Klassifizierung organischer Reaktionen mittels neuronaler Netze zur Anwendung in Reaktionsvorhersage und Syntheseplanung

Dissertation

Oliver Sacher

2001

Klassifizierung organischer Reaktionen mittels neuronaler Netze zur Anwendung in Reaktionsvorhersage und Syntheseplanung

Den Naturwissenschaftlichen Fakultäten der
Friedrich-Alexander-Universität Erlangen-Nürnberg

zur
Erlangung des Doktorgrades

vorgelegt von

Oliver Sacher

aus Nürnberg

Als Dissertation genehmigt von
den Naturwissenschaftlichen Fakultäten der Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	07.02.2001
Vorsitzender der Promotionskommission:	Prof. Dr. A. Magerl
Erstberichterstatter:	Prof. Dr. J. Gasteiger
Zweitberichterstatter:	Prof. Dr. A. Hirsch
Drittberichterstatter:	Prof. Dr. J. Zupan

Meinem Doktorvater

Herrn Prof. Dr. Johann Gasteiger

danke ich für die vielfältige Unterstützung und die wertvollen Anregungen, ohne die diese Arbeit nicht möglich gewesen wäre.

Weiteren Dank schulde ich

Herrn Dr. Lingran Chen für die grundlegenden Arbeiten auf dem Gebiet der Reaktionsklassifizierung mit Kohonen-Netzen

meinen Kollegen der Reaktionsvorhersage- und Syntheseplanungs-Gruppe Herrn Dr. Robert Höllering, Herrn Thomas Kostka, Herrn Norbert Karg, Herrn Achim Herwig, Herrn Dr. Ralf Fick, Herrn Dr. Matthias Pförtner und Herrn Markus Sitzmann für die gute Zusammenarbeit und die vielen interessanten Diskussionen

Herrn Dr. Markus Wagener, Herrn Dr. Robert Höllering, Herrn Dr. Paul Selzer, Herrn Dr. Andreas Teckentrup und Herrn Dr. Lothar Terfloth für die Administration einer stabilen Unix-Plattform

Herrn Dr. Jan Schuur, Herrn Norbert Karg, Frau Angelika Hofmann sowie Herrn Jörg Marusczyk für die gute Zusammenarbeit bei der Administration der PC-Domäne

Herrn Achim Herwig und Herrn Markus Sitzmann für das Einrichten und Warten einer effektiven Linux-Arbeitsumgebung

Herrn Dr. Wolf-Dietrich Ihlenfeldt für die zahlreichen Hilfestellungen bei programmiertechnischen Problemen

Herrn Dr. Markus Wagener, Herrn Dr. Andreas Teckentrup und Herrn Thomas Kleinöder für die Betreuung des KMAP- und PETRA-Programmsystems

unserer Sekretärin Frau Angela Döbler für die Unterstützung bei den administrativen Tätigkeiten im universitären Alltag und allen nicht namentlich genannten Kolleginnen und Kollegen für die sehr gute Arbeitsatmosphäre

meiner Freundin Frau Christiane Kohler für die Unterstützung und die Geduld, die sie mir beim Verfassen dieser Arbeit entgegengebracht hat.

Für die finanzielle Unterstützung dieser Arbeit gebührt Dank:

der Friedrich-Alexander-Universität Erlangen-Nürnberg für die Gewährung eines 2-jährigen Promotionsstipendiums

der Deutschen Forschungsgemeinschaft DFG.

Für meine Eltern

1 Einleitung

Fachzeitschriften bilden die wichtigste Informationsquelle für Chemiker. Als die ersten Chemie-Zeitschriften erschienen, 1778 das *Chemische Journal* oder 1789 die französischen *Annales de Chimie*, war die chemische Information noch überschaubar. Noch bevor im Jahre 1867 die *Chemischen Berichte* erstmals veröffentlicht wurden, erkannte man bereits die Notwendigkeit zur Einrichtung eines Referateorgans, um Publikationen aus dem chemischen oder chemieverwandten Fachgebiet möglichst umfassend zu erschließen. Ein solcher Referatedienst wurde ab dem Jahre 1830 zunächst unter dem Namen *Pharmaceutisches Centralblatt* gegründet und ab dem Jahre 1897 unter dem Namen *Chemisches Zentralblatt* herausgegeben. Mit Hilfe dieser kurzen Inhaltszusammenfassungen war man bis ins Jahr 1969 – als dieses Referateorgan eingestellt wurde – in der Lage, die exponentiell zunehmende Informationsmenge recherchierbar zu halten. Heutzutage wird hauptsächlich das amerikanische Pendant, der seit 1907 bestehende *Chemical Abstracts Service* (CAS) der American Chemical Society, verwendet[1]. Die Anzahl der in CAS aufgenommenen Abstracts steigt weiterhin exponentiell an. Neben der wachsenden Anzahl an chemischen Veröffentlichungen nimmt natürlich auch die Zahl der chemischen Verbindungen und Reaktionen exponentiell zu. Aktuelle Zahlen belegen die immense Informationsflut:

- 20.000.000 chemische Verbindungen sind derzeit registriert
- 500.000 chemische Verbindungen kommen pro Jahr hinzu
- 600.000 Veröffentlichungen erscheinen pro Jahr im Bereich der Chemie

Die vorhandene und jährlich neu hinzukommende chemische Information hat bereits seit einigen Jahrzehnten ein Ausmaß angenommen, das nur noch mit elektronischen informationsverarbeitenden Systemen bewältigt werden kann. Computer sind heutzutage unter anderem bei der Verwaltung oder Visualisierung der riesigen Datenmengen unersetzlich geworden.

Relativ spät erst begann man die publizierten Reaktionen in elektronischer Form zu speichern. Um der bis dahin bereits unüberschaubar gewordenen Flut an Reaktionen zu begegnen, wurden Reaktionsdatenbanken aufgebaut, in denen Reaktionen der Organischen Chemie in elektronisch recherchierbarer Form abgespeichert wurden. Den ersten Datenbanken, wie REACCS (***Re**action **Acc**ess **S**ystem*) von MDL und SYNLIB (***Syn**thesis **Lib**rary*) im Jahre 1982 folgten weitere bekannte Reaktionsdatenbanken, beispielsweise die ChemInform RX Datenbank von FIZ Chemie im Jahre 1991. Diese Datenbanken enthielten einige tausend bis mehrere hunderttausend Reaktionen, die bis in das Jahr 1946 zurückreichten. Mit dem Erscheinen von CrossFire*plus*Reactions von Beilstein Information Systems im Jahre 1996

hatte man erstmals elektronischen Zugang zu über 5.000.000 Reaktionen, die sogar aus den ältesten Chemie-Zeitschriften ab dem Jahr 1789 entnommen wurden.

Angesichts dieser ständig zunehmenden Informationsmenge stellt sich natürlich die Frage, ob man das in Reaktionsdatenbanken gespeicherte Wissen nicht nutzbringend extrahieren und für chemische Anwendungen dienlich einsetzen kann. Als Einsatzmöglichkeiten sind alle chemischen Anwendungen denkbar, bei denen organische Reaktionen behandelt werden. Zum einen ist hier die computergestützte Syntheseplanung zu nennen, die zu einem vorgegebenen Zielmolekül die entsprechenden Ausgangsverbindungen ermittelt, die sich über mehrere Reaktionsschritte unter bestimmten Reaktionsbedingungen in möglichst hoher Ausbeute in das Targetmolekül überführen lassen. Zum anderen stehen auch bei der Reaktionsvorhersage organische Reaktionen im Vordergrund des Interesses. In diesem Fall werden vorgegebene Ausgangsverbindungen unter bestimmten Reaktionsbedingungen umgesetzt, wobei das Produkt und eventuell weitere Nebenprodukte ermittelt werden.

Das Zusammenfassen ähnlicher Objekte zu Gruppen, das sogenannte Klassifizieren von Objekten, stellt eine Möglichkeit der Wissensextraktion dar. Die den Reaktionsdatenbanken entnommenen Reaktionen werden dabei nach ihrer Ähnlichkeit in gemeinsame Gruppen zusammengefaßt. Die Ähnlichkeit von Reaktionen kann je nach Aufgabenstellung oder Einsatzgebiet unterschiedlich definiert sein, beispielsweise topologisch basierend auf der Konnektivität oder auf elektronischen Eigenschaften. In dieser Arbeit wird ein Ähnlichkeitsvergleich vorgestellt, der hauptsächlich auf physikochemischen Kriterien basiert.

Zur Wissensextraktion aus Datenbanken wurden bereits mehrere Methoden entwickelt. Sowohl mit induktiven Lernalgorithmen (ISOLDE), als auch mit deduktiven Verfahren (TRISTAN) generalisierte man zunächst die Einzelreaktionen einer Datenbank[2]. Aufbauend auf diesen Arbeiten wurden weitere Klassifizierungsmethoden, wie HORACE, erarbeitet, die auf einer hierarchischen Clusteranalyse basieren[3],[4]. Seit dem Aufkommen künstlicher neuronaler Netze und ihrer raschen und intensiven Weiterentwicklung steht eine Methode zur Verfügung, die im Bereich der Wissensextraktion eine Reihe von Vorteilen aufzuweisen hat. Hauptsächlich zeichnen sich neuronale Netze durch ihre Lernfähigkeit aus. Ähnlich dem Lernvorgang eines Chemikers, der aus einer Reihe von Einzelbeobachtungen sich sein chemisches Wissen aneignet, erwerben auch neuronale Netze induktiv Wissen aus Einzelobjekten eines Datensatzes. Aus diesem Grund wurde von Chen et al. ein verbessertes Klassifizierungsverfahren entwickelt, das sich diese Lernfähigkeit der neuronalen Netze zunutze macht[5]. Allerdings mußte diese Methode für jeden zu untersuchenden Reaktionstyp neu angepaßt werden, eine gemeinsame Klassifizierung verschiedener Reaktionstypen – wie sie in Reaktionsdatenbanken vorkommen – war nicht möglich. Am Beispiel der nucleophilen aliphatischen Substitution von Acylchloriden und Michael-Additionen wurde dieses Klassifizierungsverfahren ausführlich vorgestellt[6].

Im Rahmen dieser Arbeit sollte ein Klassifizierungsverfahren basierend auf der Lernfähigkeit neuronaler Netze entwickelt werden, mit dem beliebig viele unterschiedliche Reaktionstypen gleichzeitig eingeteilt werden können (siehe Abbildung 1-1). Das aus Reaktionsdatenbanken erworbene Wissen soll dabei nutzbringend in chemischen Anwendungen, wie der Syntheseplanung und Reaktionsvorhersage, eingesetzt werden.

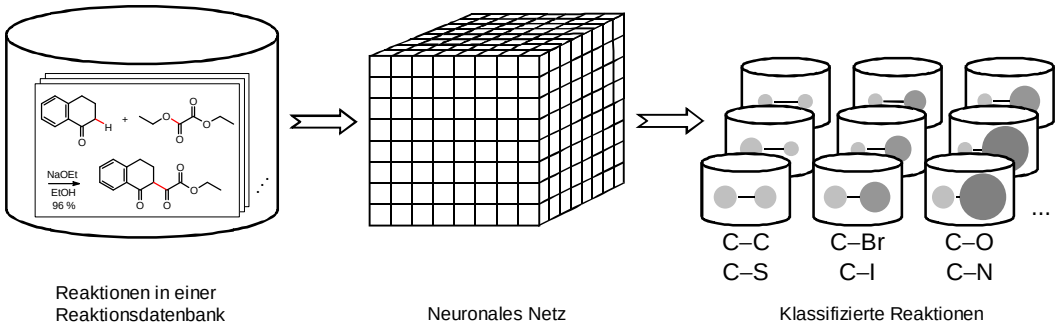


Abb. 1-1: Schematischer Ablauf der Reaktionsklassifizierung mittels neuronaler Netze.

Die vorliegende Arbeit gliedert sich in folgende Abschnitte:

- In Kapitel 2 werden zunächst wichtige Begriffe definiert, und anschließend die in dieser Arbeit verwendeten Reaktionsdatenbanken vorgestellt. Weiterhin wird auf die Funktionsweise der neuronalen Netze eingegangen, und das eingesetzte Programmsystem zur Berechnung physikochemischer Deskriptoren erläutert.
- Im Methodenteil in Kapitel 3 wird das im Rahmen dieser Arbeit entwickelte Codierungsverfahren für Reaktionen beschrieben, das prinzipiell sehr variabel konfiguriert werden kann, jedoch meist nach einem Standardverfahren angewendet wird.
- Als erstes Anwendungsbeispiel der Reaktionsklassifizierung wird in Kapitel 4 der Vergleich von Reaktionsdatenbanken diskutiert. Es werden dazu Reaktionen aus verschiedenen Reaktionsdatenbanken klassifiziert und miteinander verglichen. Die zeitliche Entwicklung einer Reaktionsdatenbank wird dargestellt, um Tendenzen in der organischen Synthesechemie aufzuzeigen.
- Kapitel 5 befaßt sich mit einer neu entwickelten Methode zur Planung organischer Synthesen, die auf der Reaktionsklassifizierung basiert. Anhand zweier Beispiele wird sowohl die Bestimmung strategischer Bindungen als auch deren Bewertung erörtert.
- Als drittes Anwendungsbeispiel steht bei der Reaktionsvorhersage in Kapitel 6 die Vorhersage des Hauptreaktionsproduktes im Vordergrund. Am Beispiel der Pyrazolsynthese ermittelt ein neuronales Netz für vorgegebene Ausgangsverbindungen das bevorzugt gebildete Regioisomer.
- Die seit einigen Jahren vor allem in Pharmakonzernen intensiv eingesetzte kombinatorische Chemie treibt das rasante Wachstum an chemischer Information explosionsartig

voran. In Kapitel 7 wird anhand zweier kombinatorischer Datensätze erläutert, wie die Reaktionsklassifizierung in diesem Bereich bei der Planung und der Analyse sinnvoll eingesetzt werden kann.

- Die hier vorgestellten Arbeiten fließen in ein neu entwickeltes Programmsystem namens CORA (***C****lassification of* ***O****rganic* ***R****eactions for* ***A****pplications*) ein, auf das ausführlich in Kapitel 8 eingegangen wird.
- In abschließenden Kapiteln wird die im Rahmen dieser Arbeit entwickelte Methode zur Codierung und Klassifizierung organischer Reaktionen diskutiert und zusammengefaßt.

Bei der Darstellung der Ergebnisse wurde darauf Wert gelegt, etablierte informationstechnische Mittel einzusetzen, die eine weltweite Einsichtnahme ermöglichen. Daher sind viele Ergebnisse, besonders wenn deren Umfang den Rahmen dieser Arbeit einschließlich des Anhangs sprengen würde, im Detail im World-Wide-Web abrufbar. Auf einer Übersichtsseite im Anhang sind deshalb alle Webseiten, die zusätzliche Information zu Ergebnissen dieser Arbeit liefern, zusammengefaßt.

2 Grundlagen

In diesem Kapitel werden zunächst wichtige Begriffe, die bei der Behandlung elektronisch erfaßter Reaktionen auftreten, erläutert, sowie die in dieser Arbeit eingesetzten Reaktionsdatenbanken vorgestellt. Danach wird auf die Funktionsweise neuronaler Netze eingegangen, insbesondere auf die Kohonen-Netze.

2.1 Begriffserläuterungen

2.1.1 Reaktionszentrum

Unter dem *Reaktionszentrum* versteht man die Atome und Bindungen, die beim Ablauf einer chemischen Reaktion direkt an einer Reorganisation von Elektronen beteiligt sind. Im Falle der in Abbildung 2-1 dargestellten Friedel-Crafts-Alkylierung sind insgesamt vier Bindungen am Bindungsumordnungsprozess beteiligt: Zwei Bindungen werden gebrochen und zwei neu geknüpft.

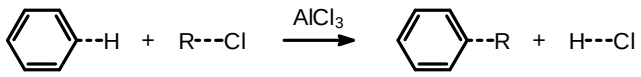


Abb. 2-1: Friedel-Crafts-Alkylierung mit hervorgehobenem Reaktionszentrum.

2.1.2 Reaktionstyp

Ein *Reaktionstyp* ist eine Gruppe von Reaktionen, die aufgrund ähnlicher Substituenten an den Atomen des Reaktionszentrums oder aufgrund gleicher Reaktionsmechanismen zusammengefaßt werden.

Reaktionen eines Reaktionstyps werden häufig mit Namens- oder Schlagwortreaktionen bezeichnet, wie Friedel-Crafts-Alkylierung, Diels-Alder-Reaktion, nucleophile Substitutionsreaktion etc.[7].

2.1.3 Reaktionsgeneratoren

Reaktionsgeneratoren finden ihren Einsatz beim Generieren von Reaktionen. Sie geben das Reaktionszentrum einer Reaktion in einer sehr allgemeinen Form wieder. Die wichtigsten zwei Generatoren, die den Großteil des Reaktionsverlaufs organischer Reaktionen beschreiben können, werden im folgenden kurz vorgestellt[8]. Die Reaktionsgeneratoren werden nach der Anzahl der gebrochenen und geknüpften Bindungen benannt.

2.1.3.1 Reaktionsgenerator 2.2

Bei diesem Reaktionsgenerator werden im Reaktionsverlauf in den Eduktmolekülen zwei Bindungen gebrochen und auf der Produktseite zwei Bindungen geknüpft (siehe Abbildung 2-2).

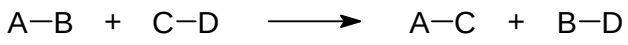


Abb. 2-2: Der Reaktionsgenerator 2.2, bei dem zwei Bindungen auf der Eduktseite gebrochen und zwei Bindungen auf der Produktseite gebildet werden.

Bei der in Abbildung 2-1 dargestellten Friedel-Crafts-Alkylierung werden auf der Eduktseite eine Kohlenstoff-Wasserstoff-Bindung und eine Kohlenstoff-Chlor-Bindung gebrochen sowie auf der Produktseite eine Kohlenstoff-Kohlenstoff-Bindung und eine Wasserstoff-Chlor-Bindung neu gebildet. Diese Reaktion kann deshalb mit dem Reaktionsgenerator 2.2 erzeugt werden.

2.1.3.2 Reaktionsgenerator 3.3

Eine Reaktion läuft nach dem Reaktionsgenerator 3.3 ab, wenn während der Reaktion drei Bindungen in den Eduktmolekülen gebrochen und drei Bindungen in den Produktmolekülen geknüpft werden. Der Bindungsumordnungsprozeß ist in allgemeiner Form in Abbildung 2-3 dargestellt.

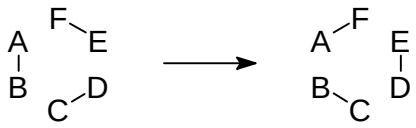


Abb. 2-3: Der Reaktionsgenerator 3.3, bei dem drei Bindungen auf der Eduktseite gebrochen und drei Bindungen auf der Produktseite gebildet werden.

Bei einer Diels-Alder-Reaktion wird formal die Bindungsordnung der Doppelbindung im Dienophil und die der beiden Doppelbindungen im Dien um eins erniedrigt, während im Produktmolekül zwei Einfachbindungen neu gebildet werden und die Bindungsordnung der ursprünglichen Einfachbindung im Dien um eins erhöht wird. In Abbildung 2-4 ist das Reaktionszentrum für die Synthese von Cyclohex-4-en-1,2-dicarbonsäureanhydrid aus Butadien und Maleinsäureanhydrid dargestellt.

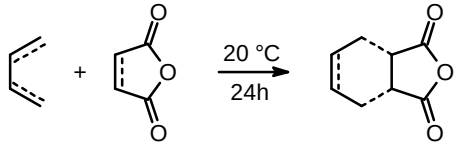


Abb. 2-4: Diels-Alder-Reaktion als Beispiel für eine Reaktion nach dem Reaktionsgenerator 3.3.

2.2 Reaktionsdatenbanken

In Reaktionsdatenbanken sind chemische Reaktionen in computerlesbarer Form abgespeichert. Über sogenannte Retrieval-Systeme hat der Benutzer Zugang zu den Reaktionen in einer Datenbank. Gegenwärtig sind ca. 14 Millionen Reaktionen in Datenbanken erfaßt, davon sind allein ca. 10 Millionen in der CrossFire*plus*Reactions Datenbank gespeichert. Obwohl eine Version der CrossFire Datenbank mit insgesamt 5.139.668 Reaktionen zugänglich ist, ist eine Untersuchung von Reaktionen aus dieser Datenbank nicht leicht möglich, da die elektronischen Exportmöglichkeiten am Computer-Chemie-Centrum aus lizenzrechtlichen Gründen eingeschränkt sind. Eine uneingeschränkte Exportfunktionalität bietet dagegen das ISIS-Retrieval System von Molecular Design Ltd. (MDL). Die Reaktionsdatenbanken dieses Anbieters stehen als Inhouse-Datenbanken zur Verfügung. In diesem Kapitel werden alle in dieser Arbeit eingesetzten Reaktionsdatenbanken vorgestellt und beurteilt.

