesamtsystems abgeschlossen werden. Innerhalb desselben wird das Bayes-Netz genutzt, um gegeben die Position eines Peaks aus dem Spektrum und die Summenformel der untersuchten Substanz eine Klassifikation der *ipso*-Position sowie der Kombination der beiden benachbarten *ortho*-Positionen vorzunehmen.

Auf diese Weise ergeben sich aus den sechs zu einem Benzolring korrespondierenden Peaks sechs dreigliedrige Ringfragmente. Diese überlappend zusammenzufügen, um so das gesuchte Substitutionsmuster zu erhalten, ist Aufgabe des Moduls zur Hypothesengenerierung. Für dieses wurde unter der Annahme korrekter Klassifikationsresultate ein Verfahren zum Aufbau des Benzolrings entworfen (vgl. Kapitel 9); darüber hinaus wurden einige Überlegungen angestellt, wie im Falle von Konflikten während des Aufbaus verfahren werden sollte. Es zeigte sich jedoch schnell, daß in diesem Zusammenhang umfangreiche Überlegungen anzustellen und zahlreiche Aspekte gegeneinander abzuwägen sind, so daß die Entwicklung einer geeigneten Revisionsstrategie in den Bereich zukünftiger Arbeiten zu verweisen ist.

Die Verläßlichkeit der Klassifikationsresultate spielt dabei eine nicht unbedeutende Rolle, was unmittelbar zur Evaluation des Bayes-Netzes als Klassifikator führt. Gleichwohl wurde dieselbe an dieser Stelle in erster Linie durchgeführt, um zum einen die grundsätzliche Eig-

nung des gewählten Ansatzes für ein Strukturaufklärungssystem zu untersuchen, und um zum anderen die Auswirkung einzelner Modellierungsentscheidungen zu bewerten und festzustellen, welche Bedeutung ihnen bei der Wiedergabe des komplexen Zusammenhangs zwischen Struktur und Spektrum eines organischen Moleküls zukommt.

Zunächst war daher das entwickelte kausale Modell der Zusammenhänge zwischen spektroskopischen Eigenschaften und der Molekülstruktur möglichst einfach und übersichtlich gehalten worden (vgl. Kapitel 6, Abbildung 6.9). Es betrachtet die beobachtete chemische Verschiebung eines Benzolringatoms als aus unterschiedlichen Beiträgen zusammengesetzt, und zwar der Grundverschiebung von 128,5 ppm, einem Beitrag der Atome der ersten Sphäre und einem Beitrag der Atome der zweiten Sphäre. Dieser ist wiederum in einen *ipso*- und einen *ortho*-Anteil untergliedert. Ganz analog wird die Molekülstruktur betrachtet (*s1ipso*-, *s2ipso*- und *s2ortho*-Atome). Sie hängt zunächst von der Information der Summenformel ab, welche als das Vorhandensein oder Nichtvorhandensein der betrachteten chemischen Elemente repräsentiert wird.

Aufgrund der eher groben Strukturrepräsentation, welche die Umgebung des untersuchten Benzolringatoms nur bis zu einer Entfernung von zwei Bindungen in die Betrachtung einbezieht (während andere, jedoch auf dem klassischen zweischrittigen Ansatz basierende Systeme fünf Bindungen entfernte Atome noch einbeziehen) war es nicht verwunderlich, daß die erste Evaluation des initialen Modells noch keine herausragenden Ergebnisse erzielte. Bei der Untersuchung alternativer Entscheidungen in den einzelnen Modellierungsschritten konnten jedoch in großem Umfang Verbesserungen erzielt werden. Die grundsätzliche Eignung des wissensbasierten, am menschlichen Vorgehen orientierten Ansatzes für ein Musteranalysesystem zur Strukturaufklärung organischer Verbindungen sowie von Bayes-Netzen im speziellen zur Realisierung dieses Ansatzes kann somit als bestätigt angesehen werden.

Die beste Erkennungsleistung wurde bei der Untersuchung ausschließlich monosubstituierter Benzolderivate erzielt. Bei dieser Teilmenge der Gesamtstichprobe stimmte die Modellierungsgenauigkeit am besten mit den tatsächlichen Gegebenheiten der im Molekül wirkenden Einflüsse überein. Außerdem wurde neben der Gesamtstichprobe eine Auswahl von Verbindungen untersucht, die ausschließlich „kurze“ Substituenten mit einer maximalen Länge von drei Bindungen enthielt.

Es wurden zudem nicht nur Varianten des Modells, sondern auch seiner Initialisierung einander gegenübergestellt. Die Strukturvariablen konnten empirisch oder gleichverteilt und die Zuordnung von Strukturmerkmalen zu Inkrementbeträgen konnte hart oder weich vorgenommen werden. Außerdem konnte die Summenbildung der einzelnen Beiträge exakt oder mit einem gewissen Toleranzintervall durchgeführt werden. Hinsichtlich der Initialisierungsvarianten konnte festgestellt werden, daß durch eine Aufweichung (weiche bzw. tolerante Initialisierungen) vielfach eine Verbesserung der Erkennung erzielt werden konnte. Wie angestrebt führte also die so eingebrachte Toleranz dazu, daß auch dann auf die richtigen strukturellen Eigenschaften geschlossen werden konnte, wenn die spektroskopischen Befunde nicht ganz genau mit den damit assoziierten Inkrementen übereinstimmten. Dennoch bleibt dies bei zukünftigen Weiterentwicklungen im Einzelfall zu überprüfen, da bei einer zu ungenauen oder ungeeigneten Wissensrepräsentation auch der gegenteilige Effekt eintreten kann.

Außerdem ist festzuhalten, daß bei den ersten Modellvarianten stets die empirische Initialisierung des Zusammenhangs zwischen Molekülstruktur und Summenformel der gleichverteilten überlegen war. Dies bedeutet, daß das allein in der Netzstruktur repräsentierte Wissen unvollkommen ist, das heißt daß innerhalb der Zusammenhänge zwischen Spektrum und Struktur Faktoren wirken, die von dem Modell zunächst nicht erfaßt werden, sondern nur im Rahmen der empirischen Initialisierung in Gestalt statistischer Zusammenhänge zutagetre-

ten. Folglich waren auch die Auswirkungen von Änderungen des Modells am deutlichsten bei den Evaluationsergebnissen gleichverteilt initialisierter Netze zu beobachten.

Die variierten Modellierungsentscheidungen betrafen vorwiegend zwei Aspekte: die Repräsentation des Zusammenhangs zwischen Summenformel und Molekülstruktur, das Zusammenwirken der Beiträge der Strukturmerkmale zur beobachteten chemischen Verschiebung. Außerdem wurde die Repräsentation der *ortho*-Positionen variiert, indem die Substituenten nach ihren Einflüssen auf die chemische Verschiebung kategorisiert betrachtet wurden. Dabei zeigte sich eine bessere Erkennung der Kategorien gegenüber den einzelnen Substituentenkombinationen, jedoch wirkte sich die durch die Kategorisierung eingebrachte zusätzliche Unschärfe negativ auf die Erkennung des *ipso*-Substituenten aus.

Betreffend die Summenformelrepräsentation wurde zum einen eine Umkehr des Kausalzusammenhangs zur Molekülstruktur untersucht (also die Summenformel nun als Konsequenz der strukturellen Gegebenheiten betrachtet) und zum anderen der Zusammenhang als ungerichtete Abhängigkeit repräsentiert, die der Tatsache Rechnung trägt, daß eine Übereinstimmung zwischen Struktur und Summenformel gegeben sein muß, ohne jedoch das eine als Ursache oder Konsequenz des anderen zu betrachten. Beide Realisierungen ermöglichten außerdem die Unterscheidung der Anzahl der Atome jedes beteiligten chemischen Elements und nicht nur die Feststellung von dessen An- oder Abwesenheit. Insbesondere diese zusätzliche Information führte zu einer deutlichen Verbesserung, darüber hinaus konnte eine weitere Steigerung der Zuverlässigkeit durch die Repräsentation als ungerichtete Abhängigkeit erreicht werden. Für diese Modellvariante wurde das beste Resultat für die Klassifikation der *ipso*-Position erzielt: Die Korrektklassifikationsrate betrug 96,64% bei empirischer Initialisierung, weicher Strukturmerkmal-Inkrement-Zuordnung und Zusammenführung der einzelnen Einflüsse auf die chemische Verschiebung ohne Toleranzinterval.