2.2.1 Theilheimer Reaktionsdatenbank

Die eingesetzte Theilheimer Reaktionsdatenbank[9] trägt die Versionsbezeichnung 90.1.4 und umfaßt insgesamt 46.785 Reaktionen. Sie basiert auf den Jahrbüchern „Synthetic Methods of Organic Chemistry“ (Karger), Ausgabe 1-35 der Jahre 1946 bis 1980. Diese jährlichen Abstracts beinhalten Informationen im Bereich der synthetischen, organischen Chemie. Die Theilheimer Reaktionsdatenbank und die originalen Jahrbücher wurden aus Reaktionsdaten von über 600 Zeitschriften zusammengetragen. Zusätzliche Information floß aus den Chemical Abstracts mit ein. In der Tabelle 2-1 sind die Fachzeitschriften und Sekundärdienste aufgeführt, die in der Theilheimer Datenbank am häufigsten genannt sind. Man findet darin die renommiertesten internationalen Zeitschriften der Chemie, wobei allein die ersten beiden Zeitschriften der American Chemical Society annähernd 30% ausmachen.

Rang	Name der Zeitschrift	Literaturstellen	Literaturst. [%]
1	J. Am. Chem. Soc.	19.106	16,0
2	J. Org. Chem.	15.361	12,9
3	J. Chem. Soc.	8.125	6,8
4	Chem. Abstr.	6.937	5,8
5	Ber. Dtsch. Chem. Ges.	6.695	5,6
6	Tetrahedron. Lett.	5.485	4,6
7	Synth. Met.	4.667	3,9
8	Helv. Chim. Acta.	4.193	3,5
9	Org. Synth.	2.824	2,4
10	Annalen der Chemie	2.659	2,2

Tab. 2-1: Zitierte Fachzeitschriften in der Theilheimer Reaktionsdatenbank; es sind die 10 häufigsten Zeitschriften angegeben. Chemical Abstracts zählt zu den Sekundärdiensten.

Die zu jeder Reaktion gespeicherten Literaturstellen geben generell Auskunft über das Jahr, in dem die Reaktion entdeckt oder eingehender untersucht wurde. Meistens übersteigt die Zahl der Literaturstellen die Anzahl der Reaktionen, da pro Reaktion häufig mehrere Literaturstellen aufgeführt sind. In der Theilheimer Reaktionsdatenbank sind für die 46.785 Reaktionen beispielsweise 107.968 Literaturstellen abgespeichert worden. In Abbildung 2-5 ist die Verteilung der Literaturjahrgänge in Form eines Histogramms für diese Reaktionsdatenbank dargestellt.

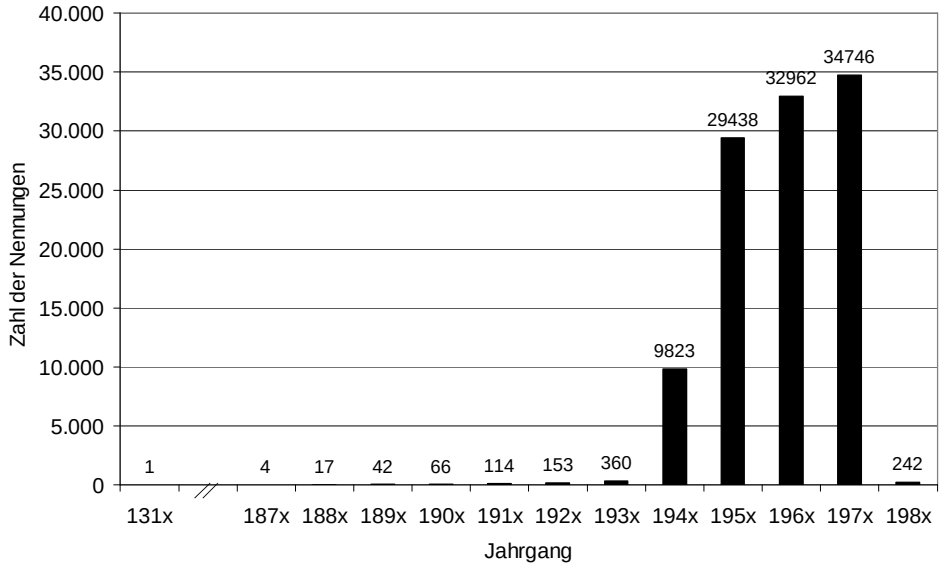


Abb. 2-5: Histogramm nach Jahrzehnte der in der Theilheimer Datenbank aufgeführten Literaturjahrgänge (x steht für eine Ziffer von 0 bis 9).

Man erkennt, daß in der Theilheimer Reaktionsdatenbank zu über 99% Reaktionen aus den Jahren 1940 bis 1979 eingegangen sind, wobei das Hauptgewicht bei den Reaktionen der Jahre 1970 bis 1979 liegt. Es fehlen dagegen alle Reaktionen der vergangenen 20 Jahre. Auffallend sind zum einen vier Reaktionen (RTHE00005959, RTHE00006707, RTHE00004601 und RTHE00002738), bei denen auf Literaturstellen der Jahre 1878 bzw. 1879 verwiesen wird. Zum anderen stammt die älteste Reaktion aus dem 13. Jahrhundert (siehe Abbildung 2-6). Bei dieser Reaktion handelt es sich um eine alchimistische Reaktion, die wohl nur die Ursprünge der modernen Chemie ins Bewußtsein zurückrufen möchte!

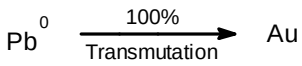


Abb. 2-6: Reaktion #46774 (RTHE00011499) aus der Theilheimer Reaktionsdatenbank.

In die Theilheimer Reaktionsdatenbank wurden vor allem Reaktionen aufgenommen, die folgende Kriterien erfüllten:

- neu oder ungewöhnlich
- bekannte Verfahren, die aber von neuen Reagenzien oder verbesserten synthetischen Methoden geprägt sind
- Überführung funktioneller Gruppen, die von allgemeinem Interesse sind
- Reaktionen zum Aufbau von Kohlenstoffgerüsten
- Aufbau neuer Verbindungsklassen und funktioneller Gruppen, die von synthetischem Interesse sind
- bekannte Reaktionen mit interessanten Erweiterungen und Anwendungen
- Reviews (Übersichtsveröffentlichungen), die sich synthetischen Beschreibungen der organischen Chemie widmen
- ergänzende Information, wie beispielsweise die Darstellung der Reagenzien und detaillierte experimentelle Beschreibungen.

Im Bereich der heterocyclischen Chemie wurden allerdings Reaktionen, die den Aufbau von neuen Ringsystemen beschreiben, nicht mit aufgenommen. Ebenso wurden in den meisten Fällen keine metallorganischen Reaktionen berücksichtigt[10].

In die elektronische Version von William Theilheimer's „Synthetic Methods of Organic Chemistry" wurden aber nicht nur Reaktionen aufgenommen, die mindestens eine der obigen Bedingungen erfüllten, sondern sie mußten darüber hinaus immer in hoher Ausbeute ablaufen. Um dieses Kriterium zu überprüfen, stellt man die Verteilung der Reaktionsausbeute als Histogramm dar. In der Datenbank fand man zu 36.328 (77,6%) von insgesamt 46.785 Reaktionen jeweils eine Angabe zur Ausbeute. Wie man anhand Abbildung 2-7 erkennt, ist das Kriterium einer hohen Ausbeute in der Theilheimer Reaktionsdatenbank sehr gut verwirklicht worden. Nur einige hundert Reaktionen weisen eine Ausbeute kleiner als 50% auf. Das Maximum dieser Verteilungsfunktion liegt bei 85% bis 90%.

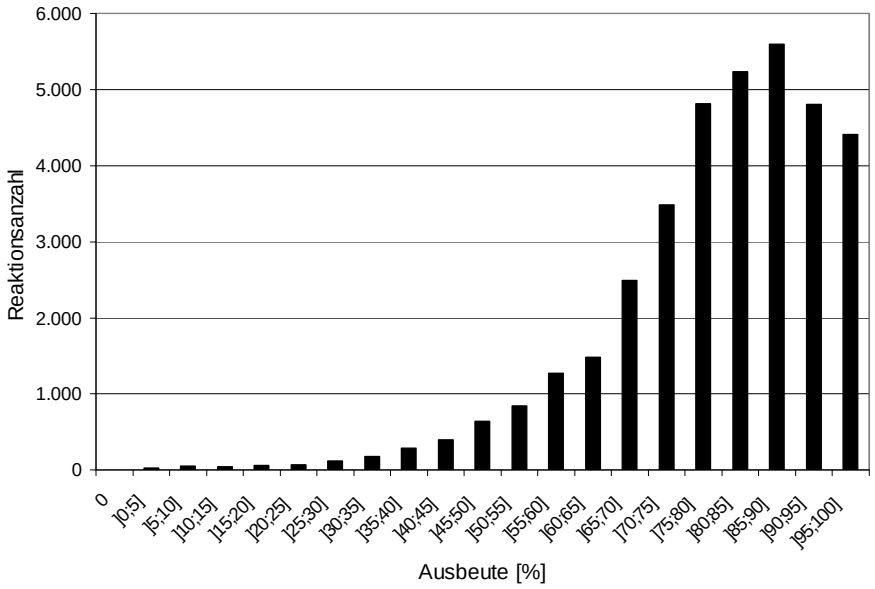


Abb. 2-7: Verteilung der Reaktionsausbeuten in der Theilheimer Datenbank.

Da diese Reaktionsdatenbank nicht mehr aktualisiert wird stellt sie eine abgeschlossene Datenbank dar. Sowohl die Vielzahl der eingegangenen Journale, der überstrichene Zeitraum von rund 40 Jahren und die Aufnahmekriterien machen die Theilheimer Reaktionsdatenbank zu einem Reaktionsreservoir, in dem die wichtigsten Reaktionstypen der organischen Chemie enthalten sind. Aus diesem Grund wird diese Datenbank in der vorliegenden Arbeit oft als Referenzdatenbank herangezogen (siehe Kapitel 4.1).

2.2.2 ChemInform RX-Reaktionsdatenbank

Die ChemInform RX-Reaktionsdatenbank[11] basiert auf dem wöchentlich erscheinenden „Chemischen Informationsdienst“, der zur Zeit vom Fachinformationszentrum Chemie GmbH und der Gesellschaft Deutscher Chemiker herausgegeben wird. Dieser renommierte Referatedienst wird seit 1969 von der Wiley-VCH Verlagsgesellschaft verlegt. Der Informationsdienst zitiert organische und elementorganische Veröffentlichungen aus der klassischen Synthese. Die Einträge in ChemInform sind mit kurzen Zusammenfassungen (Abstracts) des Inhalts der jeweils zitierten Publikation versehen. In ChemInform werden vor allem Reaktionen aufgenommen, die folgende Kriterien erfüllen:

- neue Reaktionen und Synthesen, einschließlich enzymatischer oder mikrobieller Prozesse
- Anwendung bekannter Reaktionen bei der Synthese neuer Verbindungen oder Substanzklassen

- verbesserte, synthetische Methoden, neue Reagenzien
- Synthese von Naturstoffen von allgemeinem Interesse
- Synthese neuer Organometall-Verbindungen und neuer Katalysatoren
- Reaktionen aus Übersichtsartikeln (Reviews).

Publikationen über Polymere oder Polymerisierungen werden aber ebensowenig aufgenommen wie biochemische Beschreibungen, die keine neuen präparativen Methoden enthalten.

Die elektronische Version dieses gedruckten Referatedienstes wird in Form einer Reaktionsdatenbank herausgegeben, die ChemInform RX genannt wird[12]. Jährlich werden ca. 60.000 Reaktionen in einem eigenständigen Jahrgang zusammengefaßt, wobei die Inhouse-Version halbjährlich aktualisiert wird. So fließen beispielsweise die Hefte des Jahres 1996 in die ChemInform RX-Datenbank des Jahres 1997, abgekürzt CIRX97, ein.

In dieser Arbeit kamen die in Tabelle 2-2 aufgeführten ChemInform RX-Reaktionsdatenbanken zum Einsatz.

Jahrgang	Version	Reaktionsanzahl
CIRX92	92.1.3	76.421
CIRX93	93.1.1	67.740
CIRX94	94.1.1	56.236
CIRX95	95.1.1	64.187
CIRX96	96.1	70.271
CIRX97	97.1.1	70.061

Tab. 2-2: Die in dieser Arbeit verwendeten ChemInform RX-Reaktionsdatenbanken.

Neben diesen Inhouse-Versionen, die über MDL Information Systems, Inc. vertrieben werden, gibt es auch eine Online-Version, die über das Fachinformationszentrum (FIZ) Chemie in Berlin zugänglich ist. In der vorliegenden Arbeit wurden allerdings nur die in Tabelle 2-2 aufgeführten Inhouse-Datenbanken verwendet.

Wegen des jährlichen Erscheinens dieser Datenbank stammen die Reaktionen aus einem engen Zeitintervall. So enthalten die Jahrgänge einer ChemInform RX-Datenbank hauptsächlich Reaktionen, die vor 1-2 Jahren veröffentlicht wurden (siehe Abbildung 2-8). In seltenen Fällen gehen auch einige wenige Literaturstellen in die Datenbank ein, die 3 Jahre zurückliegen.

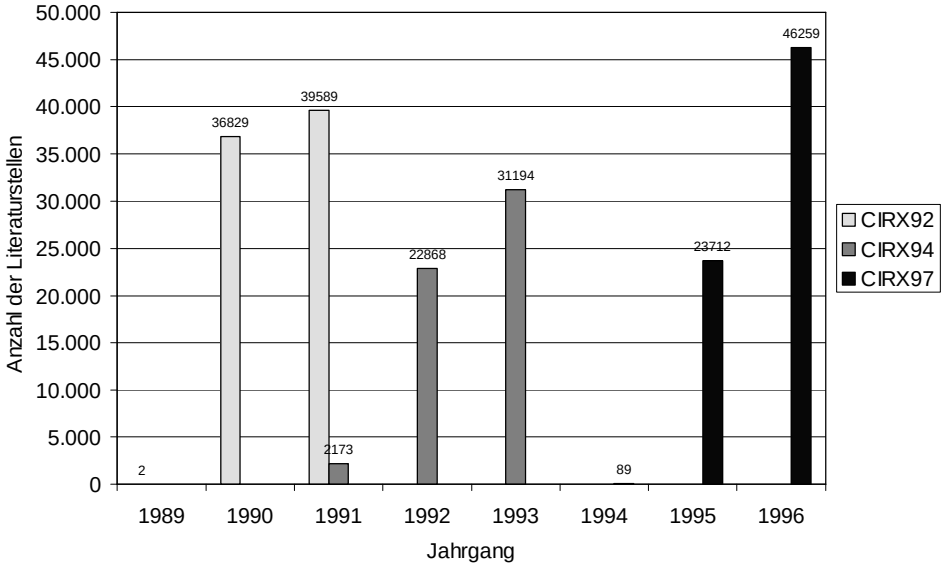


Abb. 2-8: Verteilung der Literaturjahrgänge in drei ausgewählten ChemInform RX-Reaktionsdatenbanken.

In Abbildung 2-9 ist stellvertretend für die ChemInform RX-Datenbanken das Histogramm für die CIRX94 Datenbank wiedergegeben. Im Gegensatz zur Theilheimer Reaktionsdatenbank liegt bei den ChemInform RX-Datenbanken das Maximum der Ausbeutenverteilung bei 75-80%. Auffallend ist außerdem, daß viele Ausbeutewerte kleiner als 50% sind, nämlich 23.419 (40,9%) von insgesamt 57.327 Datenpunkten. Für diese Datenbank wurde kein Ausbeutekriterium festgelegt, so daß alle veröffentlichten Reaktionen unabhängig von ihrer Ausbeute in diese Datenbank aufgenommen werden. Zum anderen übersteigt die Anzahl der Ausbeuten die Anzahl der Reaktionen. Diese Tatsache kann man damit erklären, daß für viele Reaktionen nicht nur das Hauptreaktionsprodukt angegeben wurde, sondern auch Nebenprodukte, die in geringer Ausbeute gebildet werden. Auch die rund 1.000 Fälle von 0% sind darauf zurückzuführen, daß neben dem Hauptprodukt auch meist ein isomeres Nebenprodukt angegeben wurde, das allerdings nicht gebildet wird.

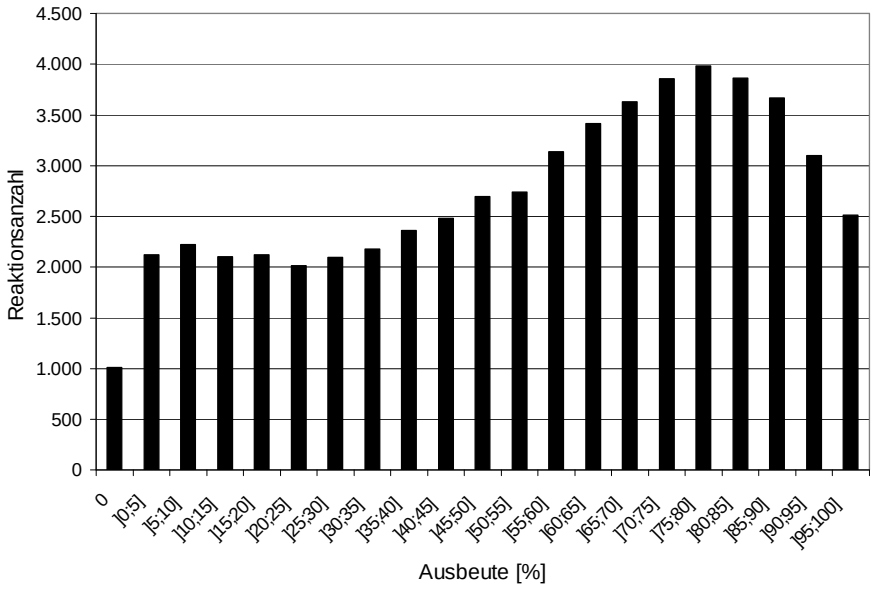


Abb. 2-9: Verteilung der Reaktionsausbeuten in der ChemInform RX-Datenbank des Jahres 1994.

Bei der Auswahl der Zeitschriften, deren Beiträge in Abstractform in ChemInform aufgenommen werden, liegt der Schwerpunkt auf der Synthese und den präparativen Methoden in der organischen und anorganischen Chemie. Das spiegelt sich auch in der folgenden Tabelle wider, in der die zehn Zeitschriften aufgeführt sind, die am häufigsten in der ChemInform RX-Datenbank genannt werden.

Rang	Name der Zeitschrift	Literaturstellen	Literaturst. [%]
1	Tetrahedron Lett.	57.310	14,2
2	J. Org. Chem.	40.097	9,9
3	Tetrahedron	29.798	7,4
4	J. Chem. Soc., Perkin Trans. 1	15.371	3,8
5	Synth. Commun.	14.220	3,5
6	Synthesis	13.180	3,3
7	J. Chem. Soc., Chem. Commun.	12.922	3,2
8	J. Heterocycl. Chem.	12.413	3,1
9	Synlett	11.061	2,7
10	Heterocycles	11.042	2,7

Tab. 2-3: Zitierte Zeitschriften in den ChemInform RX-Reaktionsdatenbanken 92-97, sortiert nach der Häufigkeit ihres Auftretens.

Zusammenfassend kann man festhalten, daß die Jahrgänge der ChemInform RX-Reaktionsdatenbanken Zugriff auf zahlreiche Reaktionsbeispiele aus einem engen Zeitintervall der 90er Jahre bieten.

2.2.3 SPORE Reaktionsdatenbank

Die *„**S**olid-**P**hase **O**rganic **Re**actions“* Reaktionsdatenbank[13] wurde aus Reaktionen zusammengetragen, die sich ausschließlich mit der Synthese kleiner organischer Moleküle an fester Phase beschäftigen. Diese Reaktionsdatenbank wurde speziell für die Informationsbedürfnisse von Chemikern entwickelt, die im Bereich der Organischen Synthese an fester Phase tätig sind. Die Datenbank eignet sich somit auch besonders zur Entwicklung kombinatorischer Bibliotheken an festen Trägern.

Die SPORE Datenbank wird jährlich aktualisiert, wobei ca. 1.500 neue Reaktionen hinzukommen. In dieser Arbeit wird die Inhouse-Datenbank mit der Versionsnummer 99.2 mit insgesamt 6.502 Reaktionen verwendet.

Die Datenbank enthält seit der ersten Festphasenreaktion von Merrifield[14] im Jahre 1963 bis zur Gegenwart alle Reaktionen, die folgende Kriterien erfüllen:

- Kupplungsreaktionen an die feste Phase und Ablösereaktionen von der festen Phase
- Einführen von Linkern und Spacer
- Einführen und Entfernen von Schutzgruppen an funktionellen Gruppen
- Reaktionen, die funktionelle Gruppen an fester Phase umwandeln.

Diese Reaktionen wurden aus rund 75 Zeitschriften zusammengetragen, wobei die in Tabelle 2-4 genannten 10 Zeitschriften und Patentschriften in der SPORE Datenbank der Version 99.2 am häufigsten genannt sind.

Rang	Name der Zeitschrift	Literaturstellen	Literaturst. [%]
1	Tetrahedron. Lett.	3.059	37,3
2	J. Org. Chem.	1.053	12,8
3	J. Am. Chem. Soc.	656	8,0
4	Patent, US	589	7,2
5	Patent, PCT Int. Appl.	390	4,8
6	Tetrahedron	240	2,9
7	J. Chem. Soc., Chem. Commun.	229	2,8
8	Mol. Diversity	211	2,6
9	Synlett	200	2,4
10	Bioorg. Med. Chem. Lett.	194	2,4

Tab. 2-4: Die zehn häufigsten zitierten Zeitschriften und Patentschriften in der SPORE Reaktionsdatenbank.

Seit der ersten Veröffentlichung von Merrifield im Jahre 1963 im *Journal of American Chemical Society*[14] nahm die Anzahl der Veröffentlichungen laufend zu, wobei in den letzten Jahren ein geradezu explosionsartiges Anwachsen zu verzeichnen ist. Dies spiegelt sich auch in dem Histogramm über die in der SPORE Reaktionsdatenbank enthaltenen Literaturjahrgänge wieder (siehe Abbildung 2-10). Nach einem anfänglichen Zuwachs der Literatur-

stellen bis ins Jahr 1979, ebbte die Begeisterung für Festphasenreaktionen in den folgenden 10 Jahren etwas ab, bevor ab dem Jahre 1994 eine geradezu explosionsartige Zunahme an Veröffentlichungen einsetzte. Diese Zunahme hängt vor allem mit dem Aufkommen der kombinatorischen Chemie anfangs der 90er Jahre zusammen, die zunächst auf Festphasenreaktionen ausgerichtet war.

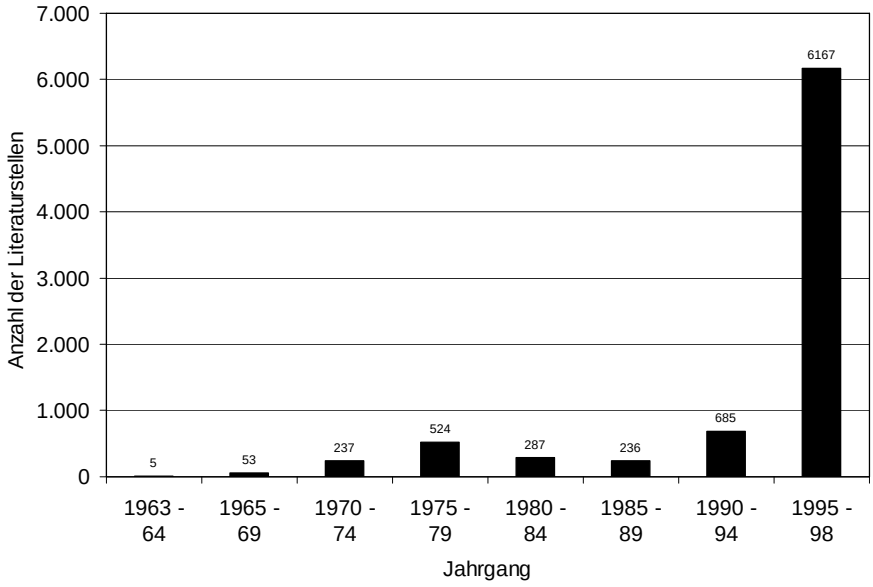


Abb. 2-10: Histogramm zur Anzahl der Literaturstellen in der SPORE Reaktionsdatenbank mit 5-Jahres-Intervallen.

In der SPORE Reaktionsdatenbank findet man insgesamt 2.391 Angaben zur Ausbeute, das entspricht 36,8% bei einer Gesamtreaktionsanzahl von 6.502 Reaktionen. Das in Abbildung 2-11 wiedergegebene Histogramm zur Verteilung der Reaktionsausbeuten zeigt einen generellen Anstieg der Reaktionsanzahl zu hohen Ausbeuten. 631 Ausbeutewerte (26,4%) liegen in dieser Datenbank unter 50% und die meisten Reaktionen weisen eine Ausbeute zwischen 95% und 100% auf. Auch in dieser Reaktionsdatenbank sind Reaktionen mit einer Ausbeute von 0% abgespeichert worden. Wie schon bei der ChemInform RX-Reaktionsdatenbank erklärt, wird diese Ausbeute meist für ein isomeres Nebenprodukt angegeben, das nicht gebildet wird.

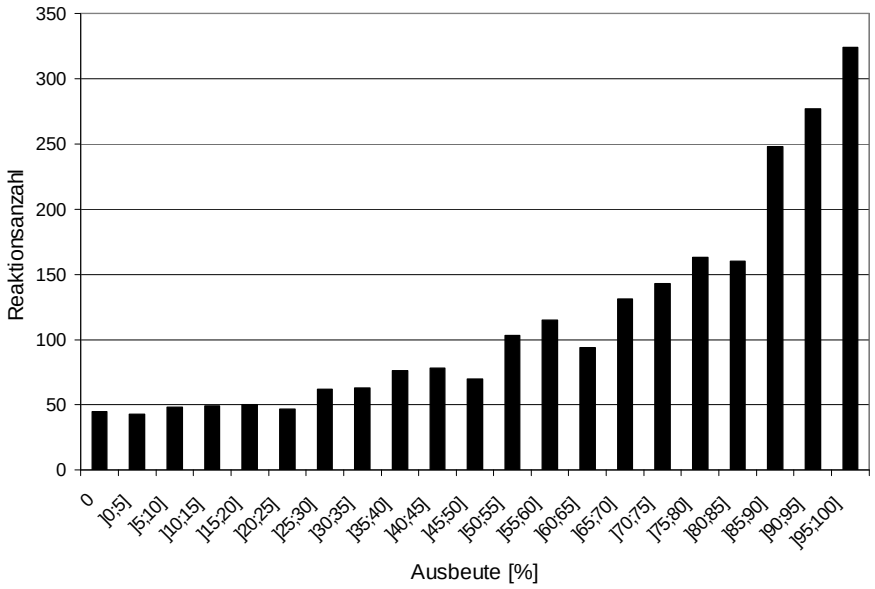


Abb. 2-11: Verteilung der Reaktionsausbeuten in der SPORE Reaktionsdatenbank.

Die SPORE Reaktionsdatenbank ist eine auf Festphasenreaktionen spezialisierte Datenbank, die alle Reaktionen an fester Phase ab dem Jahr 1963 bis zur Gegenwart enthält.

2.2.4 CrossFire*plus*Reactions

Beilstein Informationssysteme GmbH bietet mit CrossFire nach dem *Chemical Abstracts Service* das derzeit umfassendste elektronische Informationssystem der Chemie an. Die darin enthaltene Information kann man drei Bereichen zuteilen, die untereinander im Informationsaustausch stehen. Dazu gehört erstens eine Substanzdatenbank, die 7,5 Millionen organische Strukturen enthält, welche bis ins Jahr 1771 zurückreichen, sowie annähernd 1 Million anorganische bzw. metallorganische Verbindungen. Zu den Substanzen sind weiterhin chemische, physikalische und biologische Eigenschaften sowie präparative Methoden abrufbar. Insgesamt sind 20 Millionen solcher Einträge aufgenommen worden. Zweitens wurden in einer Reaktionsdatenbank bisher ca. 10 Millionen Reaktionseinträge aufgenommen, wobei davon ca. 5 Millionen Reaktionen über Reaktionssubstrukturen zugänglich sind. Der dritte Eckpfeiler dieses Informationssystems bildet eine Literaturdatenbank. Innerhalb dieser hochwertigen Datenbank kann man über insgesamt 100 Millionen Hyperlinks auf die Informationen zugreifen[15]. Da mit dem hier am Computer-Chemie-Centrum zur Verfügung stehenden Client-Server-System der Export von Reaktionen im MDL SDfile-Format nicht zur Verfügung steht, konnte diese Datenbank – mit einer Ausnahme – nicht für Untersuchungen herangezogen werden. Im Kapitel 6.4 wird auf einen Datensatz aus CrossFire*plus*Reactions zurückgegriffen, der allerdings nicht stellvertretend für die gesamte Reaktionsdatenbank herangezogen werden

kann. Somit kann an dieser Stelle auch keine Charakterisierung dieser Datenbank erfolgen. Für diese Arbeit wurde auf die Version mit der Bezeichnung BS0001PR zugegriffen.

2.2.5 Beurteilung der Reaktionsdatenbanken

Die in den Datenbanken gespeicherten Reaktionen wurden nicht für Computerchemiker, die die Datenbanken statistisch erschließen wollen, eingegeben, sondern in erster Linie für synthetisch arbeitende Chemiker. Aus diesem Grund wurde bei der Erfassung der Reaktionen nicht viel Wert auf eine korrekte Stöchiometrie und Massenbilanz gelegt, da der geübte Chemiker die fehlende Information gedanklich schnell ergänzen kann. In den Datenbanken ist also viel implizite Information enthalten, die häufig eine informationstechnische Verarbeitung erschwert.

Anhand der Massenbilanz kann man schnell feststellen, ob Reaktionen korrekt abgespeichert wurden. Nur wenn die Molekülmassen des Eduktensembles und des Produktensembles identisch sind, kann man davon ausgehen, daß die Reaktion vollständig codiert abgespeichert wurde. Tabelle 2-5 zeigt das Ergebnis einer solchen Untersuchung für verschiedene Reaktionsdatenbanken. Es ist neben dem Absolutwert auch der Prozentsatz an Reaktionen mit korrekter Massenbilanz für die jeweilige Reaktionsdatenbank aufgeführt.

Datenbank	Gesamtreaktions-anzahl	Reaktionen mit korrekter Massenbilanz	Anteil der Reaktionen mit korrekter Massenbilanz in %
Theilheimer	46.785	4.542	9,7
CIRX93	67.740	5.285	7,8
CIRX94	56.236	4.197	7,5
CIRX95	64.187	4.520	7,0
CIRX96	70.271	4.953	7,0
CIRX97	70.061	5.174	7,4
SPORE	6.502	315	4,8
Beilstein CrossFire	5.139.668	478.000*	9,2*

Tab. 2-5: Aufstellung der Massenbilanzen für verschiedene Reaktionsdatenbanken.
* Der Prozentsatz für die Beilstein CrossFire Datenbank wurde mit einer Stichprobe ermittelt, die Anzahl der Reaktionen mit korrekter Massenbilanz kann nur abgeschätzt angegeben werden.

Bei allen untersuchten Reaktionsdatenbanken liegt also der Prozentsatz an Reaktionen mit korrekter Massenbilanz unter 10%. Bei über 90% der aus Reaktionsdatenbanken entnommenen Reaktionen müßte man zunächst die abgespeicherte Reaktionsgleichung ergänzen. Entweder sind die Koeffizienten der Edukte bzw. Produkte in der Reaktionsgleichung falsch oder es fehlen Edukt- bzw. Produktmoleküle. Meist handelt es sich dabei um einfache Moleküle, wie Wasser, Ammoniak, Hydrazin etc., die nicht mit abgespeichert wurden.

Neben der stöchiometrisch meist falschen Aufnahme der Reaktionen in die Datenbanken zeigen diese weitere Mängel, die eine datenverarbeitende, statistische Analyse erschweren.

Eine abgespeicherte Reaktionsgleichung kann beispielsweise aus mehreren Einzelreaktionen bestehen. Eine solche aus Folgereaktionen zusammengesetzte Reaktionsgleichung enthält die Edukte wie bei einer Einzelreaktion, aber Produkte, die man erst nach einer Reihe von Einzelreaktionen erhalten würde. Fast immer werden zusätzliche Informationen wie Reaktionsbedingungen, Ausbeuten etc. verallgemeinert für die gesamte Reaktionssequenz angegeben, so daß eine Zuordnung auf einzelne Reaktionsschritte nicht möglich ist.

Mehrere miteinander konkurrierende Reaktionen werden ebenfalls meist nur mit einer einzigen Reaktionsgleichung erfaßt. In diesen sogenannten Parallelreaktionen reagieren Edukte eventuell über unterschiedliche Reaktionsmechanismen zu verschiedenen Produkten. Meistens werden auch hier für die Einzelreaktionen keine getrennten Angaben zu den Reaktionsbedingungen und Ausbeuten gemacht. Eine Aufspaltung von Folge- und Parallelreaktionen in Einzelreaktionen wäre daher wünschenswert. Da die Unterscheidung zwischen Folge- und Parallelreaktionen für datenverarbeitende Systeme keine triviale Aufgabe ist, sollte man in Zukunft Reaktionsdatenbanken nur noch mit Einzelreaktionen aufbauen und Parallel- und Folgereaktionen in einzelnen Reaktionen zerlegt abspeichern. Bei der jüngsten Konzeption einer Reaktionsdatenbank über biochemische Prozesse war man sich dieser Probleme bewußt und hat von Anfang an auf eine korrekte Eingabe der Reaktionsgleichungen geachtet[16].

Einen weiteren Mangel bei den Reaktionsdatenbanken findet man bei der Angabe des Reaktionszentrums. Das Reaktionszentrum einer Reaktion kann beispielsweise unvollständig angegeben sein. Da nur spezielle Wasserstoffatome abgespeichert sind, fehlen nahezu alle Wasserstoffatome im Reaktionszentrum, falls Bindungen zu diesen gebrochen oder geknüpft werden. Für diese Atome sind, wie manchmal auch bei anderen Atomen, keine Atom-Atom-Mapping-Nummern abgespeichert, so daß keine leichte Zuordnung einzelner Atome zwischen den Edukt- und Produktmolekülen gegeben ist.

Alle diese Codierungsmängel könnten mit einem Programmsystem ausgeglichen werden, das zum einen unvollständige Reaktionen komplettieren kann. Hierzu wurde von Bargon et al. ein Programm entwickelt, das Reaktionen vervollständigen und das jeweilige Reaktionszentrum exakt bestimmen kann[17]. Da im Arbeitskreis von Gasteiger keine Programme zur Komplettierung bzw. Separation von Reaktionsgleichungen zur Verfügung stehen, dies andererseits auch nicht Gegenstand dieser Arbeit war, wurden die Reaktionsgleichungen ohne weitere Korrekturen den Datenbanken entnommen. Auf die Konsequenzen wird in den entsprechenden Kapiteln näher eingegangen.

2.3 Methoden zum Identitätsvergleich von Reaktionsdatenbanken

Angesichts der inzwischen zahlreich erhältlichen Reaktionsdatenbanken mit Millionen von Reaktionen wird ein Vergleich der Reaktionsdatenbanken immer wünschenswerter. Dabei kann man prinzipiell zwischen einem Identitätsvergleich und einem Ähnlichkeitsvergleich einzelner Reaktionseinträge in verschiedenen Datenbanken unterscheiden.

Eine Möglichkeit zum Identitätsvergleich bietet die Überprüfung der Trefferliste einer exemplarischen Reaktionsanfrage an verschiedene Reaktionsdatenbanken[18]. Eine andere Möglichkeit, die von der Auswahl der Reaktionsanfrage unabhängig ist, ist der systematische Vergleich aller Reaktionseinträge verschiedener Reaktionsdatenbanken miteinander. Dabei verwendet man die für jede Reaktion angegebene Literaturstelle zum Identitätsvergleich. Identische Reaktionseinträge liegen dann vor, wenn die in eine einheitliche Form überführte und eventuell korrigierte Literaturstelle zweier Reaktionen aus verschiedenen Datenbanken gleich sind[19]. In einer Untersuchung von Hendrickson et al. zum Vergleich von Datenbanken wurde eine Methode eingesetzt, die sowohl einen Identitätsvergleich als auch einen Ähnlichkeitsvergleich gestattet[20]. Jede Reaktion wird dabei mit Hilfe des COGNOS Programmsystems in einen aus fünf Elementen bestehenden Zahlencode transformiert. Vier dieser fünf Elemente codieren das Reaktionszentrum auf einem unterschiedlichen Abstraktionsniveau (Reaktionsklasse, Reaktionstyp, Umgebung des Reaktionszentrums, abgespaltene oder hinzugefügte Gruppen). Das fünfte Element beschreibt den Teil des Moleküls, der nicht zum Reaktionszentrum zählt, mit einer Eigenschaftsliste. Je mehr dieser fünf Elemente paarweise für je zwei Reaktionen übereinstimmen, desto ähnlicher sind diese Reaktionen zueinander. Selbst wenn alle fünf Elemente übereinstimmen, müssen die Edukte bzw. Produkte in beiden untersuchten Reaktionen nicht notwendig identisch sein. Daher nahm auch Hendrickson et al. für den Identitätsvergleich eine einheitliche und eventuell berichtigte Form der originalen Literaturstellen hinzu.

Der Vergleich verschiedener Reaktionsdatenbanken auf identische Reaktionseinträge ist beispielsweise hilfreich, wenn man möglichst viele bisher publizierte und elektronisch erfaßte Reaktionen zu einer Reaktionsanfrage berücksichtigen will. In diesem Fall sollte man in möglichst vielen Reaktionsdatenbanken, die nur sehr wenige identische Reaktionseinträge gemeinsam haben, die Suchanfrage stellen. Auch beim Kauf verschiedener Reaktionsdatenbanken möchte man natürlich keine Datenbanken erwerben, die untereinander viele identische Reaktionseinträge aufweisen.

Neben diesen Einsatzmöglichkeiten für einen Identitätsvergleich von Reaktionsdatenbanken gibt es auch viele Anwendungsgebiete für einen Ähnlichkeitsvergleich. Reaktionen eines Reaktionstyps zeichnen sich häufig durch zueinander ähnliche Reaktionszentren aus, nicht durch identische. Beispielsweise kann eine nucleophile Substitutionsreaktion an einer Kohlenstoff-Sauerstoff- oder einer Kohlenstoff-Chlor-Bindung erfolgen. Auch beim Aufbau einer

kombinatorischen Bibliothek möchte man gleichzeitig eine Vielzahl von Molekülen synthetisieren, die nicht nach identischen, sondern nach ähnlichen Reaktionen ablaufen. Daher wurden bereits Klassifizierungsverfahren entwickelt, die einen Ähnlichkeitsvergleich von Reaktionen ermöglichen. Ein solches Verfahren stammt von InfoChem und wird im nächsten Kapitel vorgestellt.

2.4 Klassifizierungsverfahren von InfoChem

Einige Anbieter von Reaktionsdatenbanken haben bereits vor Jahren die Bedeutung von Klassifizierungsverfahren bei der Bewältigung der Informationsflut erkannt, und bieten daher immer häufiger Datenbanken an, die Reaktionen mit Klassifizierungscodes beinhalten. Während vor rund einem halben Jahr nur wenige Reaktionsdatenbanken von MDL, wie ChemInform RX oder SPORE, mit Klassifizierungseinträgen für die einzelnen Reaktionen angeboten wurden, so bietet MDL inzwischen für alle von ihr vertriebenen Reaktionsdatenbanken diese Option an. Dies unterstreicht die zunehmende Bedeutung, die der Reaktionsklassifizierung auch von Seiten der Datenbankanbieter beigemessen wird. Das von MDL eingesetzte Klassifizierungsverfahren, namens CLASSIFY, wurde von der Firma InfoChem lizensiert.

CLASSIFY bestimmt zunächst für jede Reaktion die Atom-Atom-Mapping-Nummern und das Reaktionszentrum, das zur Klassifizierung benötigt wird. Bei der anschließenden Klassifizierung werden für jede Reaktion drei Klassifizierungscodes berechnet. Jeder Klassifizierungscode beschreibt die strukturelle Umgebung des Reaktionszentrums, wobei unterschiedlich viele Sphären berücksichtigt werden. Der sogenannte *Broad Classification Code* repräsentiert nur die Atome und Bindungen des Reaktionszentrums, während in dem zweiten Code, *Medium Classification Code* genannt, auch die erste Nachbarschaftssphäre um das Reaktionszentrum eingeht. Bei der sogenannten *Narrow Classification* werden schließlich alle Atome und Bindungen zur Berechnung eines Klassifizierungscodes herangezogen, die maximal 2 Bindungen vom Reaktionszentrum entfernt sind.

Anhand der in Abbildung 2-12 dargestellten Reaktionen, für die jeweils drei Klassifizierungscodes angegeben sind, soll diese Klassifizierungsmethode näher erklärt werden. Der erste Klassifizierungscode beschreibt nur das Reaktionszentrum, also die Atome und Bindungen, die am Bindungsumordnungsprozeß beteiligt sind. Das Reaktionsbeispiel 1 unterscheidet sich deshalb in allen drei Klassifizierungscodes von allen anderen Beispielen, weil beim ersten Beispiel eine C-Br- anstatt einer C-Cl-Bindung geknüpft wird. Da die ausgewählten Beispiele 2 bis 5 alle dasselbe Reaktionszentrum besitzen, sind die *Broad Classification Codes* alle gleich. Diese vier Reaktionen besitzen auch alle dieselbe erste Sphäre um das Reaktionszentrum, so daß auch der zweite Klassifizierungscode bei den Reaktionen 2 bis 5 identisch ist. Erst der dritte Code vermag zwischen den Reaktionen zu unterscheiden.