Das beste Resultat für die Klassifikation der Kombination der *ortho*-Positionen betrug 77,92% und wurde im Zuge der Untersuchung im Bereich des Zusammenspiels der Inkremente und Teilinkremente, wiederum für monosubstituierte Verbindungen, erreicht. Die betreffende Modellvariante verzichtet zusätzlich zu obigem auf die Untergliederung des Beitrags der Atome der zweiten Sphäre in einen *ipso*- und einen *ortho*-Anteil. Die Initialisierung des Zusammenhangs zwischen Summenformel und Molekülstruktur wurde ebenfalls empirisch vorgenommen und die einzelnen Inkremente ohne Toleranzintervall zusammengeführt, jedoch eine harte Strukturmerkmal-Inkrement-Zuordnung gewählt. Für den *ipso*-Klassifikationsschritt wurde zugleich mit einer Korrektklassifikationsrate von 96,23% ein Resultat erzielt, das dem oben erwähnten Bestwert nahekommt.

Diese Modellvariante führte auch bei den gleichverteilten Initialisierungsvarianten zu einer weiteren Verbesserung, sie erreichten abhängig von der untersuchten Stichprobe nun sogar dieselben Bereiche wie bei empirischer Initialisierung. Dies ist einerseits ein klarer Hinweis darauf, daß das Zusammenspiel der Einflüsse in der Molekülstruktur noch nicht adäquat wiedergegeben ist und hier weiterer Entwicklungsbedarf besteht. Andererseits führt der mit der beschriebenen Modellvariante beschrittene Weg in die falsche Richtung, da statt einer Verfeinerung der Modellierungsgenauigkeit statistischen Zusammenhängen ein stärkeres Gewicht verliehen wird. Dies entspricht nicht der ursprünglichen Idee eines an Fachwissen orientierten Ansatzes.

Um nun das explizit repräsentierte Wissen demgegenüber wieder stärker zu betonen, wurde eine andere Alternative versucht: Anstatt ein Teilinkrement der höchsten repräsentierten Sphäre empirisch zu gewinnen und diesem somit implizit die Einflüsse aller weiteren, nicht repräsentierten Strukturanteile zuzuschlagen, wurden empirische und aus Literaturwissen gewonnene Anteile von einander getrennt. Beide Teilinkremente wurden nun anhand von

Literaturwissen bestimmt, und es wurde eine neue Variable für alle sonstigen Einflüsse eingeführt, die vom Substitutionsmuster (also *ipso*-Position und *ortho*-Positionen) abhängt und deren zugehörige bedingte Wahrscheinlichkeit empirisch ermittelt wird.

Auf diese Weise konnte ebenfalls erreicht werden, daß sich empirische und gleichverteilte Initialisierung in ihren Fehlerraten nur noch wenig unterschieden, gleichwohl blieben die Resultate hinter der vorgenannten Modellvariante zurück. Dies ist jedoch verständlich, wenn man bedenkt, daß diese verstärkt auf statistische Zusammenhänge setzte in denen sich nicht explizit repräsentierte Faktoren widerspiegeln konnten. Bei der letzten beschriebenen Alternative dagegen wäre eine Erhöhung der Modellierungsgenauigkeit nötig, um bessere Resultate zu erzielen.

Derartige Verfeinerungen des Kausalmodells, vor allem durch Berücksichtigung weiter als zwei Sphären entfernter Atome und insbesondere der Substituentenklassen der *meta*-Positionen und der *para*-Position, sind sicherlich eine der wichtigsten weiterführenden Arbeiten, die nun vorzunehmen sind. Durch die bis hierher durchgeführten Untersuchungen wurde mit der zuletzt beschriebenen Modellvariante ein Stadium erreicht, das die Integration entsprechender zusätzlicher Variablen besonders unproblematisch erlaubt. Zudem steht mit den im Zuge der Evaluierung gemachten Beobachtungen ein gewisser Erfahrungsschatz zur Verfügung, der das weitere Vorgehen erleichtert.

Damit wurden die Ziele der vorliegenden Arbeit erreicht: Die grundsätzliche Eignung des am menschlichen Vorgehen orientierten und auf die Nutzung explizit repräsentierten Fachwissens konzentrierten Ansatzes sowie von Bayes-Netzen als Methode zu seiner Umsetzung wurden gezeigt. Es wurde ein prinzipiell einsatzfähiges Gesamtsystem als Rahmen der Untersuchungen und Grundlage weiterer Entwicklungen geschaffen. Und es steht im Rahmen desselben ein Basismodell der entsprechenden Kausalzusammenhänge zur Verfügung, auf Grundlage dessen durch die Untersuchung und Bewertung einzelner Modellierungsentscheidungen Erfahrungen gesammelt wurden, die nun für seine Weiterentwicklung genutzt werden können.

Glossar

α-Kohlenstoff 29
Bei der systematischen Betrachtung von *Derivaten* einer Verbindung dasjenige Kohlenstoffatom, das den (systematisch variierten) *Substituenten* trägt.

γ-Effekt 29
Tatsache, daß durch eine sterisch induzierte Polarisierung innerhalb des Moleküls der γ-Kohlenstoff in substituierten (vgl. *Substituent*) *Alkanen* anders als der β- und α-*Kohlenstoff* eine Verringerung seiner *chemischen Verschiebung* aufweist.

π-Orbital 11
π-*Molekülorbital*; π-Bindung; Anteil einer *Doppel-* oder *Dreifachbindung*, der durch Linearkombination der hantelförmigen *p-Orbitale* der Bindungspartner entsteht.

σ-Orbital 11
σ-*Molekülorbital*; σ-Bindung; entsteht durch Linearkombination von *s-Orbitalen* oder/und *s*, *p*-Hybridorbitalen (vgl. *Hybridisierung*). *Einfachbindung*, jedoch enthalten auch *Doppel-* und *Dreifachbindungen* einen σ-Anteil.

***p*-Orbitale** 10
Atomorbitale mit hantelförmiger Gestalt, die energetisch über den *s-Orbitalen* derselben Hauptquantenzahl liegen. Die nach ihrer räumlichen Orientierung unterschiedenen *Orbitale* p_x, p_y und p_z sind energetisch äquivalent.

***s*-Orbitale** 10
Atomorbitale mit kugelförmiger Gestalt; energetisch niedrigste *Orbitale* innerhalb jeder Hauptquantenzahl.

a-priori-Wahrscheinlichkeiten 60
Im Kontext von *Bayes-Netzen* Wahrscheinlichkeitsverteilungen, die für diejenigen Variablen benötigt werden, die keine Eltern haben, das heißt die Ereignisse repräsentieren, deren Ursachen nicht bekannt sind.

abgeschirmt 25
Im Kontext der *NMR-Spektroskopie* eine Eigenschaft von Atomkernen mit einer erhöhten Elektronendichte in ihrer Umgebung; sie besitzen ein *hohes Abschirmfeld* (*hochfeldverschobene* Absorptionen im *Spektrum*).

Adding-One 123
Im Bereich statistischer Methoden ein Verfahren zur Vermeidung von Nullwerten bei der Bestimmung *relativer Häufigkeiten*. Es wird stets ein Sockelbetrag von 1 auf die Anzahl der Beobachtungen in der Stichprobe aufgeschlagen, da das Nichtbeobachten eines Ereignisses nicht gleichbedeutend mit dessen genereller Unmöglichkeit ist.

AFFN 100
ASCII Free Format Numeric; ein Zahlenformat, das innerhalb von *JCAMP*-Dateien verwendet wird. Beinhaltet neben Ziffern nur die Zeichen `+`, `-`, den Dezimalpunkt und `E` für die Exponentenschreibweise. `?` steht für unbekannte oder außerhalb des Meßbereichs liegende Werte.

Aliphaten 17
Verzweigte oder unverzweigte kettenförmige organische Verbindungen.

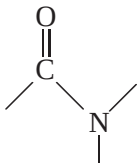
Alkane 16
Organische Verbindungen ohne *Doppel-* oder *Dreifachbindungen* und *Heteroatome*, also *gesättigte reine Kohlenwasserstoffe*.

Alkene 26
Reine Kohlenwasserstoffe mit *Doppelbindungen*.

Alkine 26
Reine Kohlenwasserstoffe mit *Dreifachbindungen*.

Alkohole 15
Verbindungen, die OH-Gruppen enthalten.

Alkylreste 16
Substituenten, die *Alkanen* gleichen.

Amide 15
Verbindungen, die die Gruppe  enthalten.

Amine 16
Organische Verbindungen, welche Stickstoff als (verzweigendes oder nicht verzweigendes) Skelettatom enthalten.

Aromaten 13
aromatische Verbindungen; Verbindungen, die durch ein *aromatisches System* charakterisiert sind.

aromatisches System 13
Delokalisiertes π-Elektronensystem; Konstellation, deren besonderer Elektronenzustand die Stoffgrupe der *Aromaten* charakterisiert.