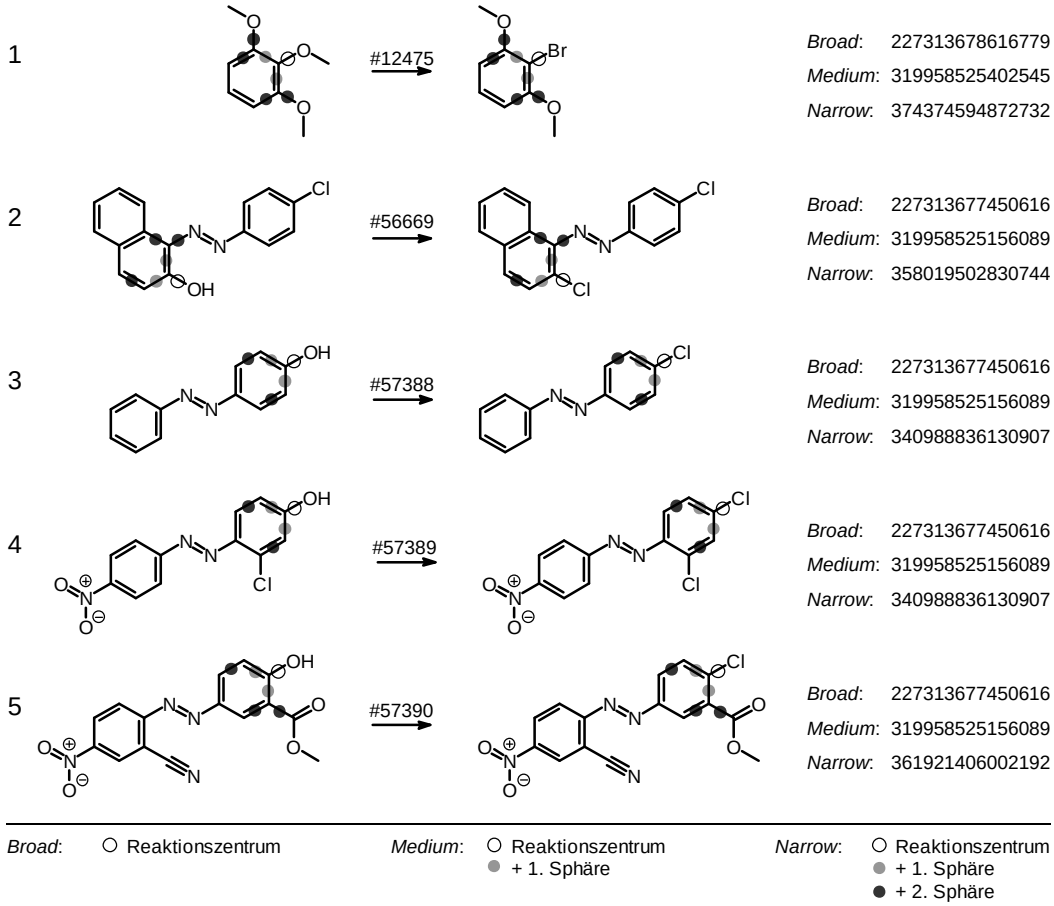


Abb. 2-12: Fünf ausgewählte Reaktionsbeispiele aus der ChemInform RX-Reaktionsdatenbank (Version 1992) mit jeweils drei Klassifizierungscodes, die mit CLASSIFY ermittelt wurden.

Ein Nachteil dieser Klassifizierungsmethode wird beim Vergleich der dritten und vierten Reaktion in Abbildung 2-12 offensichtlich. Diese beiden Reaktionen stimmen in allen drei Klassifizierungscodes überein, obwohl die aromatischen Systeme, an denen die Reaktion abläuft, unterschiedliche elektronische Eigenschaften aufweisen und somit den Reaktionsverlauf maßgeblich beeinflussen können. Da das zweite Chloratom in meta-Position, d.h. in der vierten Nachbarschaftssphäre, zur neu geknüpften Bindung steht, kann dieses Atom mit den drei Klassifizierungscodes nicht mehr erfaßt werden. Ein weiterer Nachteil dieser Klassifizierungsmethode liegt in den Zahlenwerten der berechneten Klassifizierungscodes. Diese Zahlenwerte lassen keinen Rückschluß auf die Ähnlichkeit zweier Reaktionen zu. Die vierte und fünfte Reaktion haben zwar beide ein elektronenarmes aromatisches System an dem die Kohlenstoff-Chlor-Bindung neu geknüpft wird, aber völlig verschiedene *Narrow Classification Codes*. Ein dritter Nachteil ist in der Codierung des gesamten Reaktionszentrums zu sehen. Reaktionen, bei denen entweder die Ausgangsverbindungen oder die Produkte zunächst unbekannt sind, verfügen noch nicht über ein vollständiges Reaktionszentrum. Daher können diese nicht mit den klassifizierten Reaktionen in den Datenbanken verglichen werden.

Obwohl bereits Methoden für einen Ähnlichkeitsvergleich von Reaktionsdatenbanken oder Reaktionsdatensätzen existieren, sollte im Rahmen dieser Arbeit ein weiteres Verfahren entwickelt werden, das die oben genannten Nachteile nicht zeigt. Im Gegensatz zu dem Klassifizierungsverfahren von InfoChem beruht das entwickelte Klassifizierungsverfahren nicht auf einer Charakterisierung des Reaktionszentrums aufgrund von Atom- und Bindungstypen, sondern auf physikochemischen Effekten von Atomen bzw. Bindungen, die dem Reaktionszentrum angehören. Dieses Verfahren wird ausführlich in Kapitel 3 vorgestellt.

2.5 Neuronale Netze

2.5.1 Biologische Grundlagen

Nach der Neuronentheorie, die gegen Ende des letzten Jahrhunderts entwickelt wurde, geht die Gehirntätigkeit auf die Kommunikation der voneinander elektrisch getrennten Nervenzellen, auch Neuronen genannt, zurück[21]. Die Nervenzellen im menschlichen Gehirn kann man in Stern- oder Gliazellen, die hauptsächlich die Stoffwechselversorgung sichern, und den Pyramidenzellen, die die elektrischen Impulse verarbeiten, unterteilen. Pyramidenzellen besitzen bis zu 12 kleine, astartige Auswüchse, die sogenannten Dendriten, welche die elektrischen Impulse, die sie erhalten, an den eigentlichen Zellkörper, das sogenannte Soma, weiterleiten (siehe Abbildung 2-13a). Wenn die Erregung, d.h. die elektrische Spannung, einen bestimmten Grenzwert überschreitet, so pflanzt sich ausgehend vom Zellkörper eine Änderung des Membranpotentials auf dem Axon bis in die entferntesten Verzweigungen fort. In den terminalen Axon-Endigungen überspringt der elektrische Impuls mittels Neurotransmitter, wie Acetylcholin, den synaptischen Spalt und wird von benachbarten Dendriten oder Muskelzellen erneut aufgenommen und weitergeleitet[22]. Die Weiterleitung des elektrischen Impulses innerhalb einer Nervenzelle beruht auf einer Änderung des Membranpoten-tials, welches wiederum durch eine rasche und kurzzeitige Änderung der Membranleitfähigkeit vor allem für Na^+- und K^+-Ionen verursacht wird. Nach Einwirkung eines überschwel-ligen Reizes wird in der Depolarisationsphase durch einen schnellen, kurzzeitigen Na^+-Ein-strom in die Zelle das Membranpotential von ca. –60 mV auf ca. +30 mV erhöht. Wegen des Anstiegs des Membranpotentials öffnen sich danach spannungsabhängige K^+-Kanäle, so daß K^+-Ionen in den Extrazellularraum ausströmen können. Das Membranpotential wird dadurch wieder negativ und gleicht sich in der Repolarisationsphase wieder dem Ruhepotential an[23]. Nach rund 3 ms ist die Nervenzelle von neuem erregbar. Diese als Aktionspotentiale bezeichneten periodischen Entladungen haben unabhängig von der Reizstromstärke immer dieselbe Stärke und sind somit von Natur aus binär, da sich das Aktionspotential entweder vollständig oder gar nicht ausbildet (Alles-oder-Nichts-Gesetz).

Die enorme Leistungsfähigkeit des menschlichen Gehirns beruht nicht – wie bei Computern – auf einer enormen Verarbeitungsgeschwindigkeit, sondern auf der Komplexität der Neuronenverbindungen. Jede Nervenzelle kann mit bis zu 100.000 Nachbarzellen eine synaptische Verbindung eingehen, so daß bei rund 10–100 Milliarden Nervenzellen insgesamt ca. 10^{15} synaptische Kontakte geknüpft werden können. Diese synaptischen Kontakte können in der Wirksamkeit durch biochemische Mechanismen moduliert werden. Zu diesen Mechanismen gehört beispielsweise die sogenannte Langzeitpotenzierung, bei der infolge des gleichzeitigen Eintreffens zweier Signale an einem Neuron die Synapsengewichte angepaßt werden, d.h. die Verbindung zwischen zwei Neuronen gestärkt wird. Diese Veränderbarkeit der Neuronenstärke bildet die Basis für die Lernfähigkeit des menschlichen Gehirns.

Einigen Bereichen im menschlichen Gehirn kann man bestimmte Funktionen zuordnen, wie beispielsweise der motorische, somato-sensorische und der visuelle Cortex. Beim somato-sensorischen Cortex werden die in einer bestimmten Körperregion existierenden Nervenzellen letztlich in einer bestimmten Gehirnregion abgebildet. Auffallend ist hierbei, daß die Größe der Gehirnregion mit der Anzahl der Neuronen korreliert: Organe mit sehr vielen Nervenzellen pro Quadratzentimeter, wie beispielsweise die Hand, nehmen auch auf dem somato-sensorischen Cortex einen viel größeren Platz ein als große Organe mit nur wenigen Nervenzellen pro Quadratzentimeter, wie beispielsweise das Bein. Als Folge davon liegt im Gehirn für jeden Sinneseindruck eine verzerrte Abbildung des menschlichen Körpers vor. In Abbildung 2-13b ist die Projektion des somato-sensorischen Cortex auf die Körperregionen dargestellt.

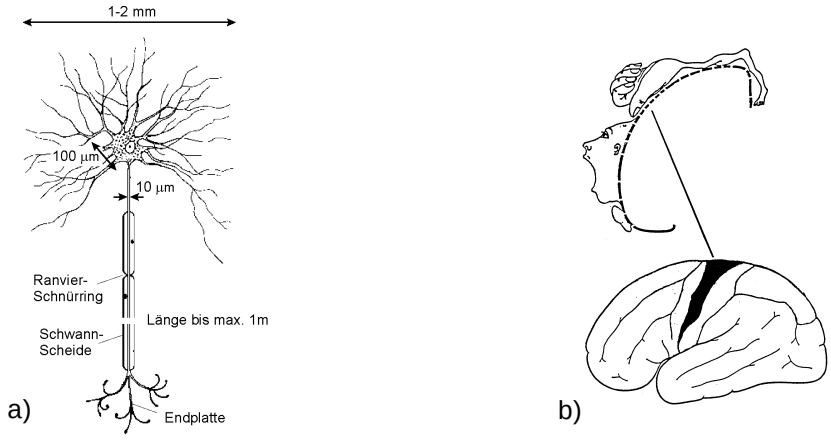


Abb. 2-13: a) Eine biologische Nervenzelle mit wichtigen Bestandteilen;
b) Projektion der Körperregionen auf den somato-sensorischen Cortex. Die Teile des Homunculus sind proportional zur Größe der sie repräsentierenden Gehirnrinde gezeichnet.

2.5.2 Grundlegende Komponenten neuronaler Netze

Künstliche neuronale Netze (KNN, engl. *artificial neural networks* ANN), oder kurz neuronale Netze, sind informationsverarbeitende Systeme, die auf neurobiologischen Modellen basieren. Sie bestehen aus einer Vielzahl unabhängiger, einfacher Einheiten, sogenannte Neuronen, die in definierter Weise miteinander verbunden sind und über gewichtete Verbindungen Informationen austauschen[24].

Neuronale Netzmodelle werden durch drei grundlegende Komponenten beschrieben, die im folgenden näher erläutert werden:

- Aufbau der Neuronen
- Topologie des Netzes
- Lernregel bzw. Propagierungsfunktion

Da in der vorliegenden Arbeit ausschließlich neuronale Netze nach Kohonen verwendet werden, wird im folgenden detaillierter auf dieses Netzmodell eingegangen.

2.5.2.1 Aufbau der Neuronen

Die Prozeßeinheiten eines neuronalen Netzes setzen sich wiederum aus einer Verbindungs-, einer Eingangs-, einer Aktivierungs- und einer Ausgangsfunktion zusammen. In Abbildung 2-14 sind diese Funktionen dargestellt.

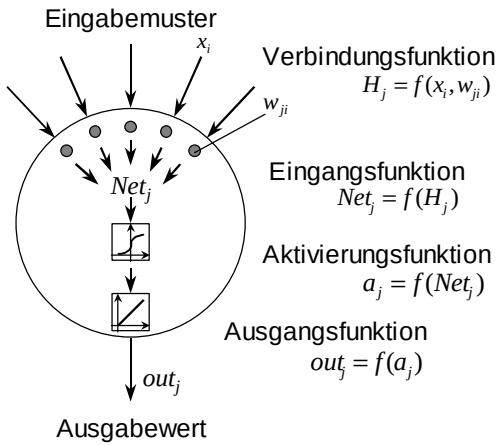


Abb. 2-14: Die Prozeßeinheiten eines neuronalen Netzes.

Signale, die ein Neuron über mehrere eingehende Verbindungen aufnimmt, werden durch die Verbindungsfunktion H_j gewichtet. Die dabei verwendeten Gewichte w_{ji}, die ein Maß für die Kopplungsstärke darstellen, werden in einem Lernprozeß ermittelt. Die Eingangsfunktion Net_j faßt all diese gewichteten Eingänge zu einem einzigen skalaren Wert – auch als Netzaktivität bezeichnet – zusammen. Meistens wird über alle Produkte aus je einem Gewicht w_{ji}

und einem Eingabewert x_i summiert. Mit Hilfe der Aktivierungsfunktion a_j wird aus diesem Wert und dem gegenwärtigen Zustand dieses Neurons ein neuer Aktivierungszustand – in Anlehnung an biologische Nervenzellen auch als Anregung bezeichnet – ermittelt. Dieser wird mittels der Ausgangsfunktion des Neurons out_j an verbundene Neuronen weitergegeben, wo sich der beschriebene Vorgang wiederholt.

Als Aktivierungsfunktion kommen lineare Schwellenwertfunktionen, semilineare Funktionen, lineare Funktionen und sigmoide Funktionen zum Einsatz (siehe Abbildung 2-15), während als Ausgangsfunktion meist eine lineare Funktion eingesetzt wird.

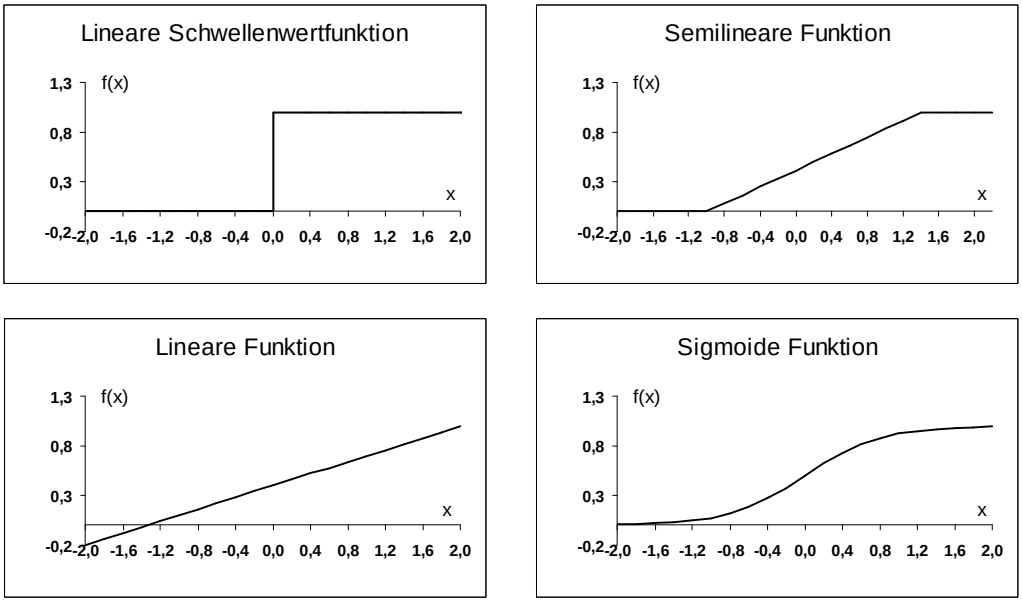


Abb. 2-15: Graphische Darstellung möglicher Aktivierungsfunktionen.

2.5.2.2 Topologie des Netzes

Die einzelnen Neuronen bilden als starre Struktur ein Netzwerk aus, wobei während der Lernphase eines neuronalen Netzes keine Strukturänderung stattfindet, d.h. die in biologischen Systemen stattfindenden dynamischen Prozesse, wie das „Absterben“ einzelner Neuronen bzw. das „Zusammenwachsen“ neuer Nervenverbindungen, werden in den künstlichen Analoga nicht berücksichtigt.

In der Regel wird eine geschichtete Netzwerkarchitektur verwendet. In diesem Fall nimmt eine Eingabeschicht (engl. *input layer*) die Eingabedaten entgegen, die anschließend in einer verborgenen Schicht oder mehreren Schichten (engl. *hidden layer*) prozessiert werden, bis sie schließlich in der Ausgabeschicht (engl. *output layer*) wieder ausgegeben werden (siehe Abbildung 2-16). Je nach der Verknüpfungsart zwischen den Schichten unterscheidet man zwischen vorwärts gerichteten (engl. *feed forward*), lateralen und rückgekoppelten (engl. *feed back*) Netzen.

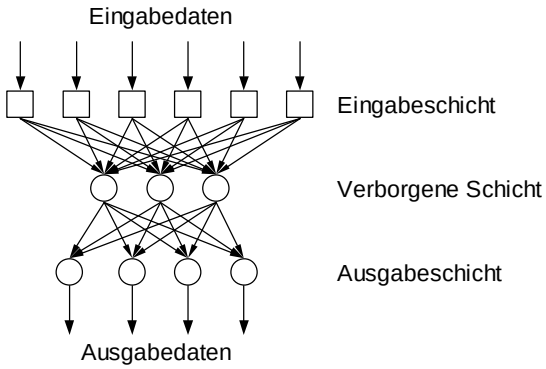


Abb. 2-16: Allgemeine Netzwerkarchitektur mit vorwärts gerichteter Verknüpfung und zwei Schichten (es werden nur die Schichten unterhalb der Eingabeschicht gezählt).

Je nach Topologie eines Kohonen-Netzes kann die Ausgabeschicht die Gestalt eines Rechtecks oder eines Torus annehmen. Die toroidale Anordnung hat gegenüber der rechteckigen Anordnung den Vorteil, daß alle Neuronen dieselben Nachbarschaftsbeziehungen aufweisen. Auch wenn man eine toroidale Anordnung verwendet, stellt man meist die Ausgabeschicht eines neuronalen Netzes als planares Gebilde dar: Durch zweifaches Aufschneiden eines toroidalen Netzwerkes gelangt man zu einem anschaulichen, zwei-dimensionalen Rechteck (siehe Abbildung 2-17).

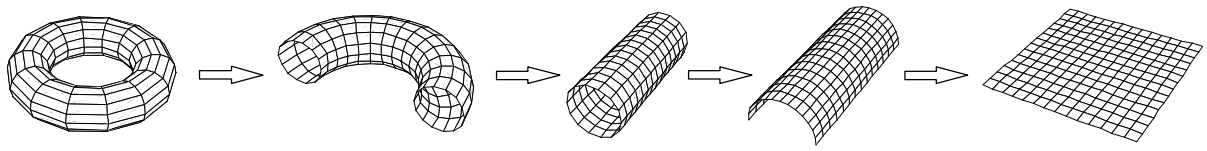


Abb. 2-17: Übergang eines toroidalen Netzes in eine quadratisch-planare Anordnung durch zweifaches Aufschneiden.

2.5.2.3 Lernregel

Wissen wird in neuronalen Netzen in Form von Gewichtsfaktoren gespeichert, indem während einer Trainingsphase die Gewichtsvektoren mittels Lernregel angepaßt werden. Die Lernregel ist also ein Algorithmus, mit dem das Netz aus einem vorgegebenen Eingabemuster ein Ausgabemuster erzeugt.

Bei den Lernverfahren kann man zwischen überwachten und unüberwachten Strategien unterscheiden. Beim überwachten Lernen, auch *supervised learning* genannt, werden dem neuronalen Netz im Trainingsprozeß sowohl die Eingabemuster als auch die gewünschten Ausgabemuster übergeben. Das Netz modifiziert dabei durch Anwendung einer Lernregel die Gewichte in der Weise, daß ein Fehlersignal minimiert wird. Das Fehlersignal berechnet das neuronale Netz aus der Differenz zwischen der erzeugten und der vorgegebenen Ausgabe. Ein solches Lernverfahren ist beispielsweise im Multilayer Perceptron mit Backpropagation-Lernregel und dem Counterpropagation-Netz verwirklicht[25].

Beim unüberwachten Lernen (engl. *unsupervised learning*), wie es beispielsweise im neuronalen Netz nach Kohonen eingesetzt wird, werden dem Netz nur die Eingabemuster präsentiert. Das Netz findet in dem Eingabemuster eigenständig die Ähnlichkeitsbeziehungen und paßt seine Gewichtsvektoren in der Weise an, daß ähnliche Eingabemuster entweder in gleichen Neuronen oder nah benachbarten Neuronen projiziert werden, während unähnliche Eingabemuster in weit voneinander entfernte Neuronen eingetragen werden.

2.5.2.4 Einsatzgebiete

Auch wenn anfangs künstliche neuronale Netze bei neurobiologischen Forschungen von Gehirnen eingesetzt wurden[26],[27], so spielt dieses Einsatzgebiet heutzutage nur noch eine untergeordnete Rolle[25]. Der Einsatzschwerpunkt verlagerte sich statt dessen in das Gebiet der „künstlichen Intelligenz“. Dort werden neuronale Netze bei der Bewältigung folgender industrieller Problemstellungen erfolgreich eingesetzt: Mustererkennung (Bild und Sprache), Mustervervollständigung und in der Steuer- und Regelungstechnik. Darüber hinaus werden neuronale Netze bei der Analyse komplexer Daten (Datenvorhersage), der Bestimmung der Ähnlichkeit zwischen Mustern oder Daten und der automatischen Klassifizierung eingesetzt.

Neuronale Netze werden seit 1988 auch im Bereich der Chemie eingesetzt[28]. Dort stieg in den vergangenen 13 Jahren die Zahl der Veröffentlichungen von anfangs 3 (1988) auf 927 (1997) exponentiell an. Einige Einsatzgebiete liegen beispielsweise in der Analytischen Chemie[29], der Korrelation von Struktur und Infrarotspektrum[30],[31] sowie in der Kontrolle chemischer Prozesse[32]. Außerdem werden neuronale Netze bei Untersuchungen zur Sekundärstruktur von Proteinen eingesetzt, bei QSAR-Untersuchungen[33], bei Untersuchungen zur chemischen Reaktivität[34], sowie zur Projektion des elektrostatischen Potentials eines Moleküls[35]. Im Umweltbereich werden neuronale Netze eingesetzt um physikochemische Eigenschaften vorherzusagen, wie beispielsweise den Siedepunkt[36] und die kritische Temperatur[37],[38], den Oktanol-Wasser-Verteilungskoeffizienten[39],[40], die Toxizität[41] und die Mutagenität[42],[43].

Einen Überblick über die vielfältigen Einsatzmöglichkeiten neuronaler Netze in der Chemie findet man in dem Übersichtsartikel von Zupan und Gasteiger[44].

2.5.3 Neuronale Netze nach Kohonen

2.5.3.1 Topologie und Trainingsprozeß

Das Konzept der selbstorganisierenden Karten, auch *self-organizing feature maps* (SOM) genannt, wurde von Teuvo Kohonen entwickelt[45],[46],[47]. Dieses neuronale Netz besitzt eine einschichtige Topologie mit einer Schicht aktiver Neuronen und gehört zu den unüberwacht lernenden Netzmodellen. Die charakteristische Eigenschaft der Kohonen-Netze ist die

Fähigkeit, einen mehrdimensionalen Informationsraum topologieerhaltend in eine zweidimensionale Kohonen-Karte abzubilden, d.h. Eingabemuster, die im mehrdimensionalen Raum benachbart sind, werden auch in benachbarte Bereiche der Kohonen-Karte projiziert.

Das Trainieren eines neuronalen Netzes läuft folgendermaßen ab: Zu Beginn der Trainingsphase sind die Gewichte der einzelnen Neuronen mit zufälligen Werten besetzt. Im Laufe des Lernprozesses werden die Eingabemuster dem Netz präsentiert, wobei für jedes Eingabemuster ein sogenanntes Gewinnerneuron mittels Euklidischer Distanzfunktion berechnet wird. Sei $\boldsymbol{W}_j$ der Gewichtsvektor (w_{j1}, w_{j2}, ... w_{jm}) und $\boldsymbol{X}_s$ das Eingabemuster (x_{s1}, x_{s2}, ... x_{sm}) sowie m die Dimension des Eingabe- bzw. Gewichtsvektors, so ermittelt man das Gewinnerneuron, c, mittels Gleichung 2-1.

$$out_c \leftarrow \min\left\{\sum_{i=1}^{m}\left(w_{ji} - x_{si}\right)^2\right\}$$

Gleichung 2-1: Berechnung des Gewinnerneurons.

Da immer nur ein einziges Neuron als Gewinnerneuron ermittelt wird, spricht man von *kompetitivem Lernen*.

Die Gewichte dieses Gewinnerneurons werden nun in der Weise geändert, daß sie dem Eingabemuster ähnlicher werden. Die zu diesem Gewinnerneuron benachbarten Neuronen werden ebenfalls abgeändert. Dies wird durch eine Distanz- oder Nachbarschaftsfunktion geregelt, die meist mit zunehmendem Abstand vom Gewinnerneuron eine lineare Abnahme der Gewichtsmodifizierung beschreibt, d.h. mit zunehmendem Abstand vom Gewinnerneuron wird die Änderung immer kleiner. Im quadratisch-planaren Nachbarschaftsgitter erfahren acht direkte Nachbarneuronen stärkere Änderungen als die 16 Nachbarn zweiten Grades usw. (siehe Abbildung 2-18). Im Falle einer rechteckigen Karte reduziert sich an den Rändern die Anzahl der Nachbarn.

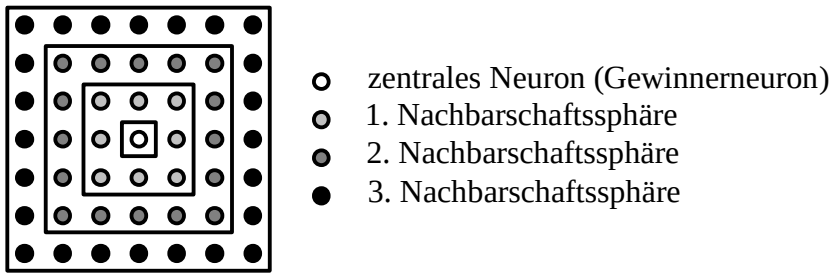


Abb. 2-18: Quadratisch-planares Nachbarschaftsgitter.

Um ein Übertrainieren des Kohonen-Netzes zu verhindern, erfolgt die Anpassung der Neuronen um so schwächer, je länger das Netz bereits trainiert wurde. Dies wird durch die sogenannte Lernrate h geregelt.

Gleichung 2-2 stellt den Anpassungsprozeß als Funktion der Zyklendurchläufe t dar: Jeder Gewichtsvektor eines Neurons $w_{ji}(t+1)$ wird aus dem Wert des Gewichtsvektors des zurück-

liegenden Zyklus $w_{ji}(t)$ berechnet, indem ein Produkt, das sich aus der Lernrate $h(t)$, einer Nachbarschaftsfunktion $f(t,d_{jc})$ und der Differenz aus dem Eingabewert und dem alten Wert zusammensetzt, addiert wird. Die Anpassung des Gewichtsvektors ist dabei um so größer, je kleiner die Distanz d_{jc} eines Neurons j zum Gewinnerneuron c ist.

$$w_{ji}(t+1) = w_{ji}(t) + h(t) \cdot f(t, d_{jc}) \cdot \left[x_i - w_{ji}(t)\right]$$

Gleichung 2-2: Die Anpassung der Neuronen ist abhängig von der Lernrate $h(t)$ und einer Nachbarschaftsfunktion $f(t,d_{jc})$.

Nachdem alle Eingabevektoren dem neuronalen Netz präsentiert wurden, werden zum Abschluß des Trainingsprozesses alle Eingabevektoren je einem Neuron des fertig trainierten Netzes unter Anwendung der Euklidischen Distanzfunktion zugeordnet. Neuronen, in welche kein einziger Eingabevektor projiziert wird, bleiben leer und werden *leere Neuronen* genannt.

2.5.3.2 Labeling

Nachdem alle Eingabevektoren je einem Neuron zugeordnet sind, kann man die resultierende Kohonen-Karte analysieren. Falls der Benutzer eines Kohonen-Netzes dieses zu Klassifizierungszwecken einsetzt, so kann er zunächst allerdings nur eine lokale Ähnlichkeit seiner Eingabemuster erkennen. Diejenigen Eingabemuster, die in dieselben Neuronen projiziert werden, gehören meistens der gleichen Klasse an. Im vorliegenden Fall weisen Reaktionen, die in denselben Neuronen projiziert werden, eine sehr hohe Ähnlichkeit auf und gehören meistens zum gleichen Reaktionstyp. Darüber hinaus können auch benachbarte Neuronen ebenfalls der gleichen Klasse angehören, d.h. Reaktionen in benachbarten Neuronen können eine hohe Ähnlichkeit aufweisen und ebenfalls demselben Reaktionstyp angehören. Der als Labeling bezeichnete Schritt weist nun jedem nicht-leeren Neuron eine Klasse zu und stellt somit sicher, daß man eine globale Ähnlichkeit der Eingabemuster erkennen kann. Erst durch dieses Einfärben der Neuronen wird deutlich, welche Bereiche der Kohonen-Karte zusammengehören und wo die Übergänge zwischen einzelnen Gebieten lokalisiert sind. Da die Eingabemuster im Gegensatz zum Counterpropagation-Netz beim Kohonen-Netz immer ohne Klassenzugehörigkeit eingesetzt werden, muß zum Einfärben auf andere Methoden zurückgegriffen werden:

- Klassenabgrenzung mittels Euklidischer Distanz: Die Euklidische Distanz ist zwischen Neuronen derselben Klasse kleiner als zwischen Neuronen verschiedener Klassen.
- intellektuelle Klassifizierung: Das Einfärben der Neuronen wird in diesem Fall vom Benutzer vorgenommen, der den Eingabemustern eine Klasse zuteilt.
- Verfahren nach Bayes: Bei diesem eigenständigen, probabilistischen Klassifizierungsverfahren, das auf dem Theorem von Bayes beruht, wird für jedes Objekt eines Datensat-

zes die Wahrscheinlichkeit berechnet, einer bestimmten Klasse anzugehören. Das Bayes-Theorem ist die Berechnung der a-posteriori-Wahrscheinlichkeit für eine Hypothese, d.h. die Wahrscheinlichkeit, daß ein Ereignis eintritt unter der Bedingung, daß ein anderes Ereignis schon eingetreten ist. In einigen Studien dieser Arbeit wurde dieses Verfahren zum Einfärben von trainierten Kohonen-Netzen erfolgversprechend angewandt.

Wenn man das trainierte Netz für Vorhersagezwecke einsetzt, so kann nach der Ermittlung eines Gewinnerneurons für das zu untersuchende Eingabemuster eine Klassenzugehörigkeit ausgegeben werden. Es empfiehlt sich in diesem Fall auch die leeren Neuronen einer Klasse zuzuordnen. Sonst würde man für ein Eingabemuster, das beispielsweise in ein leeres Neuron projiziert wird, welches zentral in einem Klassenbereich liegt, trotzdem keine Klassenzugehörigkeit finden. Im einfachsten Ansatz weist man dem leeren Neuron die Klasse zu, die am häufigsten in benachbarten Neuronen anzutreffen ist. Genauer wäre ein Verfahren, das auch der Euklidischen Distanz zwischen den Neuronen Rechnung trägt.

2.5.4 Diskussion

Zur Klassifizierung eines Datensatzes stehen viele verschiedene Methoden zu Verfügung. Man kann die Verfahren generell an der Art des Lernens in zwei Kategorien einteilen: Unüberwachtes und überwachtes Lernen. Zu den unüberwachten Klassifizierungsverfahren zählen beispielsweise die faktoriellen Methoden, wie die Hauptkomponentenanalyse (*principal component analysis*), die Clusteranalyse, graphische Verfahren sowie neuronale Netze nach Kohonen. Die Diskriminanzanalyse, die lineare Lernmaschine und die Methode der *k*-nächsten Nachbarn gehören dem überwachten Lernverfahren an[48]. Bei der Auswahl eines geeigneten Klassifizierungsverfahrens sind die Vor- und Nachteile dieser Methoden gegeneinander abzuwägen. Neuronale Netze zeigen beispielsweise folgende modellbedingten Stärken:

- Durch die Fähigkeit zur Abstraktion und Generalisierung können Objekte klassifiziert werden, die nicht genau den gelernten Beispielen entsprechen
- Durch die Lernfähigkeit können sich neuronale Netze veränderten Umgebungsbedingungen anpassen
- Neuronale Netze haben eine hohe Fehlertoleranz gegenüber verrauschten bzw. unvollständigen Eingabemustern
- Bei neuronalen Netzen ist es nicht notwendig, ein mathematisches Modell zu erstellen, da die Akquisition von Wissen direkt durch Beispiele erfolgt.

Diesen Vorteilen stehen folgende Einschränkungen und Schwächen gegenüber:

- Die zahlreichen und damit schwer nachvollziehbaren Operationen lassen beim Benutzer den Eindruck einer Black-box entstehen
- Da das erlernte Wissen in den Gewichtsvektoren des Netzes abgespeichert wird, kann es nur schwer in Form von Regeln extrahiert werden
- Die Ermittlung optimaler Netzparameter ist schwierig
- Nicht immer ist eine Konvergenz des Lernvorgangs garantiert.

Als großen Vorteil eines neuronalen Netzes kann man das Erkennen linearer und nichtlinearer Beziehungen herausstellen, während viele mathematische Klassifizierungsverfahren meist nur lineare Zusammenhänge beschreiben können.

Ein weiterer Vorteil eines neuronalen Netzes gegenüber den anderen Klassifizierungsverfahren ist das schnelle Zuordnen eines weiteren Objektes in einen bereits klassifizierten Datensatz. Einem trainierten neuronalen Netz kann man dazu einen zweiten Datensatz, einen sogenannten Testdatensatz, übergeben. Dieser Datensatz dient dann nicht wie der erste zum Anpassen der Gewichte des Netzes. Vielmehr wird für jedes Objekt des Datensatzes genau ein Neuron bestimmt, das dem Anfragevektor am ähnlichsten ist. Die klassifizierten Objekte des sogenannten Trainingsdatensatzes werden dabei nicht verändert.

Wegen dieser zahlreichen Vorteile der neuronalen Netze gegenüber anderen Klassifizierungsverfahren werden in der vorliegenden Arbeit organische Reaktionen ausschließlich mit neuronalen Netzen nach Kohonen klassifiziert (siehe dazu auch Kapitel 3.3.4).

2.5.5 Das KMAP Programmsystem

Im Rahmen dieser Arbeit wurde das im Arbeitskreis von Gasteiger entwickelte Programmsystem „***K**ohonen **Map** Simulator*“, abgekürzt KMAP, in der Version 3.0 (teilweise auch 4.0) eingesetzt. Mit diesem Programm können Datensätze nach verschiedenen Algorithmen für neuronale Netzmodelle klassifiziert werden und die Ergebnisse, d.h. die resultierenden Kohonen-Karten, graphisch dargestellt werden.

2.6 Physikochemische Deskriptoren mittels PETRA

Zur Berechnung physikochemischer Effekte von Molekülen, Atomen und Bindungen wurde PETRA (***P**arameter **E**stimation for the **T**reatment of **R**eactivity **A**pplications*) in der Version 2.6 verwendet. Dieses im Arbeitskreis von Gasteiger in den letzten 20 Jahren entwickelte Programmsystem beruht auf schnellen empirischen Verfahren und ist zur Berechnung physikochemischer Deskriptoren großer Datenmengen prädestiniert.

Generell kann PETRA Ladungseigenschaften, Polarisierbarkeiten und energetische Effekte des Gesamtmoleküls, der Atome bzw. Bindungen berechnen. Eine Aufstellung der wichtigsten zur Zeit berechenbaren Effekte ist in Tabelle 2-6 gegeben. Alle zu den Ladungseigenschaften zählende Effekte werden aus den Elektronegativitäten unter Berücksichtigung der Konnektivität des Moleküls berechnet. Die Elektronegativität wird dabei nach dem Modell von Mulliken aus dem arithmetischen Mittel der Elektronenaffinität und der Ionisationsenergie ermittelt. Zur Berechnung der einzelnen physikochemischen Effekte wurden weitere Algorithmen entwickelt, die im folgenden näher vorgestellt werden.

Die zu den Ladungseigenschaften zählende σ-Ladungsverteilung und σ-Elektronegativität für Atome und Bindungen wird mit einem iterativen Algorithmus, namens *Partial Equalization of Orbital Electronegativity* (PEOE) berechnet[49],[50],[51],[52],[53]. Das Prinzip dieses Algorithmus beruht auf der Tatsache, daß bei der Bildung einer Bindung immer Ladung vom elektropositiven zum elektronegativen Atom fließt. Sobald aber Ladung vom elektropositiven Bindungspartner abfließt, wird dessen Elektronegativität erhöht, die Elektronegativität des elektronegativen Partners erniedrigt. Der Ladungstransfer kommt zum Erliegen, sobald sich die Elektronegativitäten angeglichen haben. Ein durch Ladungstransfer induziertes elektrisches Feld wirkt ebenfalls dem Ladungsfluß entgegen, so daß der Ausgleich der Elektronegativitäten stets unvollständig bleibt. Die Konvergenz dieses extrem schnellen Verfahrens ist nach maximal 10 Iterationsschritten erreicht.

Für die π-Ladungsverteilung und π-Elektronegativität von Atomen und Bindungen existiert ein analoges, iteratives Verfahren, namens *Partial Equalization of π-Electronegativity* (PEPE)[54],[55]. Hier werden zunächst alle Resonanzstrukturen eines π-Systems ausgehend von mesomeren Zentren (mit +M- oder –M-Effekt) erzeugt. Anschließend werden die Grenzstrukturen mit topologischen und energetischen Gewichtsfaktoren beurteilt. Anhand dieser Gewichtsfaktoren wird für jede Resonanzstruktur der Einfluß auf den Ladungsausgleich berechnet. Der Ladungstransfer entlang des π-Systems verändert an jedem Atom der Resonanzstruktur die Elektronegativität (siehe PEOE-Verfahren). Die Elektronendichte wird solange verschoben bis sich die Elektronegativitäten der Atome der gewichteten Resonanzstrukturen angepaßt haben. Die mit dieser Methode berechneten physikochemischen Eigenschaften entnehme man ebenfalls der Tabelle 2-6.

Die Gesamtladung eines Atoms, $q_{A,tot}$, oder einer Bindung, $\Delta q_{AB,tot}$, setzt sich additiv aus der σ- und π-Ladung zusammen.

Resonanzeigenschaften werden in PETRA über die Berechnung der π-Elektronegativität und der Elektronegativität der freien Elektronenpaare ermittelt. Die Elektronegativitäten werden für alle Atome berücksichtigt, die mit den Atomen des polaren Bindungsbruchs in Konjugation stehen. Es kann sowohl die Stabilisierung positiver R^{+}_{AB}, und negativer R^{-}_{AB} Ladungen oder deren Summe $R^{+/-}_{AB}$ berechnet werden. Eine Weiterentwicklung der Ladungsstabilisierungsenergie stellt die Delokalisationsenergie D^{+}_{AB}, D^{-}_{AB} und $D^{+/-}_{AB}$ dar.

Die Berechnung der Polarisierbarkeit basiert nicht auf den Elektronegativitäten der Atome des Moleküls. Für diesen Effekt wird auf ein Inkrementverfahren nach Kang und Jhon[56] zurückgegriffen. Die Summe aller Atominkremente, welche vom Hybridisierungszustand abhängig sind, wird für jede Sphäre berechnet und mit einem Gewichtsfaktor multipliziert. Diese einzelnen Beiträge einer topologischen Sphäre werden aufaddiert und liefern die sogenannte effektive Atompolarisierbarkeit α_{jd}[57]. Die effektive Bindungspolarisierbarkeit α_b berechnet sich aus dem Mittelwert der effektiven Atompolarisierbarkeiten der an der Bindung beteiligten Atome.

Die Standardversion von PETRA (Version 2.6) kann physikochemische Deskriptoren von Molekülen berechnen, die maximal 1.000 Atome enthalten.

Physikochemischer Effekt	Abkürzung	Symbol	Einheit	Methode
Bindungsdissoziationsenergie	BDE(A-B)		kJ/mol	Inkrement
Gesamtbindungs-dissoziationsenergie	TBDE(A-B)		kJ/mol	Inkrement
σ-Partialladung	QSIG(A)	$q_{A,\sigma}$	e	PEOE
σ-Elektronegativität	ENSIG(A)	$\chi_{A,\sigma}$	eV	PEOE
σ-Ladungsdifferenz der Atome der Bindung	DQSIG(AB)	$\Delta q_{AB,\sigma}$	e	PEOE
σ-Elektronegativitätsdifferenz der Atome der Bindung	DENSIG(AB)	$\Delta\chi_{AB,\sigma}$	eV	PEOE
Verschobene Ladungsmenge bei Sigma-Berechnung	SQIT(AB)	$Q_{AB,\sigma}$	e	PEOE
π-Partialladung	QPI(A)	$q_{A,\pi}$	e	PEPE
π-Elektronegativität	ENPI(A)	$\chi_{A,\pi}$	eV	PEPE
π-Ladungsdifferenz der Atome der Bindung	DQPI(AB)	$\Delta q_{AB,\pi}$	e	PEPE
π-Elektronegativitätsdifferenz der Atome der Bindung	DENPI(AB)	$\Delta\chi_{AB,\pi}$	eV	PEPE
Lone-Pair-Elektronegativität	ENLP(A)	$\chi_{A,LP}$	eV	PEPE
Atomare Gesamtladung	QTOT(A)	$q_{A,tot}$	e	PEOE u. PEPE
Gesamtladungsdifferenz der Atome der Bindung	DQTOT(AB)	$\Delta q_{AB,tot}$	e	PEOE u. PEPE
Resonanzstabilisierung einer positiven Ladung an Atom A	PSTAB(AB)	R^{+}_{AB}	–	
Resonanzstabilisierung einer negativen Ladung an Atom A	NSTAB(AB)	R^{-}_{AB}	–	
Gesamte Resonanz-stabilisierung	STABRES(AB)	$R^{+/-}_{AB}$	–	
Delokalisationsstabilisierung einer positiven Ladung an Atom A	PDELOC(AB)	D^{+}_{AB}	–	

Tab. 2-6: Eine Auswahl der von PETRA berechneten physikochemischen Deskriptoren.