Atomorbitale 11
Modell der mit einem Atom assoziierten *Elektronen*. Durch ihre Linearkombination entstehen *Molekülorbitale*, welche *Elektronenpaarbindungen* beschreiben.

Ausprägungen 59
Im Kontext kausaler Modellierung in *Bayes-Netzen* Varianten eines Ereignisses, in welchen dieses eintreten kann; z.B. das Ereignis „Wetter“ und die Ausprägungen „sonnig“, „bewölkt“, „regnerisch“.

Azoverbindungen 16
Organische Verbindungen, die die Teilstruktur –N=N– enthalten.

Bayes-Klassifikator 51
Ansatz der *Musterklassifikation*, der die Zuordnung basierend auf der *Rückschlußwahrscheinlichkeit* der einzelnen Klassen vornimmt. Ähnelt darin dem *Risikominimierungsansatz*, Fehlklassifikationen werden jedoch unterschiedlich bewertet.

Bayes-Netz 58
Graphenmodell mit dem Schwerpunkt der Betrachtung von Kausalbeziehungen zwischen Ereignissen. Diese sind über bedingte Wahrscheinlichkeiten quantifiziert. Erlaubt sowohl den kausalen Vorwärtsschluß als auch den diagnostischen Rückschluß von der Beobachtung auf die Ursache (vgl. *diagnostische/kausale Evidenzen*).

Bedingt unabhängig 61
Eigenschaft zweier Ereignisse *A* und *C*, wenn gegeben ein drittes Ereignis *B* keinerlei Kenntnisse über *C* die Wahrscheinlichkeit von *A* beeinflussen und umgekehrt; siehe Gleichung (4.13)

Belief Updating 75
Neuberechnung der Wahrscheinlichkeiten aller Aussagen innerhalb eines *Bayes-Netzes* bei gegebenen Beobachtungen.

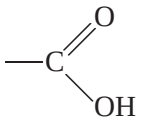
Bezeichner 99
In einer *JCAMP*-Datei der erste, von ## und = eingeschlossene Teil eines *LDR*s, der die Art der enthaltenen Information kennzeichnet. Der JCAMP-Standard gibt diverse Bezeichner vor; durch Voranstellen von $ können zusätzliche definiert werden.

Block 100
Zusammengehörige Information innerhalb einer *JCAMP*-Datei. Man unterscheidet *Link-Blocks* und *Datenblocks*. `##TITLE=` und `##END=` kennzeichnen jeweils Anfang und Ende eines Blocks, wobei eine Schachtelung möglich ist.

Blocks 77
Informationseinheiten innerhalb einer Datei im *BNIF-Format*. Schlüsselwörter (`network`, `variable` und `probability`) kennzeichnen den Bezug der Information eines besimmten Blocks.

BNIF-Format 76
Bayesian Network Interchange Format; Vorschlag zur Standardisierung der Repräsentation von *Bayes-Netzen*. Trotz der Entwicklung hin zu einem XML-basierten Ansatz werden die BNIF-Konventionen in dieser Arbeit herangezogen, da sie in der genutzten Implementierung des *Bucket-Elimination* Algorithmus benutzt werden.

Bucket-Elimination 75
Algorithmischer Rahmen, der Prinzipien des *Dynamic Programming* verallgemeinert. Grundidee ist die Gruppierung von Termen nach dem höchsten vorkommenden Variablenindex in sogenannte Buckets („Behälter“) und ihre anschließende absteigende Eliminierung, je nach Anwendung durch Einsetzen oder Summation. Für *Bayes-Netze* einsetzbar zum *Belief Updating* sowie zur Berechnung von *mpe* oder *map*.

Carbonsäuren 15
Organische Verbindungen, die die *funktionelle Gruppe*  enthalten.

chemische Äquivalenz 26
Eigenschaft zweier Atomkerne, die sich in gleicher chemischer Umgebung befinden. Sie weisen dieselbe *chemische Verschiebung* auf.

chemische Bindung 5
Feste Verknüpfung zweier Atome.

chemische Verschiebung 24
Abhängigkeit der Larmorfrequenz eines Kernspins von dessen chemischer Umgebung (Elektronendichte); Lage des *Peaks* auf der *x*-Achse im *NMR-Spektrum*.

Chromophore 22
In der *Spektroskopie* strukturelle Untereinheiten eines Moleküls, die bei der spektroskopischen Untersuchung Absorptionen hervorrufen.

Compound JCAMP 100
JCAMP-Datei, die mehrere *Blocks* enthält, welche archivartig aufgelistet oder mit Hilfe von *Link-Blocks* hierarchisch untergliedert sein können.

Cyanate 15
Organische Verbindungen, welche die Gruppe –O–C≡N enthalten.

Cyanide 16
Organische Verbindungen, welche die Gruppe –C≡N enthalten.

Cycloverbindungen 17
Ringförmige Verbindungen.

d-separiert 62
Eigenschaft zweier Ereignisse *A* und *C* in einem *Bayes-Netz*, die gegeben die eingetragenen Evidenzen *bedingt unabhängig* sind. Siehe Abbildung 4.6.

Datenblock 100
In *JCAMP*-Dateien ein solcher *Block*, der die gewünschten Informationen (im Gegensatz zu organisatorischen Angaben in *Link-Blocks*) enthält.

delokalisiertes π-Elektronensystem 13
System alternierender *Doppel-* und *Einfachbindungen*, in welchem die unhybridisierten *p-Orbitale* (vgl. auch *Hybridisierung*) der beteiligten Atome eine solche Orientierung haben, daß ihre Kombination zu π-*Molekülorbitalen* in beiden Richtungen entlang des Kohlenstoffskeletts stattfindet, so daß die Doppelbindungen nicht eindeutig zu lokalisieren sind. Siehe Abbildung 2.7.

Delokalisierungsenergie 18
Energiedifferenz zwischen dem Elektronenzustand abwechselnder *Doppel-* und *Einfachbindungen* und einem *aromatischen System*; das heißt der Betrag, um welchen der aromatische Elektronenzustand günstiger ist als die *mesomeren Grenzformeln*.

Derivate 5
Abkömmlinge einer organischen Verbindung, die durch geringfügige strukturelle Modifikation aus der ursprünglichen Verbindung hervorgehen.

diagnostische Evidenzen 59
Im Kontext von *Bayes-Netzen* solche Beobachtungen, bei deren Auswertung entgegen der Kausalrichtung (von der beobachtbaren Wirkung auf die mögliche Ursache) geschlossen wird.

disjunkte Ereignisse 61
Im mathematischen Sinne Ereignisse, welche kein einziges gemeinsames *Elementarereignis* enthalten.

disubstituiert 20
Eigenschaft eines *Derivats*, bei welchem zwei der möglichen Positionen der Ausgangsverbindung substituiert sind (vgl. auch *Substituenten*).

Divorcing 67
Konzept zur Anpassung der Modellierung kausaler Zusammenhänge.

Doppelbindung 13
Zwei Bindungen zwischen denselben beiden Atomen, bestehend aus einem σ- und einem π-Anteil; vgl. σ-*Orbital*, π-*Orbital*.

Dreifachbindung 13
Drei Bindungen zwischen denselben beiden Atomen, bestehend aus einem σ- und zwei π-Anteilen; vgl. σ-*Orbital*, π-*Orbital*.

Dynamic Programming 75
Bereich der Informatik, der sich der Laufzeitoptimierung von Algorithmen widmet. Das dabei genutzte Grundprinzip ist die Rückführung komplexerer Probleme auf einfachere: Die Lösung des eigentlichen Problems kann ermittelt werden, indem die optimalen Lösungen der Subprobleme ausgenutzt werden.

Eckengradfolge 35
Geordnete Liste aller Eckengrade eines Graphen; der Eckengrad eines Knotens gibt die Zahl mit ihm verbundener Kanten an, wobei Verbindungen mit sich selbst (Schleifen) doppelt gezählt werden.

Edelgaskonfiguration 6
Elektronenzustand einer voll besetzten äußersten Schale, der dem der Edelgase entspricht. Energetisch günstiger Zustand, den Atome durch die Ausbildung von Bindungen anstreben.

Einfachbindungen 12
Einzelne Bindung zwischen zwei Atomen; σ-Bindung; vgl. σ-*Orbital*.

Eintrag 99
Im Zusammenhang des *JCAMP-Formats* der Bestandteil eines *LDR*s, der die unmittelbaren Daten enthält. Deren Format (etwa *AFFN*, *String* oder frei formatierter *Text*) wird durch den *Bezeichner* festgelegt, der die Art des LDRs charakterisiert.

Elektron 5
Negativ geladenes subatomares Teilchen.

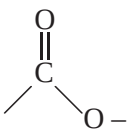
Elektronegativität 6
Einheitenlose Naturkonstante, die die Neigung eines chemischen Elements bezeichnet, *Valenzelektronen* an sich zu ziehen. Sie ist im allgemeinen um so größer, je mehr Valenzelektronen das chemische Element besitzt.

Elektronenhülle 28
Gesamtsystem der Elektronen eines Moleküls, dessen Gestalt (Dichte) vorrangig von den in Bindungen involvierten *Valenzelektronen* charakterisiert wird.

Elektronenpaarbindung 6
Bindung zwischen Elementen, deren *Elektronegativitäten* sich nicht oder nur mäßig stark unterscheiden, so daß sich die Bindungspartner die *Valenzelektronen* teilen.

Elementarereignis 60
Im mathematischen Sinne ein Ereignis, das nicht als Kombination anderer Ereignisse aufgefaßt werden kann, z.B. das „Würfeln einer 1“ im Gegensatz zum „Würfeln einer ungeraden Zahl“.

entschirmt 25
Im Kontext der *NMR-Spektroskopie* eine Eigenschaft von Atomkernen mit verringerter Elektronendichte in ihrer Umgebung; diese besitzen ein *tiefes Abschirmfeld* (*tieffeldverschobene* Absorptionen im *Spektrum*).

Ester 15
Verbindungen, die die Gruppe  enthalten. *Derivate* von *Carbonsäuren*, in denen das Wasserstoffatom der Säuregruppe (–COOH) durch einen organischen Rest ersetzt ist.

Ether 15
Verbindungen, in denen ein Sauerstoffatom innerhalb des Kohlenstoffskeletts eingebaut ist.

Expertensysteme 58
Wissensbasierte Systeme; Systeme, die mit Hilfe einer Wissensbasis von Produktionsregeln und einem Inferenzsystem zum Anstellen von Schlußfolgerungen das Vorgehen menschlicher Experten nachbilden sollen. Der ehrgeizige Ansatz der 1960er Jahre, menschliche Experten durch präzisere Maschinen zu ersetzen, erreichte in dieser Form jedoch im *unsicheren Schließen* seine Grenzen.

Färbung 35
In der Graphentheorie Zuordnung einer natürlichen Zahl zu jedem Knoten des Graphen; repräsentiert der Graph eine Molekülstruktur, so werden dadurch die chemischen Elemente der einzelnen Atome codiert.

False Positives 43
Bei der Evaluierung eines *Mustererkennung*ssystems Bezeichnung für solche Testbeispiele, die fälschlich eine positive Bewertung erhalten; bezüglich Strukturaufklärungssystemen sind dies solche Strukturen, die eine bessere Bewertung erhielten als die tatsächlich korrekte Molekülstruktur.

Fermionen 23
Atomkerne mit ungerader *Massenzahl*, z.B. ^{1}H und ^{13}C, für welche die *Spinquantenzahl I* ein Vielfaches von $+\frac{1}{2}$ ist.

funktionelle Gruppen 14
Teile eines Moleküls, die mit einer bestimmten chemischen Funktionalität assoziiert sind, welche oftmals vom Vorhandensein von *Heteroatomen* ausgeht, da diese aufgrund ihrer von Kohlenstoff und Wasserstoff abweichenden *Elektronegativität* wie Markierungen im Molekül wirken.

Gemischtverteilungsklassifikator 51
Ansatz im Bereich der *Musterklassifikation*, welcher den mustererzeugenden Prozeß über eine gewichtete Summe von Normalverteilungen modelliert.

Genetische Algorithmen 40
Von der biologischen Evolution inspiriterte Methodengattung in der Informatik im Bereich der Optimierung. Die interessante Information (z.B. Molekülstrukturen) wird in „Genen“ codiert und diese mit Hilfe einer Fitness-Funktion bewertet. Je nach dieser Bewertung geht ein Gen durch Klonierung, Rekombination oder Mutation in die nächste Generation ein.

Gepaarte Elektronen 10
Zwei Elektronen gegensätzlichen Spins, die gemeinsam ein und dasselbe *Orbital* besetzen. Dieser Zustand ist aufgrund der Kompensationseffekte der Einzelspins besonders günstig zu bewerten.

Gesättigte Verbindungen 16
Organische Verbindungen, die ausschließlich *Einfachbindungen* enthalten.

Halogenierte Verbindungen 15
Organische Verbindungen, in denen Halogene (Fluor, Chlor, Brom, Iod) ein oder mehrere Wasserstoffatome ersetzen.

Heteroaromaten 18
Aromatische Verbindungen, in welchen *Heteroatome* am *aromatischen System* beteiligt sind.

Heteroatom 5
Fremdatom eines anderen chemischen Elements als Kohlen- oder Wasserstoff, das in einer organischen Verbindung vorkommt.

Heterosubstituierte Kohlenwasserstoffe 15
Organische Verbindungen, in welchen ein oder mehrere Positionen durch *Heteroatome* ersetzt sind (lat. *substituere*: „ersetzen“).

Hidden Markov Modelle 52
Methode der *Mustererkennung*, die besonders für die Betrachtung von Folgen von Mustern (etwa bei der Sprach- oder Handschrifterkennung) geeignet ist, da sie Segmentierung und Klassifikation in einem Schritt integriert.

hochfeldverschobenen 26
vgl. *abgeschirmt*

hohes Abschirmfeld 25
vgl. *abgeschirmt*

HOSE-Code 41
engl. *hierarchically ordered system of spherical environments*: „hierarchisch geordnetes System sphärischer Umgebungen“; wurde entwickelt, um eine eindeutige, systematische und kompakte Beschreibung von Molekülstrukturen zu ermöglichen. Die betrachtete Struktur wird dabei in sogenannte *Sphären* untergliedert (siehe dort).

Hybridisierung 11
Intraatomare Linearkombination des 2*s*- und eines, zweier oder aller drei 2*p*-*Orbitale* des Kohlenstoffs. Je nachdem entstehen zwei *sp*-Hybride, drei sp_2-Hybride oder vier sp_3-Hybride, welche jeweils energetisch äquivalent sind.

Imine 16
Organische Verbindungen, welche die Gruppe  enthalten.

induktive Effekte 14
Tatsache, daß *Heteroatome* durch ihre von Kohlenstoff und Wasserstoff deutlich verschiedene *Elektronegativität* Bindungen polarisieren und so eine Veränderung der Elektronendichte verursachen. Die in der klassischen organischen Chemie am häufigsten vorkommenden Heteroatome haben eine höhere Elektronegativität als Kohlenstoff, so daß das Heteroatom negativ und Kohlenstoff positiv polarisiert wird (*negativer induktiver Effekt = −I-Effekt*).

Inkrementmethoden 37
Methodengattung zur Vorhersage von *NMR-Spektren*. Ausgehend von der Annahme, daß die Einflüsse einzelner Strukturfragmente auf die Verschiebung eines bestimmten *Chromophors* additiv sind, kommen Tabellen zum Einsatz, welche die Einflüsse verschiedener Gruppen in einer bestimmten relativen Position auf bestimmte Grundtypen von Kernen enthalten.

Ionen 6
Positiv (Kationen) oder negativ (Anionen) geladene Teilchen (Atome oder Moleküle).

Ionenbindung 6
Bindung zwischen Elementen, deren *Elektronegativitäten* sich so stark unterscheiden, daß der elektronegativere Partner die *Valenzelektronen* vollständig an sich zieht, so daß beide Partner als *Ionen* vorliegen.

ipso 20
Bezeichnung für diejenige Position in *monosubstituierten* Benzol*derivaten*, die den *Substituenten* trägt. Darüber hinaus kann derselbe Begriff zur Bezeichnung des Fokuspunktes bei der Beschreibung von *Substitutionsmustern* verwendet werden.

Isomere 7
Verbindungen mit gleicher *Summenformel*, die eine unterschiedliche Struktur haben

Isotop 30
Variante eines chemischen Elements mit einer abweichenden Anzahl von *Neutronen* im Kern.

JCAMP-Format 74
Joint Committee on Atomic and Molecular Physical Data; die JCAMP-Datenformate JCAMP-DX (Spektraldaten) und JCAMP-CS (Strukturdaten) dienen dazu, es den Anwendern spektroskopischer Systeme zu ermöglichen, ihre Daten zwischen unterschiedlichen Systemen zu transferieren.

k-fach-Kreuzvalidierungsverfahren 144
Verfahren zur Evaluation eines Klassifikators. Die Stichprobe wird in k Teilmengen unterteilt, und es wird k-mal evaluiert, wobei jeweils eine andere Teilmenge als Testmenge dient. Die übrigen werden zur Initialisierung verwendet. Besonders vorteilhaft bei knapper Datenlage, da die Durchschnittsbildung den Einfluß einer besonders vorteilhaften oder besonders ungünstigen Trainings- oder Testmenge relativiert.

kausale Evidenzen 59
Im Kontext von *Bayes-Netzen* solche Beobachtungen, die bei der Auswertung mit der Ursache-Wirkungs-Richtung propagiert werden.