Physikochemischer Effekt	Abkürzung	Symbol	Einheit	Methode
Delokalisationsstabilisierung einer negativen Ladung an Atom A	NDELOC(AB)	D^{-}_{AB}	–	
Gesamte Delokalisationsstabilisierung	SDELOC(AB)	$D^{+/-}_{AB}$	–	
Mittlere molekulare Polarisierbarkeit	POLARIZA	α_{mol}	$Å^3$	Inkrement
Effektive Atompolarisierbarkeit	APOLARIZ(A)	α_{jd}	$Å^3$	Inkrement
Bindungspolarisierbarkeit	BPOLARIZ(AB)	α_{b}	$Å^3$	Inkrement

Tab. 2-6: Eine Auswahl der von PETRA berechneten physikochemischen Deskriptoren.

3 Klassifizierungsverfahren

In diesem Kapitel wird zunächst auf die Codierungsmöglichkeiten für Reaktionszentren eingegangen, die bei der Reaktionsklassifizierung zum Einsatz kommen. Danach wird das im Rahmen dieser Arbeit ausgearbeitete Standardverfahren zur Klassifizierung vorgestellt.

3.1 Codierung des Reaktionszentrums für einen Identitätsvergleich

Für jede Reaktion sollte neben einem Codierungsvektor, der die Ähnlichkeit von Reaktionen zu beschreiben vermag, auch eine Codierungsform vorliegen, die zur Prüfung auf Identität zweier Reaktionszentren dient. Daher wird für jede Reaktion das Molekülfragment auf der Edukt- und Produktseite ermittelt, das die Atome und Bindungen enthält, die am Reaktionsgeschehen beteiligt sind. Für diese beiden Fragmente wird danach ein Hashcode berechnet.

Unter einem Hashcode versteht man in der Chemie die komprimierte Darstellung einer Datenstruktur, wie Atome, Moleküle oder Ensembles, mit konstanter Länge. Der erste Schritt bei der Berechnung eines Hashcodes ist nach Ihlenfeldt[58] die Ladungsäquilibrierung, damit verschiedene Resonanzstrukturen nicht zu unterschiedlichen Hashcodes führen. Danach wird für jedes Atom ein Anfangswert, der sogenannte *Seed*, ermittelt. Dieser berechnet sich aus mehreren Atomeigenschaften, wie die Zahl der Nachbarn, Zahl der Wasserstoffnachbarn, Ordnungszahl etc., die unter Benutzung einer Primzahlentabelle in Zahlenwerte übertragen werden, welche miteinander multipliziert werden. Aus diesem Anfangswert für ein Atom wird nun der Atomhashcode durch mehrere Rotations- und Exklusiv-Oder-Operationen berechnet. Die Rotationsprozeduren spreizen den anfänglichen Zahlenwert über die gesamte Bitvektorlänge, während durch die XOR-Verknüpfungen mit den Hashcodes benachbarter Atome die Umgebung des Atoms erfaßt wird. Bei dem Komprimierungsvorgang, der auch als Projektion eines Punktes im viel-dimensionalen Eigenschaftsraum auf eine einzige diskontinuierliche Koordinate beschrieben werden kann, tritt ein Informationsverlust ein. Deshalb ist es unmöglich, einen Hashcode in die ursprüngliche Datenstruktur zurück zu transformieren. Außerdem ist es wegen des begrenzten Wertebereichs nie ausgeschlossen, daß Kollisionen auftreten. Eine Kollision liegt vor, wenn für zwei verschiedene Datenstrukturen ein und derselbe Hashcode berechnet wird. Die Wahrscheinlichkeit einer Kollision ist um so geringer, je größer der mögliche Wertebereich im Vergleich zu der Zahl der zu transformierenden Daten ist. Mit dem zur Zeit als Standard geltenden 64-bit Code können rund 10 Millionen Datenstrukturen kollisionsfrei kodiert werden. Hashcodes werden in der Chemie zur Prüfung auf Identität zweier Atome, Moleküle etc. oder als Zugriffsschlüssel in großen Datenbanken eingesetzt.

In der vorliegenden Arbeit wird der Hashcode zur Prüfung auf Identität zweier Reaktionszentren herangezogen. Dazu berechnet man für die beiden Teile des Reaktionszentrums auf der Edukt- und Produktseite jeweils einen Hashcode. Stimmen zwei Reaktionen in beiden Hashcodes überein, so haben sie ein identisches Reaktionszentrum; haben sie nur einen Hashcode gemeinsam, so haben sie auf Edukt- oder Produktseite gleich reagierende Molekülfragmente. Mit Hilfe des Hashcodes ist somit einerseits ein schneller Zugriff auf Reaktionen mit identischen Reaktionszentren möglich. Andererseits lassen sich leicht Duplikate in einer Reaktionsdatenbank feststellen, oder in verschiedenen Datenbanken finden (siehe Kapitel 7.3.2).

3.2 Codierung des Reaktionszentrums für einen Ähnlichkeitsvergleich

3.2.1 Beschränkung auf einen Teil des Reaktionszentrums

Der Verlauf organischer Reaktionen wird von vielen Faktoren beeinflußt: Zuallererst wird das entstehende Produkt natürlich von den Edukten selbst bestimmt. Die Existenz bzw. Abwesenheit funktioneller Gruppen in den Edukten nimmt genauso Einfluss auf den Reaktionsweg wie die zahlreichen Reaktionsbedingungen, zu denen beispielsweise Lösungsmittel, Katalysator, Temperatur, Druck, Licht etc. zählen. Darüber hinaus können auch sterische, kinetische oder thermodynamische Effekte eine große Rolle spielen. Jeder dieser Effekte kann den Reaktionsweg so stark beeinflussen, daß anstelle eines geplanten Produkts eine ganz andere Verbindung entsteht. Viele kommerziell erhältliche Reaktionsdatenbanken gestatten meistens nur Rückschlüsse auf den Einfluß der Edukte. Es werden zwar stets für jede aufgenommene Reaktion die Edukte und Produkte gespeichert, doch nicht jede Reaktionsdatenbank liefert weitere Angaben zu den Reaktionen, wie beispielsweise die Reaktionsbedingungen. Aus diesem Grund muß man bei der Codierung von Reaktionen mit den strukturellen Merkmalen der Edukte und Produkte auskommen, will man nicht für viele Reaktionen ohne Reaktionsbedingungen die fehlenden Angaben aus der Originalliteratur ergänzen. Wenn man Reaktionen beschreiben will, muß die Molekülstruktur der Edukte und Produkte aber nicht in ihrer Gesamtheit erfaßt werden, denn während einer Reaktion werden meist nur einige wenige Bindungen im Molekül verändert. Somit ist es ausreichend, wenn man sich auf den reagierenden Molekülausschnitt auf Edukt- und Produktseite, das sogenannte Reaktionszentrum (siehe Kapitel 2.1.1), konzentriert.

Im Hinblick auf die beiden Haupteinsatzgebiete der Reaktionsklassifizierung, der Reaktionsvorhersage und der Syntheseplanung, wird stets nur ein Teil des Reaktionszentrums zur Codierung herangezogen. Bei der Reaktionsvorhersage sind zunächst nur die Ausgangsverbindungen bekannt, während das Produkt erst bestimmt werden muß. Bei der Codierung der

Reaktionen beschränkt man sich daher auf die am Reaktionsgeschehen beteiligten Atome und Bindungen in dem Eduktensemble. Andererseits ist bei der Syntheseplanung nur das Produkt bekannt; daher wird hier der Teil des Reaktionszentrums herangezogen, der aus dem Produktensemble stammt. Beim Datenbankenvergleich kann man den Teil des Reaktionszentrums frei wählen: Es kann entweder der Teil des Reaktionszentrums auf der Edukt- oder Produktseite herangezogen werden, oder auch das komplette Reaktionszentrum. In der vorliegenden Arbeit wird der Teil auf der Produktseite ausgewählt, da sich vor allem die unvollständige Codierung der Reaktionszentren auf der Eduktseite nachteilig auf die Klassifizierung auswirkt.

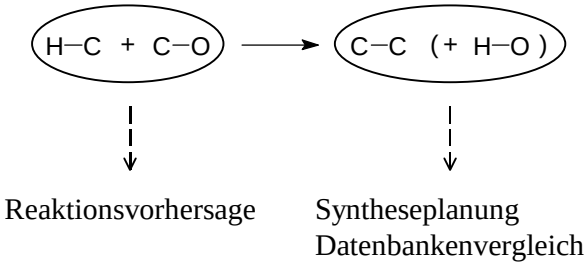


Abb. 3-1: Bei der Codierung von Reaktionen wird stets nur ein Teil des Reaktionszentrums herangezogen: Der Teil auf der Eduktseite für die Reaktionsvorhersage, der Teil auf der Produktseite für die Syntheseplanung und den Datenbankenvergleich.

3.2.2 Anforderungen an einen Codierungsalgorithmus

Das Reaktionszentrum einer Reaktion umfaßt alle Bindungen auf der Eduktseite, die gebrochen werden und alle Bindungen auf der Produktseite, die gebildet werden. In den Reaktionsdatenbanken werden häufig Reaktionssequenzen mit mehreren hintereinander durchgeführten Reaktionen wiedergegeben. Somit besteht auch das Reaktionszentrum der Gesamtreaktion aus den kumulativen Zentren der Einzelreaktionen. Daher ist es nicht verwunderlich, daß die in Reaktionsdatenbanken gespeicherten Reaktionszentren nicht nur aus je zwei oder drei Bindungen auf Edukt- und Produktseite bestehen, sondern häufig sehr viele Bindungen umfassen können. Die Größe der Reaktionszentren kann daher sehr unterschiedlich sein (siehe Abbildung 3-5). Andererseits können viele mathematisch-statistische Verfahren prinzipiell nur Objekte gleicher Länge vergleichen. Die Codierung unterschiedlich großer Reaktionszentren muß also neben einer Transformation der Struktur in einen mathematisch-statistisch lesbaren Zahlencode auch eine einheitliche Vektorlänge gewährleisten.

Ein Codierungsalgorithmus für Reaktionszentren muß des weiteren für Reaktionen mit gleichen Reaktionszentren einen untereinander vergleichbaren Codierungsvektor liefern. Es muß in jedem Falle sichergestellt sein, daß eine gleichartige Ausrichtung der Reaktionszentren und somit eine vergleichbare Anordnung der Vektorelemente vorliegt.

Die Konnektivität der Atome darf bei dem Codierungsvorgang nicht verändert werden, d.h. Reaktionszentren, die auf Produktseite das Fragment eines Ethylalkohols aufbauen, müs-

sen einen anderen Codierungsvektor aufweisen als Reaktionen, bei denen beispielsweise eine Ethergruppe aufgebaut wird. Außerdem muß man auch für chemisch ähnliche Reaktionszentren einen untereinander vergleichbaren Vektor erhalten. Unter chemisch ähnlichen Reaktionszentren versteht man zum Beispiel solche Zentren, die Elemente enthalten, welche chemisch ähnlich reagieren. Das chemisch ähnliche Verhalten einer Kohlenstoff-Chlor-Bindung im Vergleich zu einer Kohlenstoff-Brom-Bindung sollte daher auch im Codierungsvektor zum Ausdruck kommen (siehe Abbildung 3-2).

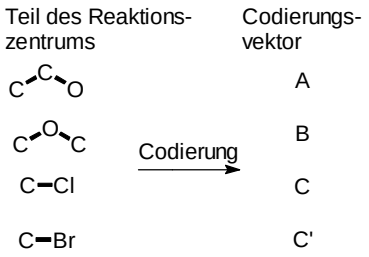


Abb. 3-2: Anforderungen an einen Codierungsalgorithmus: Zueinander isomere Teile zweier Reaktionszentren sollten möglichst unterschiedliche Codierungsvektoren A und B erhalten, während Fragmente mit ähnlichem chemischen Verhalten dagegen möglichst ähnliche Vektoren C und C' haben sollten. Die reagierenden Bindungen sind fett markiert.

Außerdem sollte der Codierungsvektor leicht interpretierbar sein um Rückschlüsse auf das Reaktionszentrum zu ermöglichen.

3.2.3 Realisierung des Codierungsalgorithmus

Die zuvor dargelegten Anforderungen an einen Codierungsalgorithmus werden mit neu entwickelten bzw. modifizierten Methoden realisiert, die im folgenden näher erläutert werden.

3.2.3.1 Die Linearnotation SMILES

SMILES (***S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification*)[59] ist eine leicht erlernbare, flexible Linearnotation, die es ermöglicht, chemische Strukturen in einer computerlesbaren Form darzustellen. Dave Weininger entwickelte diese Sprache in den frühen 80er Jahren an der U.S. Environmental Protection Agency in Duluth, MN, später im Medicinal Chemistry Project am Ponoma College, Claremont, CA. Heutzutage ist sie fester Bestandteil des „Daylight Toolkits“, das von der 1997 gegründeten Firma Daylight Chemical Information Systems, Inc. vertrieben wird.

Ein großer Vorteil von SMILES gegenüber anderen kanonischen Repräsentationen ist seine gute Lesbarkeit. SMILES beschreibt chemische Strukturen mittels der Graphentheorie, wobei Atome als Knoten und Bindungen – nach dem Valence-Bond Modell – als Kanten dargestellt werden. Ein SMILES-String setzt sich aus Atomsymbolen zusammen, wobei neben-

einander stehende Atome in der chemischen Struktur benachbart sind. Zur Beschreibung cyclischer Systeme werden Ringschlußbindungen aufgebrochen, und die an diesen Bindungen beteiligten Atome mit einer Ziffer, zur Andeutung der Nachbarschaft, markiert. Für Cyclohexan lautet der SMILES-String beispielsweise C1CCCCC1. Zwischen den Atomsymbolen können auch noch weitere ASCII-Zeichen eingeschoben sein, um Verzweigungen oder Bindungstypen abzubilden[60].

In der Zwischenzeit wurde SMILES zu einer Sprachenfamilie weiterentwickelt; dazu gehören beispielsweise SMIRKS (***SMI**LES **r**ea**k**tion **s**pecification*), die zur Beschreibung von Reaktionen der Form Ausgangsstoffe > Agens > Produkte entwickelt wurde. Ein anderer Sprachdialekt ist USMILES (***u**nique **SMILES***). USMILES erzeugt im Gegensatz zu SMILES aus einer chemischen Struktur immer einen eindeutigen SMILES-String, d.h. zwei identische chemische Strukturen besitzen ein- und denselben USMILES-Code, während die Abfolge der Atome in den SMILES-Codes unterschiedlich sein kann. Die Generierung eines USMILES-Strings wird mit einem Verfahren namens CANGEN erreicht, das sich aus einem zweistufigen Algorithmus, CANON und GENES, zusammensetzt[61]. Im ersten Schritt (CANON) werden die Atome der chemischen Struktur mit kanonischen Ziffern versehen, wobei auch hier die Struktur mit Hilfe der Graphentheorie beschrieben wird. Im zweiten Schritt (GENES) werden diese Ziffern verwendet, um aus der Graphendarstellung eine Baumdarstellung abzuleiten, indem beim Startatom begonnen wird und nach den kanonischen Ziffern verzweigt wird. Daraus erhält man schließlich den USMILES-String.

3.2.3.2 Modifikation des USMILES-Algorithmus

Für die Reaktionsklassifizierung wurde die Berechnung des im vorangegangenen Kapitel 3.2.3.1 vorgestellten USMILES-Codes geändert. Um die SMILES-Codes verschiedener Reaktionszentren miteinander vergleichen zu können, muß nicht nur für ein und dasselbe Zentrum ein eindeutiger SMILES-Code existieren, sondern die SMILES-Codes müssen auch chemisch äquivalent sein. USMILES vermag zwar einen eindeutigen USMILES-String auszugeben, doch für chemisch ähnliche Moleküle werden nicht immer chemisch äquivalente USMILES-Codes ermittelt. Chemisch äquivalent sind beispielsweise die beiden USMILES-Codes in Tabelle 3-1 deshalb nicht, weil durch die Reihenfolge der Atome auch die Bindungsrichtung festgelegt wird. Während beim LiC die Primärrichtung vom Metall zum Kohlenstoffatom zeigt, ist es beim CNa genau umgekehrt. Da die physikochemischen Eigenschaften einer Bindung von deren Richtung abhängig sind, müssen die USMILES-Codes in einer chemisch äquivalenten Reihenfolge ausgegeben werden.

Teil des Reaktionszentrums	USMILES-Code
Li~C und C~Li	[Li][C]
Na~C und C~Na	[C][Na]

Tab. 3-1: USMILES-Code für ausgewählte Beispiele. Der originale USMILES-Algorithmus führt zwar für beide Beispiele zu einem eindeutigen Code, jedoch sind die Codes chemisch nicht äquivalent. Die geschwungene Linie soll den nur teilweise kovalenten Charakter der Bindung verdeutlichen.

Dazu wird die im ersten Schritt von CANGEN ermittelte Rangfolge eines Atoms *A* modifiziert, indem man von der mit 100 multiplizierten Elektronegativität EN_A nach Pauling[62] des jeweiligen Atoms die alte Ziffer A_{rank}^{old} subtrahiert (siehe Gleichung 3-1).

$$A_{rank}^{new} = (EN_A * 100) - A_{rank}^{old}$$

Gleichung 3-1: Berechnung der Ranking-Nummern für ein Atom *A* ausgehend von der Ranking-Nummer für den USMILES-String.

Auf diese Weise erhalten die elektronegativsten Atome immer die kleinste Ziffer und bilden daher die Startatome für den zweiten Schritt der CANGEN-Methode. Andererseits bleiben die topologischen Differenzen zwischen Atomen desselben Elements erhalten. Diese Methode liefert also für jede chemische Struktur einen chemisch vergleichbaren, eindeutigen USMILES-String, wobei stets mit dem elektronegativsten Atom begonnen wird. Für die Beispiele der Tabelle 3-1 ergeben sich beispielsweise die in Tabelle 3-2 dargestellten modifizierten USMILES-Codes.

Teil des Reaktionszentrums	modifizierter USMILES-Code
Li~C und C~Li	[C][Li]
Na~C und C~Na	[C][Na]

Tab. 3-2: Modifizierter USMILES-Code für die ausgewählten Beispiele der Tabelle 3-1. Der angepaßte USMILES-Code ordnet die Atome nun in einer vergleichbaren Form an. Die geschwungene Linie soll den nur teilweise kovalenten Charakter der Bindung verdeutlichen.

Die Anordnung zueinander ähnlicher Reaktionszentren basierend auf dem modifizierten USMILES-Code hat den Vorteil, daß auch die Atome und Bindungen unabhängig von ihren physikochemischen Eigenschaften ausgerichtet werden. Somit werden auch die Reaktionszentren von Reaktionen mit umgekehrten elektronischen Eigenschaften, sogenannte Reaktionen mit inversem Elektronenbedarf, identisch ausgerichtet zu den Reaktionszentren der Reaktionen mit „normalen“ physikochemischen Eigenschaften.

Für den Fall, daß der Edukt- oder Produktteil des Reaktionszentrums symmetrisch ist, müssen weitere Algorithmenschritte durchlaufen werden, um eine eindeutige und chemisch vergleichbare Ausrichtung der Reaktionszentren zu garantieren. Falls man die Atome des betrachteten Teils des Reaktionszentrums auch in dem anderen Teil wiederfindet, so kann man diesen anderen Teil des Reaktionszentrums zur Ausrichtung heranziehen. Wenn man den nicht-untersuchten Teil in einer eindeutigen Weise anordnen kann und über die sogenannte

Atom-Atom-Mapping-Nummern die Atome im Edukt- und Produktensemble eindeutig zuordnen kann, so erhält man auch für den untersuchten Teil des Reaktionszentrums eine eindeutige Anordnung. Dies soll anhand des folgenden Beispiels erklärt werden. Die Knüpfung einer symmetrischen Kohlenstoff-Kohlenstoff-Bindung erfolgt meist aus zwei heterogenen Atompaaren, beispielsweise einer Kohlenstoff-Wasserstoff- und einer Kohlenstoff-Sauerstoff-Bindung. Man ordnet nun die zwei Atompaare des Reaktionszentrums auf der Eduktseite in einer eindeutigen Weise an – beispielsweise O–*C* und **C**–H – und hat auf diese Weise auch die Reihenfolge der beiden Kohlenstoffatome festgelegt, nämlich *C*–**C**. Mit Hilfe der eindeutigen Atom-Atom-Mapping-Nummern erhält man dann auch eine eindeutige Anordnung der beiden Kohlenstoffatome auf der Produktseite (siehe Tabelle 3-3).

Reaktionszentrum auf der Eduktseite		modifizierter USMILES-Code		resultierendes Reaktionszentrum auf der Produktseite
H–**C** + *C*–O	⇢	[O][*C*].[**C**][H]	⇢	*C*–**C**
C–O + **C**–H	⇢	[O][*C*].[**C**][H]	⇢	*C*–**C**
N–*C* + H–**C**	⇢	[N][*C*].[**C**][H]	⇢	*C*–**C**

Tab. 3-3: Ein symmetrischer Reaktionszentrumteil (C-C auf der Produktseite) wird durch eindeutige Anordnung des anderen Reaktionszentrumteils (O-C und C-H auf der Eduktseite) ausgerichtet.