Kausalnetz 58
Qualitative Beschreibung der Kausalzusammenhänge des Problemfeldes.

Kern 101
In *JCAMP*-Dateien diejenigen *LDR*s innerhalb eines *Blocks*, die gemäß dem Standard obligatorisch sind und die die sinnvollerweise erforderliche Information enthalten.

klassenspezifische Dichte 51
Im Kontext der *Musterklassifikation* die Wahrscheinlichkeitsdichte der *Merkmalsvektoren* einer gegebenen die Klasse.

Kohlenwasserstoffe 5
Substanzen, die hauptsächlich aus den Elementen Kohlenstoff (C) und Wasserstoff (H) aufgebaut sind.

Konfigurationsisomere 8
Isomere, die sich nicht in Lage und Art der Bindungen innerhalb des Moleküls, jedoch in der relativen Stellung von Atomen oder Gruppen zueinander unterscheiden.

Konformationsisomere 8
Kategorisierung wesentlich verschiedener energetisch unterschiedlicher Ausprägungen der relativen Anordnung von Teilen des Gesamtmoleküls. Aufgrund der Drehbarkeit von Bindungen sind die unterschiedlichen Konformationsisomere jedoch in einander überführbar.

konjugierte Doppelbindungen 13
Alternierende Anordnung von *Doppel-* und *Einfachbindungen*.

Konstitutionsisomere 8
Isomere, die sich in Lage und Art der Bindungen unterscheiden.

Kopplungen 29
Wechselseitige Beeinflussung der in der *NMR-Spektroskopie* betrachteten Spinzustände des untersuchten Kerns und der Kerne in seiner Umgebung. Kommt zustande, wenn sie, wie im Falle von ^{1}H und ^{13}C, dieselbe *Spinquantenzahl* haben, und führt zu einer Aufspaltung der Absorptionen in Liniengruppen (*Multipletts*).

kovalente Bindung 5
Bindung zwischen chemischen Elementen mit annähernd gleicher *Elektronegativität*, bei der sich die Bindungspartner ein Bindungselektronenpaar gleichberechtigt teilen.

Ladungszahl 23
Zahl der *Protonen* eines Atomkerns.

LDR 99
engl. *labeled data record*: „etikettiertes Datenfeld“; Informationseinheit einer *JCAMP*-Datei. Besteht aus einem *Bezeichner* und dessen zugehörigen *Eintrag*.

Link-Block 100
Art von *Blocks* innerhalb einer *JCAMP*-Datei, welche die informationsenthaltenden *Datenblocks* zusammenfaßt, gruppiert und organisiert.

Makroatom 35
Im Kontext von Strukturgeneratoren ein Strukturfragment, das wie ein einzelnes Atom behandelt wird.

map 75
maximum a-posteriori hypothesis; wahrscheinlichste Zustandskombination der Hypothesenvariablen eines *Bayes-Netzes* bei gegebenen Beobachtungen.

Massenzahl 23
Zahl der *Protonen* und *Neutronen* eines Atomkerns zusammen.

Maximum-Likelihood-Klassifikator 51
Ansatz in der *Musterklassifikation*, der die Klassenzuordnung basierend auf der *klassenspezifischen Dichte* vornimmt, das heißt es wird diejenige Klasse gewählt, für die die Wahrscheinlichkeit des gegebenen *Merkmalsvektors* maximal ist.

Merkmalsextraktion 50
In der *Mustererkennung* die Abbildung eines in Gestalt seiner Meßdaten repräsentierten Objekts auf einen *Merkmalsvektor*.

Merkmalsvektor 50
In der *Musterklassifikation* Beschreibung des zu klassifizierenden Objekts. Seine einzelnen Komponenten heißen Merkmale; ihre Zahl entscheidet über den mathematischen Aufwand bei der Klassifiaktion.

mesomere Effekte 14
Interferenz freier Elektronenpaare in der Nachbarschaft *aromatischer Systeme* mit diesen, wodurch sie in die *mesomeren Grenzformeln* der Verbindung eingehen. Dadurch verändert sich die Elektronendichte des Systems, und Partialladungen an einzelnen Positionen können entstehen. Mit Hilfe der *Valenzstrichschreibweise* ist dies durch „Umklappen“ von Elektronenpaaren zu veranschaulichen.

mesomere Grenzformeln 9
Strukturformeln, welche durch Verschieben von Elektronenpaaren in einander überführt werden können und welche wiedergeben, zwischen welchen (nicht real existierenden) Extremen sich der tatsächliche Elektronenzustand einer *mesomeriestabilisierten* Verbindung befindet.

Mesomerie 9
Phänomen, daß die Bindungsverhältnisse eines Moleküls nicht eindeutig angegeben, sondern nur durch mehrere *mesomere Grenzformeln* umschrieben werden können. Die tatsächliche Elektronenverteilung liegt *zwischen* diesen Grenzformeln und wird durch keine einzelne von ihnen komplett erfaßt (griech. *meso-*: „Mitte“).

Mesomeriestabilisierung 9
Tatsache, daß der wahre Elektronenzustand einer Verbindung energetisch günstiger ist, als es jede durch ihre *mesomere Grenzformeln* beschriebene Struktur wäre.

meta 20
Eigenschaft von durch eine unsubstituierte Position von einander getrennten *Substituenten* in disubstituierten Benzol*derivaten* („*meta*-ständig“). Darüber hinaus kann die Bezeichnung analog für die übernächste Positionen vom Fokuspunkt aus bei der Beschreibung des *Substitutionsmusters* verwendet werden.

Molekülorbitale 11
Beschreibung von *Elektronenpaarbindungen*; entstehen durch Linearkombination von Atomorbitalen (LCAO, *linear combination of atomic orbitals*) und werden von *Elektronen* besetzt, die durch ihre Beteiligung an der Bindung nicht mit einem einzelnen Atom assoziiert sind.

monosubstituiert 20
Eigenschaft eines *Derivats*, bei dem nur eine von mehreren möglichen Positionen der Ausgangsverbindung substituiert ist (vgl. auch *Substituenten*).

mpe 75
engl. *most probable explanation*: „wahrscheinlichste Erklärung“; im Zusammenhang mit *Bayes-Netzen* die wahrscheinlichste Belegung einiger Variablen bei gegebenen Beobachtungen bezüglich der übrigen Variablen.

Multigraphen 35
Graphen mit potentiell mehreren Kanten zwischen denselben Knoten.

Multipletts 29
Aufspaltung der NMR-Absorptionen in Liniengruppen durch die *Kopplung* von Kernen, durch die deren Resonanzfrequenzen um einige Hz verschoben werden.

Muster 49
Im allgemeinen Sprachgebrauch eine exemplarische Probe oder ein Schema, wodurch prototypisch Abläufe oder Objekte charakterisiert werden. Im Kontext der *Mustererkennung* die Repräsentation eines betrachteten Objekts (Meßdaten), dessen Interpretation (Kategorie bzw. symbolische Beschreibung) gesucht ist.

Musteranalyse 50
Disziplin der Informatik aus dem Bereich der *Mustererkennung*, deren Ziel es ist, das gegebene *Muster* zu einer symbolischen Beschreibung zu verarbeiten, welche auf der Beziehung der einzelnen Bestandteile des komplexen Musters zueinander basiert.

Mustererkennung 49
Disziplin der Informatik, die versucht, Wahrnehmungsleistungen, die vom Menschen oder allgemein von Lebenwesen bekannt sind, zu automatisieren, indem das „Muster“ hinter einem bestimmten Ablauf, einer Szene, einem Objekt usw. erkannt und es so als Instanz eines Oberbegriffs identifiziert oder schematisch beschrieben wird.

Musterklassifikation 50
Disziplin der Informatik aus dem Bereich der *Mustererkennung*, deren Ziel die Zuordnung eines *Musters* zu einer von endlich vielen Kategorien (Klassen) ist.

Negativliste 36
Im Kontext von Strukturgeneratoren eine Liste bei der Generierung unzulässiger Strukturbausteine; allgemein eine Liste von Elementen, die in einem bestimmten Kontext verboten/unzulässig sind.

neuronale Netze 38
Vom Aufbau des Gehirns inspirierte Methode der Informatik. Erlaubt das Lernen komplexer *Muster* ohne explizite Vorgabe der zugrundeliegenden Regeln. Das *Training* erfolgt durch Präsentation einer großen Menge von Eingabedaten und den jeweils korrekten Ausgabewerten, wobei eine Lernregel die internen Parameter anpaßt.

Neutron 5
Ungeladenes Teilchen innerhalb des Atomkerns.

Nitrosoverbindungen 16
Organische Verbindungen, die eine Nitroso-Gruppe (–N=O) enthalten.

Nitroverbindungen 16
Organische Verbindungen, die eine Nitro-Gruppe ($–NO_2$) enthalten.

NMR-Spektroskopie 21
Kernmagnetresonanzspektroskopie; eine bestimmte spektroskopische Technik (vgl. *Spektroskopie*). Im NMR-*Spektrum* wird die Reaktion der Kerne auf ihre Resonanzfrequenzen sichtbar gemacht, die Rückschlüsse auf die Struktur des Moleküls erlaubt.

Noisy Or 67
Konzept zur Anpassung der Modellierung kausaler Zusammenhänge.

Normalverteilungsklassifikator 51
Ansatz im Bereich der *Musterklassifikation*, welcher den mustererzeugenden Prozeß über eine Normalverteilung (vgl. Gleichung (4.1)) modelliert.

Notes 101
In *JCAMP*-Dateien diejenigen (nichtobligatorischen) *LDR*s innerhalb eines (spektrenbezogenen) JCAMP-DX *Datenblocks*, die dem *Kern* vorangestellt sein können, um die Beschreibung des NMR-Experiments angemessen zu vervollständigen.

Oktettregel 6
Besagt, daß Atome in der Regel mit acht *Elektronen* in der äußersten Schale *Edelgaskonfiguration* erreichen (lat. *okta*-: „acht“).

Olefine 17
Organische Verbindungen, die Mehrfachbindungen enthalten.

Orbital 9
Beschreibung eines *Elektrons* durch seine Aufenthaltswahrscheinlichkeit im Raum; vgl. *Teilchen-Welle-Dualität*.

Orbitalmodell 6
Quantenmechanisches Modell zur Wiedergabe der Elektronen- und Bindungsstruktur innerhalb organischer Moleküle.

organische Chemie 5
Die Chemie der Kohlenstoffverbindungen, das heißt derjenigen chemischen Verbindungen, deren Grundstruktur aus Kohlenstoffatomen aufgebaut ist.

ortho 20
Bezeichnung für benachbart angeordnete *Substituenten* in *disubstituierten* Benzol*derivaten*. Darüber hinaus kann die Bezeichnung für die dem Fokuspunkt benachbarte Position bei der Beschreibung des *Substitutionsmusters* verwendet werden.

para 20
Bezeichnung für gegenüberliegend angeordnete *Substituenten* in *disubstituierten* Benzol*derivaten*. Außerdem kann die Bezeichnung für die dem Fokuspunkt gegenüberliegende Position zur Beschreibung von *Substitutionsmustern* verwendet werden.

Peaks 22
Absorptionsmaxima innerhalb eines *Spektrums*, die durch ihre Lage auf der Energieachse bestimmte Eigenschaften des untersuchten Moleküls charakterisieren.

polare Bindung 6
Elektronenpaarbindung zwischen Elementen, deren *Elektronegativität* sich unterscheidet, jedoch nicht stark genug für eine *Ionenbindung*. Der elektronegativere Bindungspartner wird negativ, der andere positiv polarisiert.

Polynomklassifikator 52
Ansatz der *Musterklassifikation*; Universalapproximator, da nach dem Approximationssatz von WEYERSTRASS jede Funktion durch ein Polynom approximiert werden kann, wenn der Grad des Polynoms ausreichend hoch gewählt wird. Zu approximieren ist im Kontext der Musterklassifikation der mustererzeugende Prozeß.

Positivliste 36
Im Kontext von Strukturgeneratoren eine Liste bei der Generierung erlaubter Strukturbausteine; allgemein eine Liste von Elementen, die in einem bestimmten Kontext erlaubt/zulässig sind.

Promotion 11
Anhebung von einem der beiden 2*s*-Elektronen des Kohlenstoffs in das energetisch höhere noch unbesetzte 2*p-Orbital*, um vier ungepaarte, für die Bindungsbildung verfügbare Elektronen zu erhalten. Die dazu nötige Promotionsenergie wird durch den Energiegewinn bei der Bindungsbildung überkompensiert.

Proton 5
Positiv geladenes Teilchen innerhalb des Atomkerns.

protonenbreitbandentkoppelt 29
Eigenschaft von ^{13}C-NMR-*Spektren*, wenn beim NMR-Experiment ein Frequenzband eingestrahlt wurde, welches den gesamten Bereich der ^{1}H-Anregungen abdeckt (Breitband). Die ^{1}H-Kerne (Protonen) vollziehen dadurch ständig Zustandsübergänge, so daß es nicht zu *Kopplungen* mit den ^{13}C-Kernen kommt.

Quantenzahl 10
Index, der mögliche Zustände von Systemen numeriert, z.B. zur Unterscheidung von *Orbitalen*.

Rückschlußwahrscheinlichkeit 51
Im Kontext der *Musterklassifikation* die Wahrscheinlichkeit einer Klasse gegeben den *Merkmalsvektor*. Grundlage der Klassifikation beim *Bayes-Klassifikator* und beim *Risikominimierungsansatz*.

reine Kohlenwasserstoffe 15
Organische Verbindungen, die keine *Heteroatome* enthalten.

relative Häufigkeiten 111
Quotient $\frac{n}{N}$ der Anzahl n von Beispielen, die eine bestimmte Eigenschaft besitzen, und der Größe N der Stichprobe. Dient zur Approximation von Wahrscheinlichkeiten, vgl. Gleichung (8.1).

Relaxation 31
Allmähliche Wiederherstellung der im Ruhezustand für die Besetzung der Energieniveaus geltenden BOLTZMANN-Verteilung (vgl. Gleichung (2.4)), welche durch die gezielte Anregung der Kernspins im NMR-Experiment aufgehoben wird.

repräsentative, klassifizierte Stichprobe 52
Menge von Beispiel*mustern*, für welche jeweils die zugehörige Klasse bekannt ist, und die in ihrer Zusammensetzung die Charakteristik des betrachteten Anwendungsfeldes gut wiedergibt. Dient dem *Training* in der *Musterklassifikation*.

Risikominimierung 51
Ansatz der *Musterklassifikation*, der die Zuordnung basierend auf der *Rückschlußwahrscheinlichkeit* der einzelnen Klassen vornimmt. Ähnelt darin dem *Bayes-Klassifikator*, Fehlklassifikationen werden jedoch unterschiedlich bewertet.

Satz von BAYES 61
Fundamentaler Satz der Wahrscheinlichkeitsrechnung, Grundlage der Informationspropagierung in *Bayes-Netzen*. Siehe Gleichungen (4.7) und (4.9).

Satz von der Totalen Wahrscheinlichkeit 61
Fundamentaler Satz der Wahrscheinlichkeitsrechnung, siehe Gleichung (4.8).

Schalenmodell 6
RUTHERFORD-BOHR-SOMMERFELD-Atommodell; geht von einem schalenartigen Aufbau aus, in welchem die *Elektronen* der einzelnen Schalen den Atomkern umkreisen wie Planeten eine Sonne.

Schleifen 35
vgl. *Eckengradfolge*.

schwere Atome 34
In einer organischen Verbindung andere Atome als Wasserstoff.

Shell 101
In *JCAMP*-Dateien diejenigen *LDR*s innerhalb eines (strukturbezogenen) JCAMP-CS-*Datenblocks*, die auf den *Kern* folgen können und die nichtobligatorische Zusatzangaben betreffend die Molekülstruktur enthalten.

Simple JCAMP 100
JCAMP-Datei, die nur aus einem einzigen *Datenblock* besteht.

Spektrendatenbanken 39
Werden im Kontext der Strukturaufklärung eingesetzt, um SSC-Bibliotheken (*subspectra-substructure-correlation*) aufzubauen. Sie enthalten interpretierte *Spektren*, das heißt zusätzlich zu den Spektraldaten auch Verknüpfungen zu den verursachenden Strukturen bzw. Chromophoren.

Spektroskopie 21
Untersuchung der quantisierten Wechselwirkung von Strahlungsenergie mit Materie. Durch die Absorption charakteristischer Energiequanten kann dabei auf strukturelle Eigenschaften des untersuchten Moleküls geschlossen werden.

Spektrum 22
Graphische Darstellung der Intensität der der absorbierten Strahlung in Abhängigkeit von der eingestrahlten Energie im spektroskopischen Experiment. Besonders interessant sind die Absorptionsmaxima, die durch ihre Lage auf der Energieachse bestimmte Eigenschaften des untersuchten Moleküls charakterisieren.