Falls auch der andere Teil des Reaktionszentrums symmetrisch ist, so werden die Atome im Reaktionszentrum schließlich nach fallender σ-Elektronegativität angeordnet.

3.2.4 Einsatz physikochemischer Effekte

Nach der eindeutigen Ausrichtung der Atome und Bindungen des betrachteten Teils des Reaktionszentrums werden diese codiert.

Den Anforderungen eines Codierungsverfahrens zur Beschreibung der Reaktionszentren kommt am besten der Einsatz physikochemischer Effekte entgegen. Diese vermögen gegenüber einer topologischen Beschreibung die elektronischen Verhältnisse besser wiederzugeben. Zur Berechnung physikochemischer Effekte werden schnelle empirische Methoden eingesetzt, wie sie im PETRA Programmsystem (siehe Kapitel 2.6) realisiert sind.

Beschreibt man die den Reaktionsweg beeinflussende elektronische Situation des Reaktionszentrums mit physikochemischen Eigenschaften, so erhält man einen chemisch gut interpretierbaren Vektor. Um auch die elektronischen Einflüsse in der näheren Umgebung des Reaktionszentrums zu berücksichtigen, werden die physikochemischen Effekte nicht am isolierten Reaktionszentrum ermittelt. Vielmehr werden die physikochemischen Eigenschaften aller Atome und Bindungen im gesamten Molekül berechnet und danach die Zahlenwerte für das Reaktionszentrum herausgegriffen. In π-Systemen können beispielsweise die elektroni-

schen Eigenschaften einer Bindung durch weit entfernte funktionelle Gruppen stark beeinflußt werden (siehe Abbildung 3-3). Das Wasserstoffatom am γ-Kohlenstoffatom wird nach dem Vinylogieprinzip über die π-Bindung stark durch die elektronenziehende Carbonylgruppe beeinflusst. Analoges gilt auch für das angegebene aromatische System, bei dem die Kohlenstoff-Wasserstoff-Bindung stark durch die elektronenziehende Nitrogruppe in para-Stellung beeinflußt wird.

Abb. 3-3: Physikochemische Eigenschaften am Reaktionszentrum können durch weit entfernte Gruppen beeinflußt werden.

3.2.5 Codierung organischer Reaktionen mittels Autocorrelation

Die von Moreau[63] erstmals bei einer QSAR-Untersuchung beschriebene Autocorrelation eignet sich gut zur Codierung der Molekülkonstitution. Die sogenannten 2D-Autocorrelationsvektoren bieten einige Vorteile: Sie sind kompakt, unabhängig von der Atomnumerierung und haben unabhängig von der Molekülgröße eine konstante Länge. Zur Berechnung der Autocorrelationsfunktion *A(d)* werden die Atomeigenschaften p_i und p_j zweier Atome *i* und *j* miteinander multipliziert und über alle Atompaare summiert, wobei ein Diskretisierungsfaktor $\delta(d\text{-}d_{ij})$ berücksichtigt werden muß. Ist der Abstand *d* gleich der Anzahl der Bindungen zwischen den zwei Atomen *i* und *j* so nimmt δ den Wert 1 an, andernfalls den Wert 0 (siehe Gleichung 3-2).

$$A(d) = \frac{1}{2} \sum_{\substack{i,j \\ (i \neq j)}} p_i \cdot p_j \cdot d\left(d - d_{ij}\right)$$

Gleichung 3-2: Berechnung des Autocorrelationsvektors.

Topologische Autocorrelationsvektoren wurden im Arbeitskreis von Gasteiger bereits erfolgreich zur Ähnlichkeitsanalyse von biologisch aktiven Verbindungen eingesetzt[64].

Anhand einiger Untersuchungen zur Einsetzbarkeit von Autocorrelationsvektoren bei der Codierung von Reaktionszentren zeigte sich jedoch, daß die Codierungsvektoren nur schwer Rückschlüsse auf die ursprünglichen Zentren erlauben. Aus diesem Grund wurde diese Codierungsmöglichkeit nicht weiter verfolgt.

3.3 Beschreibung eines Standardverfahrens

Ziel dieser Arbeit ist es, ein Standardverfahren zur Reaktionsklassifizierung zu entwikkeln, das universell einsetzbar sein sollte. Die neu entwickelten bzw. modifizierten Methoden sollen sowohl beim Datenbankenvergleich, in der Reaktionsvorhersage, sowie der Syntheseplanung Verwendung finden, und zu möglichst guten Ergebnissen führen. In den meisten Fällen könnte durch eine Anpassung der physikochemischen Eigenschaften, der Netzparameter, der Lernparameter etc. eine Optimierung in Bezug auf eine chemisch korrektere Klassifizierung durchgeführt werden. Ein solcher Optimierungsprozeß könnte zwar mit zeitintensiveren Studien für jeden Fall durchlaufen werden, jedoch wird in der vorliegenden Arbeit das in den nächsten Kapiteln vorgestellte Standardverfahren angewandt, das durchwegs zu guten Resultaten führt.

3.3.1 Auswahl eines Standardsatzes an Deskriptoren

Zu einem Standardverfahren zählt auch ein Satz an Deskriptoren aus physikochemischen Effekten, der die Reaktionen für alle Einsatzgebiete zu codieren vermag. Die Auswahl der physikochemischen Eigenschaften wird einerseits durch statistische Methoden eingeschränkt, andererseits auch durch chemische Überlegungen getroffen. Nach der Analyse der Korrelationsmatrix eines Datensatzes wird jeweils eine Eigenschaft von hochkorrelierten Eigenschaftspaaren ausgeschlossen. Obwohl man mathematisch korrekt die Korrelationsmatrix für jeden Datensatz einzeln berechnen müßte, wird hier stellvertretend für alle Datensätze eine Korrelationsmatrix näher diskutiert. Dieser Datensatz entstammt der Theilheimer Reaktionsdatenbank und enthält insgesamt 75.070 Datenpunkte für 16 ausgewählte physikochemische Bindungseigenschaften (siehe Tabelle 3-4).

Anhand dieser Korrelationstabelle und einer Reihe chemischer Begründungen werden folgende Bindungseigenschaften ausgewählt, wobei der zweite bis sechste folgende Effekt von PETRA (siehe Kapitel 2.6) berechnet wird:

- Bindungsordnung, b_o

Die Korrelation mit den anderen physikochemischen Effekten ist für die Bindungsordnung nur minimal. Die Bildung oder Spaltung einer π-Bindung verläuft über andere Mechanismen als die Bildung einer Einfachbindung.

- σ-Elektronegativitätsdifferenz der Atome der Bindung, $\Delta\chi_{AB,\sigma}$
- π-Elektronegativitätsdifferenz der Atome der Bindung, $\Delta\chi_{AB,\pi}$

Die σ- und π-Elektronegativität bestimmen im wesentlichen den Charakter einer Bindung, da die an einer Bindung beteiligten Atome die Elektronen in unterschiedlichem Maße anzuziehen vermögen. Die π-Elektronegativität kann im Gegensatz zur Bindungsordnung nur zwi-

	b_o	BDE	α_b	$\Delta\chi_\pi$	$\Delta\chi_\sigma$	Δq_π	Δq_σ	Δq_{tot}	D^-	R^-	D^+	R^+	$D^{+/-}$	Q_σ	$R^{+/-}$	TBDE
b_o	1,00	-0,12	-0,14	-0,27	0,10	0,01	-0,02	-0,02	0,09	0,05	0,01	0,01	0,09	0,00	0,11	0,22
BDE	-0,12	1,00	0,06	0,15	-0,07	0,00	-0,10	-0,09	0,08	0,07	0,00	0,04	0,02	0,01	0,05	**0,76**
α_b	-0,14	0,06	1,00	-0,10	-0,34	0,01	-0,01	-0,02	0,33	0,36	0,25	0,36	0,32	0,01	0,44	-0,01
$\Delta\chi_\pi$	-0,27	0,15	-0,10	1,00	0,28	0,01	-0,05	-0,04	-0,18	-0,15	-0,09	-0,25	-0,10	-0,02	-0,20	-0,06
$\Delta\chi_\sigma$	0,10	-0,07	-0,34	0,28	1,00	-0,04	0,07	0,06	-0,22	-0,19	-0,16	-0,19	-0,17	-0,01	-0,18	0,04
Δq_π	0,01	0,00	0,01	0,01	-0,04	1,00	-0,06	0,00	0,01	0,01	0,01	0,01	0,01	-0,08	0,01	0,00
Δq_σ	-0,02	-0,10	-0,01	-0,05	0,07	-0,06	1,00	**0,90**	-0,02	-0,03	0,07	0,09	0,05	0,11	0,06	-0,14
Δq_{tot}	-0,02	-0,09	-0,02	-0,04	0,06	0,00	**0,90**	1,00	-0,02	-0,02	0,07	0,08	0,06	0,06	0,05	-0,13
D^-	0,09	0,08	0,33	-0,18	-0,22	0,01	-0,02	-0,02	1,00	**0,78**	0,66	0,70	0,64	0,00	0,59	0,09
R^-	0,05	0,07	0,36	-0,15	-0,19	0,01	-0,03	-0,02	**0,78**	1,00	0,50	0,57	0,49	0,00	0,52	0,07
D^+	0,01	0,00	0,25	-0,09	-0,16	0,01	0,07	0,07	0,66	0,50	1,00	**0,77**	**0,91**	0,00	0,64	-0,04
R^+	0,01	0,04	0,36	-0,25	-0,19	0,01	0,09	0,08	0,70	0,57	**0,77**	1,00	0,69	-0,01	**0,77**	0,00
$D^{+/-}$	0,09	0,02	0,32	-0,10	-0,17	0,01	0,05	0,06	0,64	0,49	**0,91**	0,69	1,00	0,00	**0,77**	0,02
Q_σ	0,00	0,01	0,01	-0,02	-0,01	-0,08	0,11	0,06	0,00	0,00	0,00	-0,01	0,00	1,00	0,00	0,02
$R^{+/-}$	0,11	0,05	0,44	-0,20	-0,18	0,01	0,06	0,05	0,59	0,52	0,64	**0,77**	**0,77**	0,00	1,00	0,06
TBDE	0,22	**0,76**	-0,01	-0,06	0,04	0,00	-0,14	-0,13	0,09	0,07	-0,04	0,00	0,02	0,02	0,06	1,00

Tab. 3-4: Korrelationsmatrix der PETRA-Effekte für Daten aus der Theilheimer Reaktionsdatenbank. Es wurden alle Bindungseigenschaften auf der Produktseite der Reaktionszentren berücksichtigt. Alle Korrelationskoeffizienten mit einem Zahlenwert größer 0,75 sind markiert dargestellt.

schen Einfach- und Mehrfachbindungen differenzieren, so daß beide Eigenschaften notwendig sind.

– Gesamtladungsdifferenz der Atome der Bindung, $\Delta q_{AB,tot}$

Die Gesamtladungsdifferenz zwischen zwei Atomen einer Bindung beschreibt am besten die elektronischen Verhältnisse in dieser Bindung.

– Delokalisationsstabilisierung einer negativen Ladung, D^-_{AB}

– Delokalisationsstabilisierung einer positiven Ladung, D^+_{AB}

Wie man anhand der Tabelle 3-4 sieht, reichen von den insgesamt sechs Stabilisierungsenergien (D^-, R^-, D^+, R^+, $D^{+/-}$ und $R^{+/-}$) zwei Effekte, nämlich D^- und D^+, aus, um die Stabilität der polaren Zwischenstufen zu erfassen. So besitzen beispielsweise D^- und R^- einen Korrelationskoeffizienten von 0,78, D^+ und R^+ einen von 0,77. Die gesamte Delokalisationsstabilisierung $D^{+/-}$ ist mit einem Koeffizienten von 0,91 mit der Delokalisationsstabilisierung einer positiven Ladung D^+ korreliert. Zwischen der gesamten Resonanzstabilisierung $R^{+/-}$ und der gesamten Delokalisationsstabilisierung $D^{+/-}$ besteht ebenfalls eine hohe Korrelation von 0,77.

Aus den insgesamt 16 Bindungseigenschaften werden also sechs physikochemische Effekte ausgewählt. Die restlichen zehn Eigenschaften werden zum einen deshalb nicht ausgewählt, weil sie wie im Falle der Bindungs- und Gesamtbindungsdissoziationsenergie (BDE

und TBDE) miteinander hoch korreliert sind. Zum anderen spielen bei organischen Reaktionen heterolytische Bindungsbrüche die Hauptrolle, so daß auf Deskriptoren, die homolytische Brüche beschreiben, wie die Bindungsdissoziationsenergie, verzichtet wird.

Wie sich in den Anwendungskapiteln 4 bis 7 noch zeigen wird, haben sich diese ausgewählten physikochemischen Effekte bei der Klassifizierung von Reaktionen in allen Anwendungsgebieten bewährt.

3.3.2 Skalierung des Codierungsvektors

Der Codierungsvektor jeder Reaktion muß zunächst einen Datenaufbereitungsprozeß durchlaufen, bevor er als Eingabevektor eines neuronalen Netzes eingesetzt werden kann. Als erster Schritt dieser Datenaufbereitung wird eine Skalierung der Zahlenwerte durchgeführt, bevor im zweiten Schritt eine einheitliche Länge des Codierungsvektors durchgesetzt wird (siehe Kapitel 3.3.3). In Abbildung 3-4 sind die Histogramme für die sechs ausgewählten physikochemischen Effekte dargestellt.

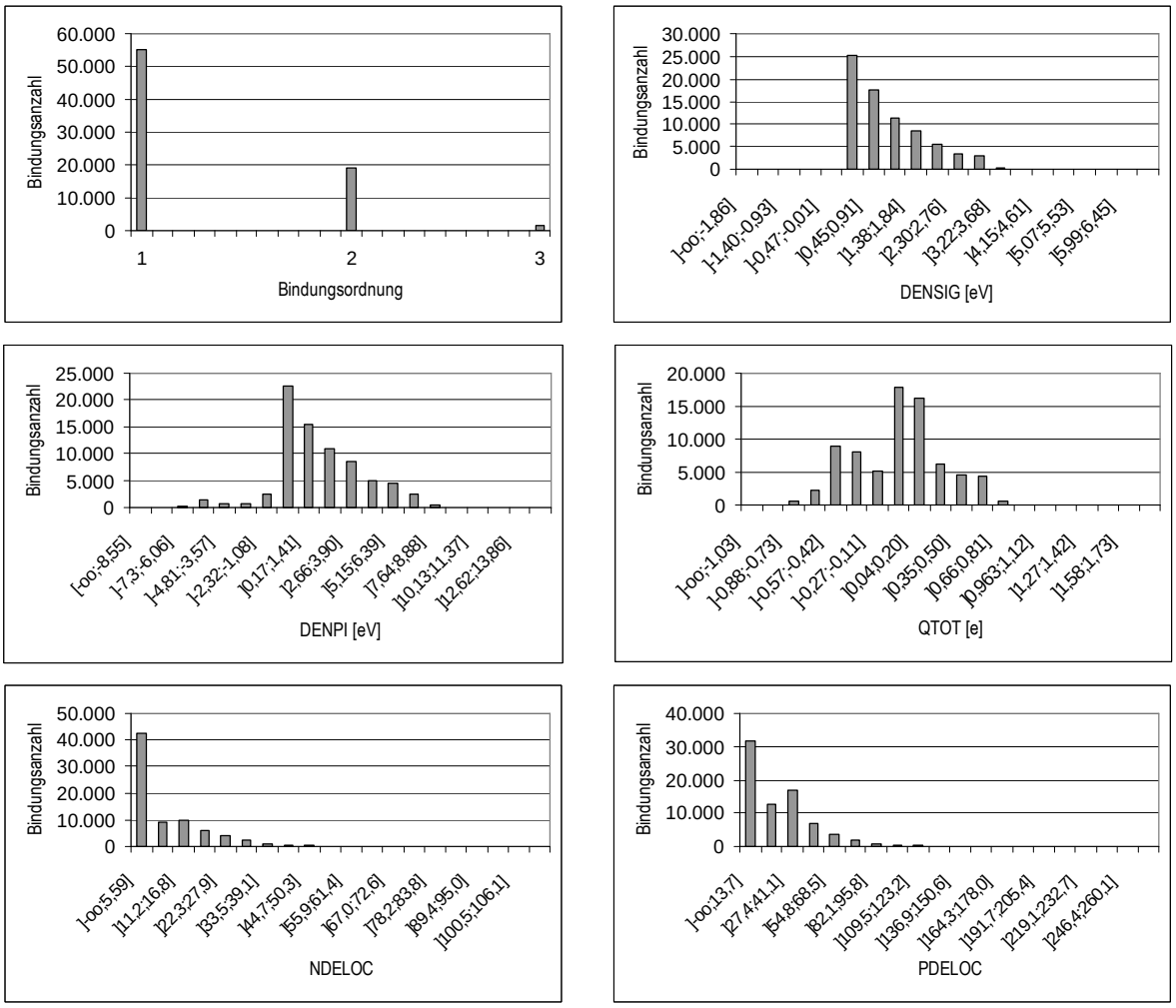


Abb. 3-4: Histogramme für die sechs ausgewählten physikochemischen Eigenschaften. Insgesamt gingen 75.070 Bindungen aus Reaktionsprodukten der Theilheimer Reaktionsdatenbank ein.

Dabei wurden alle reagierenden 75.070 Bindungen auf der Produktseite der insgesamt 33.613 codierbaren Reaktionen in der Theilheimer Reaktionsdatenbank herangezogen und als Histogramme in Abbildung 3-4 dargestellt.

Nur die Bindungsordnung weist eine diskrete Werteverteilung auf, alle anderen Bindungseigenschaften sind kontinuierlich, aber nicht normalverteilt.

Da die Wertebereiche sowohl in ihrer Lage als auch in ihrer Ausdehnung unterschiedlich sind, wird jede Eigenschaft mit einer linearen Transformationsgleichung skaliert. Die einzelnen Skalierungsparameter der Funktion y = a*x + b entnehme man Tabelle 3-5.

	b_o	$\Delta\chi_\sigma$	$\Delta\chi_\pi$	Δq_{tot}	D^-	D^+
a	1,00000	0,33333	0,142857	1,00000	0,02000	0,01000
b	-2,00000	0,00000	0,000000	0,00000	0,00000	0,00000

Tab. 3-5: Skalierungswerte für die sechs ausgewählten physikochemischen Eigenschaften nach einer linearen Gleichung y=a*x+b.

3.3.3 Länge des Codierungsvektors

Im zweiten Schritt des Datenaufbereitungsprozesses wird die Bedingung einer konstanten Länge des Codierungsvektors überprüft. Zur Festlegung einer Standard-Vektorlänge wird zunächst die Größe der Reaktionszentren auf der Produktseite untersucht. Da in Reaktionsdatenbanken häufig Reaktionssequenzen abgespeichert sind, gehören oft mehr als zwei oder drei Bindungen auf der Edukt- und der Produktseite zum Reaktionszentrum. In Abbildung 3-5 ist ein Histogramm der zum Reaktionszentrum zählenden Bindungen auf der Seite der Produkte für alle Reaktionen der Theilheimer Reaktionsdatenbank wiedergegeben.

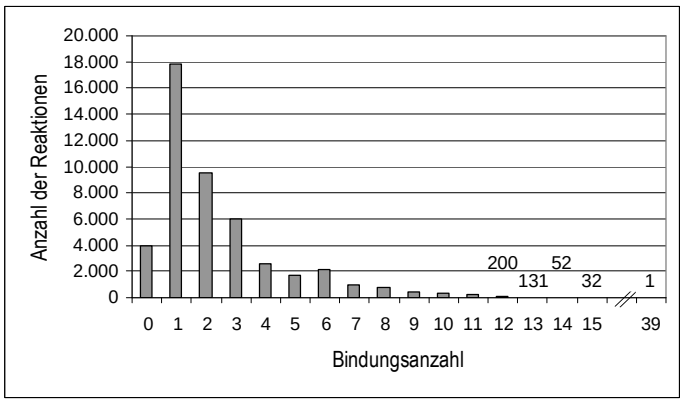


Abb. 3-5: Histogramm zur Reaktionszentrengröße. Für alle Reaktionen der Theilheimer Reaktionsdatenbank ist die Anzahl der am Bindungsumordnungsprozeß beteiligten Bindungen im Produktensemble ermittelt worden.

Aus Abbildung 3-5 geht hervor, daß nur 9.523 (20,4%) Reaktionen zwei reagierende Bindungen im Produktensemble aufweisen. Die meisten Reaktionen (17.874 oder 38,2%) weisen nur eine Bindung auf, da oft das Nebenprodukt nicht abgespeichert wird. Aus dem gleichen

Grund findet man auch relativ viele Reaktionen, bei denen scheinbar keine Bindung geknüpft wird. Hierbei handelt es sich beispielsweise um Eliminierungs- oder Spaltungsreaktionen sowie um unvollständig codierte Reaktionsbeispiele in der Datenbank. Mit zunehmender Größe des Reaktionszentrums nimmt auch die Anzahl der Reaktionsbeispiele ab; eine Reaktion weist mit 39 Bindungen im Produktensemble das größte Reaktionszentrum auf der Produktseite auf. Bemerkenswert ist bei dem Kurvenverlauf die Bindungsanzahl 6. Hier liegt die Reaktionsanzahl höher als man aus dem abfallenden Kurvenverlauf erwarten würde. Dies liegt zum einen an der energetisch begünstigten Ausbildung aromatischer Systeme, oder an speziellen elektrocyclischen Reaktionstypen, wie der Diels-Alder-Reaktion, bei denen an insgesamt sechs Bindungen Veränderungen eintreten.