Sphären 41
Dienen der systematischen Erfassung der Atome eines Moleküls bei dessen Beschreibung im *HOSE-Code*. Das Atom im Fokus der Betrachtung besetzt die nullte Sphäre. Die erste Sphäre enthält seine direkten Nachbarn, die zweite Sphäre all diejenigen Atome, die zwei Bindungen entfernt liegen, und so weiter (vgl. Abbildung 3.3).

Spinquantenzahl 23
I; bestimmt die Zahl der unterscheidbaren Spinzustände eines Atomkerns als $2I+1$. Zustandsübergänge werden bei den Untersuchungen der *NMR-Spektroskopie* angeregt. Ist $I = \frac{1}{2}$, wie bei ^{1}H und ^{13}C, so ist der vollzogene Übergang in Abhängigkeit vom eingestrahlten Energiequantum eindeutig.

Stereochemie 8
Teilbereich der *organischen Chemie*, der sich mit räumlichen Aspekten, wie der relativen Stellung einzelner Atome innerhalb eines Moleküls, befaßt.

String 100
Datenformat innerhalb von *JCAMP*-Dateien, das alphanumerische Einträge bezeichnet, die für die automatische Verarbeitung vorgesehen sind. Sie müssen also bestimmten Regeln folgen, die dem durch den *Bezeichner* des *LDRs* angegebenen Typ entsprechen.

Strukturaufklärung 1
Das Aufdecken struktureller Eigenschaften organischer Moleküle.

Strukturformel 7
Notation der Molekülstruktur, die über die *Summenformel* hinausgeht. Enthält zusätzlich Informationen über Lage und Art der Bindungen zwischen den einzelnen Atomen.

Strukturgeneratoren 34
Programme, die gegeben eine bestimmte Ausgangsinformation alle passenden Molekülstrukturen liefern. Setzen Methoden aus den Bereichen der Gruppentheorie, Kombinatorik und Graphentheorie vereinigt in einem spezialisierten Computeralgebrasystem sein.

Substituenten 14
Gruppen oder Atome, die in *Derivaten* organischer Verbindungen Wasserstoffatome oder andere Seitengruppen der Ausgangsverbindung ersetzen (lat. *substituere*: „ersetzen“); oft handelt es sich dabei um *funktionelle Gruppen*.

Substituenteninkremente 86
Terme, die den Betrag wiedergeben, um den ein bestimmter *Substituent* die *chemische Verschiebung* seines Bindungspartners beeinflußt. Gelten jeweils mit Bezug auf einen bestimmten Typ von Chromophor als Bindungspartner.

Substitutionsgrad 20
Anzahl von *Substituenten* in *Derivaten* organischer Verbindungen.

Substitutionsmuster 20
Zahl, Art und relative Anordnung von *Substituenten* in *Derivaten* organischer Verbindungen.

Substrukturanalyse 38
Untersuchung einer gegebenen Menge von Strukturen durch paarweisen Vergleich aller Moleküle mit Blick auf das größte gemeinsame Strukturfragment. Anhand von Gemeinsamkeiten können die einzelnen Elemente der Gesamtmenge in Untermengen sinnvoll gruppiert werden.

Summenformel 7
(*molecular formula*); Angabe der an einer Verbindung beteiligten chemischen Elemente und ihrer Multiplizitäten, z.B. H_2SO_4.

Tautomerie 9
Umlagerung innerhalb des Moleküls, die spontan und unter sehr geringem Energieaufwand stattfindet. Die in einer Substanzprobe der betreffenden Verbindung vorliegenden unterschiedlichen *Isomere* (Tautomere) können daher in der Praxis nicht von einander getrennt werden. Die Umlagerung kann in *Strukturformeln* (*Valenzstrichschreibweise*) durch Verschieben von Elektronenpaaren dargestellt werden.

Teilchen-Welle-Dualität 9
Feststellung, daß kleine bewegte Objekte Welleneigenschaften besitzen. Bewegte Massepunkte, z.B. *Elektronen*, können daher als (komplexe) Wellenfunktion Ψ beschrieben werden. Diese entspricht einer Wahrscheinlichkeitsamplitude, deren Quadrat die Aufenthaltswahrscheinlichkeit des durch sie beschriebenen Systems ist.

Text 100
Datenformat, das innerhalb von *JCAMP*-Dateien verwendet wird und das frei formatierte alphanumerische Einträge bezeichnet, die nicht für die automatische Verarbeitung vorgesehen sind.

tiefes Abschirmfeld 26
vgl. *entschirmt*

tieffeldverschobenen 26
vgl. *entschirmt*

Training 52
In der *Mustererkennung* die Optimierung der internen Parameter eines Systems durch Präsentation einer *repräsentativen klassifizierten Stichprobe*.

ungesättigte Verbindungen 17
Organische Verbindungen, die Mehrfachbindungen enthalten.

unsichere Information 58
Unvollständige, vage oder uneindeutige Angaben.

unsicheres Schließen 58
Das Anstellen von Schlußfolgerungen basierend auf unvollständigem Wissen oder *unsicherer Information*.

Valenzelektronen 5
Bezeichnung für die äußeren *Elektronen* eines Atoms; diejenigen Elektronen, die an der Ausbildung von Bindungen beteiligt sind.

Valenzschalenerweiterung 16
Fähigkeit mancher chemischer Elemente (z.B. des Schwefels), mehr als acht *Valenzelektronen* aufzunehmen; das heißt im besonderen auch, daß sie mehr als vier *Elektronenpaarbindungen* ausbilden können.

Valenzstrichschreibweise 7
Notation zur Wiedergabe von *Strukturformeln*, bei der jeweils ein Elektronenpaar durch einen Strich dargestellt wird. Üblich ist eine verkürzte Schreibweise, bei welcher die Bindungen zwischen Kohlenstoff- und Wasserstoffatomen nicht explizit notiert werden.

Literaturverzeichnis

[Atk96] Peter W. Atkins. *Physikalische Chemie, 2. Auflage*. VCH Verlagsgesellschaft mbH, Weinheim, 1996.

[BGH+95] C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, and T. Wieland. MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure eludication. *Anal. Chim. Acta*, 314:141–147, 1995.

[Bil03] J. Bilmes. *Mathematical Foundations of Speech and Language Processing*, chapter Graphical Models and Automatic Speech Recognition. Institute of Mathematical Analysis Volumes in Mathematics Series. Springer-Verlag, 2003.

[BK02] C. Borgelt and R. Kruse. *Graphical Models. Methods for Data Analysis and Mining*. John Wiley & Sons, 2002.

[BKL96] C. Benecke, A. Kerber, and R. Laue. Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci*, 36:431–439, 1996.

[BKLW96] C. Benecke, A. Kerber, R. Laue, and T. Wieland. Ein anwendungsgebiet der computeralgebra in der industrie: Molekulare strukturerkennung. *Computeralgebra-Rundbrief*, 19, 1996.

[BLM89] Thomas O. Binford, Tod S. Levitt, and Wallace B. Mann. Bayesian inference in model-based machine vision. In *Uncertainty in Artificial Intelligence 3*, Machine Intelligence and Pattern Recognition, pages 73–95. North-Holland, 1989.

[BO03] Olav Bangsø and Kristian G. Olesen. Applying object oriented bayesian networks to large (medical) decision support systems. In *Proceedings of the Eighth Scandinavian Conference on Artificial Intelligence*. IOS Press, 2003.

[Bos95] Karl Bosch. *Elementare Einführung in die Wahrscinelichkeitsrechnung (6. Auflage)*. Vieweg, Braunschweig, 1995.

[Bre78] W. Bremser. HOSE - a novel substructure code. *Anal. Chim. Acta*, 103:355–365, 1978.

[Bre92] Eberhard Breitmaier. *Vom NMR-Spektrum zur Strukturformel organischer Verbindungen*. B. G. Teubner, Stuttgart, 1992.

[BS91] I. N. Bronstein and K. A. Semendjajew. *Taschenbuch der Mathematik (25., durchgesehene Auflage)*. B. G. Teubner, 1991.

[Bud90] Joachim Buddrus. *Grundlagen der Organischen Chemie, 2. Aufl.* Walter de Gruyter, Berlin; New York, 1990.

[BWe91] Hans Beyer and Wolfgang Walter (ed.). *Lehrbuch der Organischen Chemie, 22., überarbeitete und aktualisierte Auflage*. S. Hirzel Verlag, Stuttgart, 1991.

[CF00] James R. Cheeseman and Æleen Frisch. Predicting magnetic properties with chemdraw and gaussian. *URL: http://www.gaussian.com*, 2000.

[Coz] Fabio Cozman. The interchange format for bayesian networks. URL http://www.cs.cmu.edu/afs/cs/user/fgozman/www/Research/InterchangeFormat.

[CR79] C. J. Colburn and R. C. Read. Orderly algorithms for generating restricted classes of graphs. *Journal of Graph Theory*, 3:187–195, 1979.

[Dec96] Rina Dechter. Bucket elimination: A unifying framework for probabilistic inference. In Eric Horvitz and Finn V. Jensen, editors, *Twelfth Conference on Uncertainty in Artificial Intelligence*, 1996.

[Die92] Dieter Ströhl, Stefan Thomas, Erich Kleinpeter, Reiner Radeglia und Joachim Brunn. ^{13}c-nmr-untersuchungen von substituenteneffekten in mehrfach substituierten benzen- und naphtalenverbindungen: Inkrementberechnungen der ^{13}c-chemischen verschiebungen. *Monatshefte für Chemie*, 123:769–777, 1992.

[DL93] Antony N. Davies and Peter Lampen. JCAMP-DX for NMR. *Applied Spectroscopy*, 47(8):1093–1099, 1993.

[Dud73] Richard O. Duda. *Pattern Classification and Scene Analysis*. Wiley Interscience, 1973.

[Ewi79] D.F. Ewing. ^{13}c substituent effects in monosubstituted benzenes. *Organic Magnetic Resonance*, 12:499–524, 1979.

[Fin03] Gernot A. Fink. *Mustererkennung mit Markov-Modellen*. Teubner Studienbücher, Wiesbaden, 2003.

[Fle87] Roger Fletcher. *Practical Methods of Optimization*. Wiley Interscience, 1987.

[FS96] K. Funatsu and S. Sasaki. Recent advances in the automated structure eludidation 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.

[gau04] 2004. URL: http://www.gaussian.com.

[GHH+91] J. Gasteiger, B. M. P. Hendiks, P Hoever, C. Jochum, and H. Somberg. JCAMP-DX: A standard exchange format for chemical structure information in computer readable form. *Applied Spectroscopy*, 45(1):4–11, 1991.

[Jac99] Peter Jackson. *Introduction to Expert Systems*. Addison-Wesley, 1999.

[Jen96] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, London, 1996.

[Jer86] M. Jerrum. A compact representation of permutation groups. *J. Alg.*, 7:60–78, 1986.

[Kat87] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

[KB94] T. M. Klingler and D. L. Brutlag. Discovering structural correlations in alpha-helices. *Protein Science*, 3(10):1847–1857, 1994.

[KLS03] F.T. Koch, P. Losso, and U. Sternberg, 2003. URL: http://www.cosmos-software.de.

[Lau96] S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[LTS94] S. L. Lauritzen, B. Thiesson, and D. J. Spiegelhalter. Diagnostic systems by model selection: A case study. In *Artificial Intelligence and Statistics IV: Selecting Models from Data*, Lecture Notes in Statistics 89, pages 143–152. Springer-Verlag, 1994.

[Mea02] J. Meiler et al. Validation of structure proposals by substructure analysis and ^{13}C NMR chemical shift prediction. *J. Chem. Inf. Comput. Sci.*, 42:241–248, 2002.

[MM02] J. Meiler and M. Meringer. Ranking MOLGEN structure proposals by ^{13}C NMR chemical shift prediction with ANALYZE. *MATCH Communications in Mathematical and in Computer chemistry*, 45:86–108, 2002.

[MMW00] J. Meiler, R. Meusinger, and M. Will. Fast determination of ^{13}C-NMR chemical shifts using artificial neural networks. *J. Chem. Inf. Comput. Sci*, 40:1169–1176, 2000.

[MMWM02] Jens Meiler, Walter Maier, Martin Will, and Reinhard Meusinger. Using neural networks for ^{13}C NMR chemical shift prediction – comparison with traditional methods. *J. Magn. Res.*, 157:242–252, 2002.

[Mor01] Charles E. Mortimer. *Chemie – Das Basiswissen der Chemie, 7. korrigierte Auflage*. Georg Thieme Verlag, Stuttgart, 2001.

[Mun98] Morton E. Munk. Computer-based structure elucidation: Then and now. *J. Chem. Inf. Comput. Sci.*, 38:997–1009, 1998.

[MW01] J. Meiler and M. Will. Automated structure elucidation of organic molecules from ^{13}C NMR spectra using genetic algorithms and neural networks. *J. Chem. Inf. Comput. Sci.*, 41:1535–1546, 2001.

[MW02] J. Meiler and M. Will. Genius: A genetic algorithm for automated structure elucidation from ^{13}C NMR spectra. *J. Am. Chem. Soc.*, 124:1868–1870, 2002.

[MWJ88] Robert S. McDonald and Paul A. Wilks Jr. JCAMP-DX: A standard form for exchange of infrared spectra in computer readable form. *Applied Spectroscopy*, 42(1):151–162, 1988.

[NEK94] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech & Language*, 8:1–38, 1994.

[Nie83] Heinrich Niemann. *Klassifikation von Mustern*. Springer-Verlag, 1983.

[Nie90] Heinrich Niemann. *Pattern Analysis and Understanding*. Springer-Verlag, 1990.

[Nil98] Nils J. Nilsson. *Artificial Intelligence. A New Synthesis*. Morgan Kaufmann Publishers, San Francisco, 1998.

[Pea86] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Francisco, 1988.

[Rea78] R. C. Read. Every-one a winner. *Annals of Discrete Mathematics*, 2:107–120, 1978.

[RHR70] E. Ruch, W. Hässelbarth, and B. Richter. Doppelnebenklassen als klassenbegriff und nomenklaturprinzip für isomere und ihre abzählung. *Theor. Chim. Acta*, 19:288–300, 1970.

[RK83] E. Ruch and D. J. Klein. Double cosets in chemistry an physics. *Theoret. Chim. Acta (Berl.)*, 63:447–472, 1983.

[RS98] Paul von Rague Schleyer, editor. *Encyclopedia of Computational Chemistry*. John Wiley & Sons, 1998.

[Sch96] Jürgen Schürmann. *Pattern Classification*. Wiley Interscience, 1996.

[Smy97] P. Smyth. Belief networks, hidden markov models and markov random fields: A unifying view. *Pattern Recognition Letters*, 18, 1997.

[Som92] Léa Sombé. *Schließen bei unsicherem Wissen in der künstlichen Intelligenz*. Vieweg, Braunschweig, 1992.

[Spe98] Specinfo, 1998. URL: http://specinfo.wiley-vch.de/SI32/start.htm.

[Spi84] R.P. Spiegelhalter, D.J und Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems. *Journal of the Royal Statistical Society, Series A*, 147:35–77, 1984.

[STK94] D. Ströhl St. Thomas and E. Kleinpeter. Computer application of an incremental system for calculating the 13 c nmr spectra of aromatic compounds. *Journal of Chemical Information and Computer Sciences*, 1994.

[Str89] Ströhl, D. and Radeglia, R. and Brunn, J. and Fanghänel, E. 1,3,4,5-tetrasubstituierte benzene. *Journal für praktische Chemie*, 331:347–353, 1989.

[ucl02] 2002. URL: http://www.chem.ucla.edu/˜bacher/General/30BL/NMR/Ccorrben.html.

[VVD+87] M. G. Voronkov, N. S. Vyazankin, E. N. Deryagina, A.S. Nakhmanovich, and V. A. Usov. *Reactions of Sulfur with Organic Compounds*. Plenum Publishing Corporation, Consultants Bureau, New York, 1987.

[Wac01] Sven Wachsmuth. *Multi-modal Scene Understanding Using Probabilistic Models*. PhD thesis, Universität Bielefeld, Technische Fakultät, 2001.

[WFR96] M. Will, W. Fachinger, and J. R. Richert. Fully automated structure elucidation - a spectroscopist's dream comes true. *J. Chem. Inf. Comput. Sci.*, 36:221–227, 1996.

[Wie97] Thomas Wieland. Konstruktionsalgorithmen bei molekularen graphen und deren anwendung. *MATCH Communications in Mathematical and in Computer Chemistry*, 36:7–157, 1997.

[Win92] Patrick Henry Winston. *Artificial Intelligence (3rd Edition)*. Addison-Wesley, 1992.

[Wit03] Raiker Witter. *Three Dimensional Structure Elucidation with the COSMOS-NMR Force Field*. PhD thesis, Friedrich-Schiller-Universität, Jena, 2003.

[WR97] M. Will and J. R. Richert. Specsolv – an innovation at work. *J. Chem. Inf. Comput. Sci.*, 37:403–404, 1997.

[wul04] 2004. URL: http://wulfenite.fandm.edu/Data /Table_34.html.