Bei der Festlegung der maximalen Vektorlänge bzw. der maximalen Reaktionszentrengröße muß man einerseits zwischen einer angestrebten, möglichst vollständigen Codierung aller Reaktionen eines Datensatzes und andererseits einer vertretbaren Vektorlänge abwägen. Um auch den wichtigen Reaktionstyp der Diels-Alder-Reaktionen zu erfassen, wird für die weiteren Untersuchungen die maximale Größe des Reaktionszentrums auf sechs Bindungen im Edukt- oder Produktensemble festgelegt. Somit hat jeder Codierungsvektor einer Reaktion, der aus 6 Bindungen zu je 6 physikochemischen Effekten besteht, die Länge 36.

Im zweiten Schritt des Datenaufbereitungsprozesses werden die Vektoren von Reaktionen, die weniger als sechs Bindungen im Reaktionszentrum auf der Edukt- oder Produktseite aufweisen, mit weiteren Vektorelementen aufgefüllt, bis die konstante Vektorlänge erreicht ist. Somit ist sichergestellt, daß jede Reaktion durch einen 36-dimensionalen Codierungsvektor repräsentiert wird. Als Zahlenwert für unbesetzte Vektorelemente wird unabhängig von der Eigenschaft der Wert –5,0 eingesetzt. Dieser Wert ist die höchste ganzzahlige Grenze, die von keinem skalierten Zahlenwert der sechs ausgewählten physikochemischen Eigenschaften von insgesamt 75.070 untersuchten Bindungen unterschritten wird.

Nachdem die Skalierung und die Dimension der Codierungsvektoren festgelegt wurde, können sie einer Klassifizierungsmethode übergeben werden, auf die im folgenden Kapitel näher eingegangen wird.

3.3.4 Klassifizierung der Codierungsvektoren

In der vorliegenden Arbeit werden zur Klassifizierung der Codierungsvektoren aus den in Kapitel 2.5.4 genannten Gründen ausschließlich neuronale Netze verwendet. Dazu werden die codierten Reaktionen einem neuronalen Netz nach Kohonen zum Trainieren übergeben. Das neuronale Netz projiziert den viel-dimensionalen Eingabevektor in einen zwei-dimensionalen Raum, und trägt somit jede Reaktion unter Erhalt der Ähnlichkeitsbeziehungen in ein Neuron ein. Wie man anhand Abbildung 3-6 erkennt, kann man sowohl aus der Richtung, als auch aus der Entfernung zweier Reaktionen, die in verschiedenen Neuronen projiziert wer-

den, Rückschlüsse auf die Ähnlichkeit ziehen. Je kleiner die Entfernung der beiden Neuronen ist, desto ähnlicher sind die Reaktionen zueinander. Aus der Richtungskomponente kann man ableiten, mit welcher anderen Klasse die Reaktion Ähnlichkeit hat.

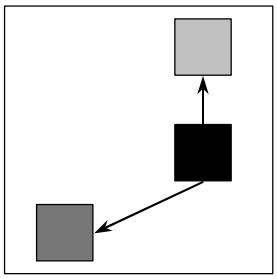


Abb. 3-6: Bei einer zwei-dimensionalen Projektion kann man sowohl aus der Richtung, als auch aus der Entfernung Ähnlichkeitsbeziehungen ableiten.

Ein trainiertes Kohonen-Netz ist auch ideal zur Vorhersage der Produkte oder Edukte einer Reaktion geeignet. Für jede Reaktion eines Testdatensatzes wird dabei genau ein Neuron bestimmt, das dem Anfragevektor am ähnlichsten ist. Falls in diesem sogenannten Gewinnerneuron während des Trainings Reaktionen projiziert wurden, so hat man die ähnlichsten Reaktionen zur Anfragereaktion bestimmt. Diese Vorhersageleistung neuronaler Netze ist sowohl in der Syntheseplanung, als auch in der Reaktionsvorhersage von großem Vorteil.

3.3.5 Festlegung der neuronalen Netzkonfiguration

Bevor die codierten Reaktionen mit neuronalen Netzen nach Kohonen klassifiziert werden können, muß zunächst das neuronale Netz konfiguriert werden. In zahlreichen Studien hat sich gezeigt, daß sich vor allem neuronale Netze mit rechteckiger Topologie bei heterogenen Datensätzen bewähren. Bei Verwendung eines solchen Netzes werden „Außenseiter“, also Objekte, die stark vom Durchschnitt abweichen, bevorzugt an den Rand des Netzes projiziert. Toroidale Netze sollte man dagegen bevorzugt bei homogenen Datensätzen einsetzen. Reaktionsdatenbanken enthalten oft Reaktionen, bei denen die physikochemischen Eigenschaften der Reaktionszentren stark vom Mittelwert abweichen. Aus diesem Grund werden in der vorliegenden Arbeit ausschließlich Kohonen-Netze mit rechteckiger, namentlich quadratisch-planarer, Topologie eingesetzt.

Eine typische Kommandoabfolge zum Trainieren eines Kohonen-Netzes mit KMAP 3.0 ist in Tabelle 3-6 wiedergegeben. Aus dieser ist auch die Konfiguration der Netzparameter wie die Lernrate etc. ersichtlich.

KMAP-Kommando	Bemerkung
create 36 n n	Anlegen eines Kohonen-Netzes der Dimension 36 x n x n, wobei n entweder aus der Quadratwurzel der Datenanzahl berechnet oder durch einen Optimierungsprozeß bestimmt wird.
set dnc 800 srdS	Dynamisches Verringern der Spannweite und der Lernrate; der Trainingsprozeß wird automatisch beendet, wenn die Lernrate kleiner als 0,1 und die Spannweite kleiner als 0,1 wird.
set dnc1 1 0.95	Änderung der Spannweite und Angabe des Lernfaktors
set par i 0.7	Festlegen der Spannweite und der Lernrate zu Beginn, wobei i so gewählt wird, daß es ca. 1/5 von n beträgt.
set top_type r	Auswahl der rechteckigen Topologie
init_net	Initialisierung der Gewichte
load_data ./train	Einlesen des Trainingsdatensatzes
train j	Trainieren des Netzes mit maximal j Zyklen
set c_type x	Konfliktneuronen werden mit einem X markiert
set color 0 1	Setzen der Farbenanzahl
show_map	Zeige das Kohonen-Netz als Text
show_net ./train	Die Konfiguration des Netzwerkes und alle Gewichtsvektoren werden abgespeichert
predict -cl ./train	Die Koordinaten des Gewinnerneurons, seine Gewichtsvektoren sowie die Nummer des Eingabevektors werden abgespeichert.
(load_data ./test)	Falls eine Vorhersage gewünscht wird, wird der Testdatensatz eingelesen. Mit weiteren nicht näher aufgeführten Befehlen kann dieser angezeigt und das resultierende Netz abgespeichert werden.
quit	Beendet KMAP

Tab. 3-6: Typische Kommandoabfolge zum Trainieren und Testen eines Datensatzes mit KMAP der Version 3.0.

4 Praktische Anwendung: Datenbankenvergleich

4.1 Angewandte Methode

Reaktionsdatenbanken kann man entweder einem Identitäts- oder einem Ähnlichkeitsvergleich unterziehen. Während der Identitätsvergleich bereits in Kapitel 2.3 erläutert wurde, werden in diesem einleitenden Kapitel zum Datenbankenvergleich die Methoden für einen Ähnlichkeitsvergleich erläutert.

Der Ähnlichkeitsvergleich beruht auf der in Kapitel 3 vorgestellten Klassifizierungsmethode. Es werden alle Reaktionen einer Reaktionsdatenbank nach dem in Kapitel 3.3 diskutierten Standardverfahren codiert. Bei diesem Verfahren werden alle Reaktionen, die auf der Produktseite höchstens 6 gebildete Bindungen aufweisen, mit sechs physikochemischen Effekten codiert. Anschließend werden diese Reaktionen in Form ihrer Codierungsvektoren einem Kohonen-Netz zur Klassifizierung übergeben. Dieses Netz paßt nun die Gewichte in der Weise an, daß die Reaktionen dieses Datensatzes möglichst über das gesamten Netz verteilt eingetragen werden, wobei aufgrund der topologieerhaltenden Eigenschaft des Kohonen-Netzes ähnliche Reaktionen im gleichen oder in benachbarten Neuronen eingetragen werden. Beim Vergleich mit anderen Datenbanken wird dieser Bereich als vollständiger Reaktionsraum betrachtet und und das trainierte Netz übernimmt die Funktion eines Referenznetzwerkes (siehe Abbildung 4-1). Eine der Zielsetzungen beim Aufbau der Theilheimer Reaktionsdatenbank war die Erfassung eines möglichst breiten Reaktionsspektrums (siehe Kapitel 2.2.1). Aus diesem Grund wird die Theilheimer Reaktionsdatenbank beim Datenbankenvergleich ausnahmslos als Referenzdatenbank herangezogen, so daß man über ein Referenznetzwerk mit einem möglichst ausgedehnten Reaktionsraum verfügt.

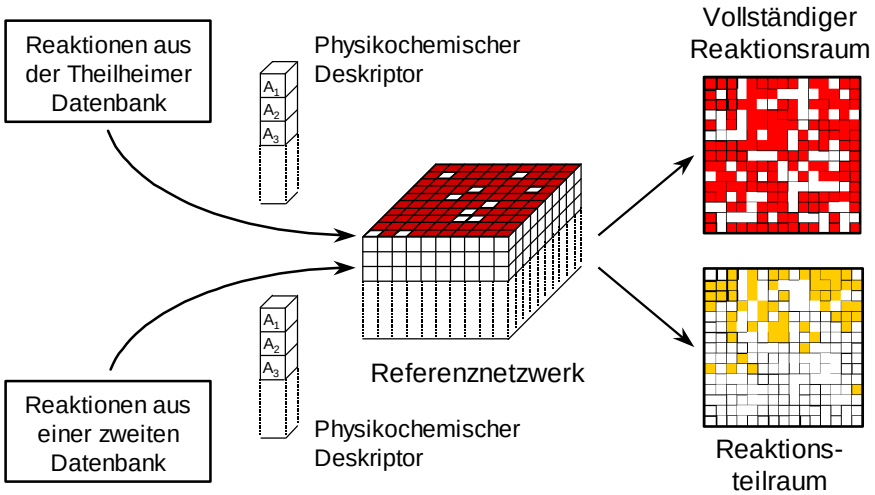


Abb. 4-1: Methode zum Vergleich von Reaktionsdatenbanken: Ein Reaktionsdatensatz (aus der Theilheimer Datenbank) wird zum Trainieren eines Kohonen-Netzes herangezogen. Dieser Datensatz deckt den Reaktionsraum vollständig ab. Ein zweiter Datensatz wird als Testdatensatz in dieses Referenznetz projiziert. Der aufgespannte Reaktionsteilraum wird schließlich analysiert.

Wird nun ein zweiter Datensatz mit diesem Referenznetzwerk verglichen, so werden die Reaktionen dieses Datensatzes auf dieselbe Weise codiert. Anschließend werden diese codierten Reaktionen aber nicht zum Trainieren eines neuen Netzes herangezogen, sondern in das bereits trainierte Referenznetzwerk projiziert. Dabei erfolgt in dem neuronalen Netz keine Änderung der Gewichte. Statt dessen werden die einzelnen Codierungsvektoren nur aufgrund der kleinsten Euklidischen Distanz (siehe Gleichung 2-1) in das entsprechende Neuron des trainierten Kohonen-Netzes eingetragen. Somit sind Reaktionen aus dem ersten und zweiten Datensatz, die im selben Neuron oder nah benachbarten Neuronen eingetragen sind, zueinander ähnlich. In der Regel erfüllt dieser zweite Teilraum den vollständigen Reaktionsraum nur zu einem Bruchteil. Diesen vom zweiten Datensatz eingenommenen kleineren Raum bezeichnet man als Reaktionsteilraum. Aus der Größe und der Lage dieses Reaktionsteilraums kann man Rückschlüsse auf die Ähnlichkeit der beiden Datensätze ziehen. Im folgenden Kapitel 4.2 wird zunächst der Aufbau eines Referenznetzwerkes aus der Theilheimer Reaktionsdatenbank beschrieben. In den Kapiteln 4.3 und 4.4 werden dann zwei andere Reaktionsdatenbanken mit diesem Referenznetzwerk verglichen.

4.2 Klassifizierung der Theilheimer Reaktionsdatenbank

Nach der in Kapitel 3.3 beschriebenen Methode wird die Theilheimer Reaktionsdatenbank (siehe Kapitel 2.2.1) klassifiziert. Zur Codierung sollen die physikochemischen Effekte der Bindungen im Reaktionszentrum auf der Produktseite herangezogen werden. Von den insgesamt 46.785 Reaktionen können 7.012 Reaktionsbeispiele nicht klassifiziert werden, da die Anzahl der Bindungen im Reaktionszentrum auf der Produktseite außerhalb des gewählten Bereichs von 1 bis maximal 6 Bindungen liegt (siehe Abbildung 3-5). Von den verbleibenden 39.773 Reaktionen kann PETRA in 2.403 Fällen keine physikochemischen Effekte berechnen, da beispielsweise nicht implementierte Elemente, wie Übergangsmetalle, im Produkt enthalten sind. Schließlich bleiben noch 3.757 Reaktionen infolge einer unvollständigen Reaktionscodierung in der Datenbank unberücksichtigt. Bei diesen Reaktionen ist beispielsweise der Teil des Reaktionszentrums auf der Produktseite symmetrisch. Wegen fehlender Atom-Atom-Mapping-Nummern in den Edukten bzw. Produkten ist keine eindeutige Festlegung der Atom- und Bindungsreihenfolge möglich. Somit werden von insgesamt 46.785 Reaktionen 33.613 (71,8%) Reaktionen codiert. Die codierten Reaktionen werden einem neuronalen Netz nach Kohonen (siehe Kapitel 2.5.3) zur Klassifizierung übergeben. Die Größe des Kohonen-Netzes errechnet sich aus der Quadratwurzel des vierten Teils der codierten Objektanzahl. Während man bei kleineren Datensätze von ca. 100 bis 200 Objekten die Quadratwurzel aus der 2-fachen Objektanzahl verwendete[6], so hat sich der vierte Teil der Objektanzahl bei Datensätzen in der Größenordnung von 30.000 bis 40.000 Objekten gut bewährt. Die anderen Netzparameter entnehme man Tabelle 4-1.

Netzparameter	Wert	Kommando
Netzwerkgröße	92 x 92	create 36 **92 92**
Netzwerktopologie	quadratisch-planar	set top_type r
Dimension der Neuronen	36	create **36** 92 92
Anzahl der maximal durchlaufenen Zyklen	80.000	train 80000
Lernrate zu Beginn h ($t=0$)	0,7	set par 18 **0.7**
Lernfaktor a	0,95	set dnc1 1 **0.95**
Spannweiten zu Beginn s_x (t=0), s_y (t=0)	18	set par **18** 0.7
Änderung der Spannweiten Ds_x, Ds_y	1,0	set dnc1 **1** 0.95
Zyklen konstanter Trainingsparameter t_s	800	set dnc **800** srdS

Tab. 4-1: Netz- und Trainingsparameter für die Klassifizierung der 33.613 Reaktionen aus der Theilheimer Reaktionsdatenbank. Die Kommandos sind für die KMAP Version 3.0 angegeben.

Für diesen Datensatz wird exemplarisch die zeitliche Entwicklung des Netztrainings nachvollzogen. Das neuronale Netz wird vor Beginn des Trainings initialisiert, indem die Gewichte der Vektoren mit zufälligen Werten gefüllt werden. Die übergebenen Objekte werden daher in diesem Netz, das bisher keinen einzigen Trainingszyklus durchlaufen hat, ohne Ähnlichkeitsbeziehungen willkürlich verteilt (siehe Abbildung 4-2a). In dieser wie in den folgenden Abbildungen sind Neuronen, in denen Reaktionen eingetragen werden, im Unterschied zu leeren Neuronen schwarz markiert. Nach 500 Zyklen hat bereits eine Anpassung der Euklidischen Distanzen der meisten Neuronen stattgefunden. Die Objekte sind nicht mehr über das ganze Netz verteilt, sondern man erkennt bereits in diesem frühen Stadium des Trainings die Ausbildung zusammengehöriger Neuronenbereiche (siehe Abbildung 4-2b).

Abb. 4-2: Entwicklung der Netzanpassung mit Zunahme der durchlaufenen Zyklen. Die Anzahl der Zyklen ist unter den Karten angegeben.

Nach 5.000 Zyklen sind keine mit den anfänglichen Zufallszahlen initialisierte Euklidischen Distanzen mehr vorhanden. Die zusammengehörigen Bereiche sind nun über die gesamte Fläche verteilt, sind aber noch sehr weit voneinander separiert (siehe Abbildung 4-3a). Die zehnfache Anzahl an Lernzyklen ist erforderlich, um diese Lücken weitgehend zu schließen, d.h. die zusammengehörigen Bereiche breiten sich aus ohne mit benachbarten Bereichen zu verschmelzen. Die Lernrate und die Spannweiten wurden inzwischen so stark verringert, daß die kleinen Veränderungen der Euklidischen Distanzen nur noch innerhalb zusammengehöriger Bereiche Auswirkungen zeigen (siehe Abbildung 4-3b).

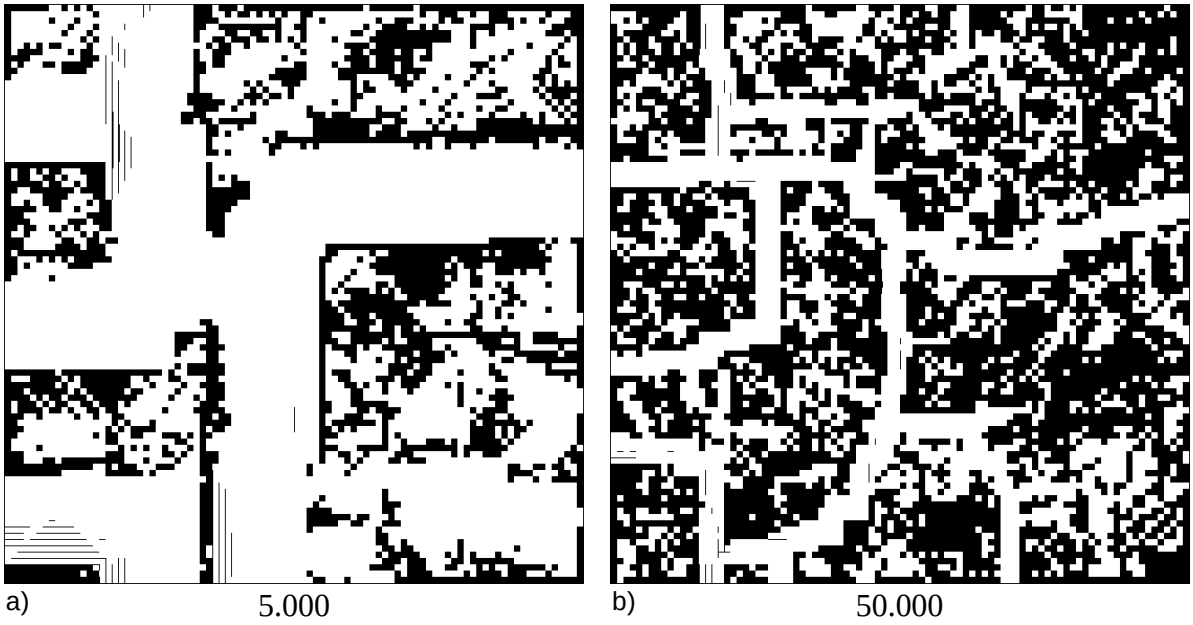


Abb. 4-3: Entwicklung der Netzanpassung mit Zunahme der durchlaufenen Zyklen. Die Anzahl der Zyklen ist unter den Karten angegeben.

Nach 72.800 Zyklen unterschreiten die Lernrate und die Spannweite den als Standardwert eingestellten Grenzwert von je 0,1 und der Trainingsvorgang wird abgebrochen.

Im Anhang A.2 sind die Rechenzeiten, die für die Extraktion der Reaktionen, deren Codierung und Klassifizierung nötig sind, exemplarisch für die Theilheimer Reaktionsdatenbank angegeben.

Die resultierende Kohonen-Karte ist in Abbildung 4-4 dargestellt. Aufgrund der Größe der Kohonen-Karte und der Größe des Datensatzes wird hier auf eine detaillierte Angabe der Reaktionsnummern für jedes Neuron verzichtet. Statt dessen ist die Anzahl der eingetragenen Reaktionen farblich abgestuft dargestellt. Die Kohonen-Karte ist im World-Wide-Web einschließlich aller eingetragener Reaktionen einsehbar (siehe Anhang A.1).

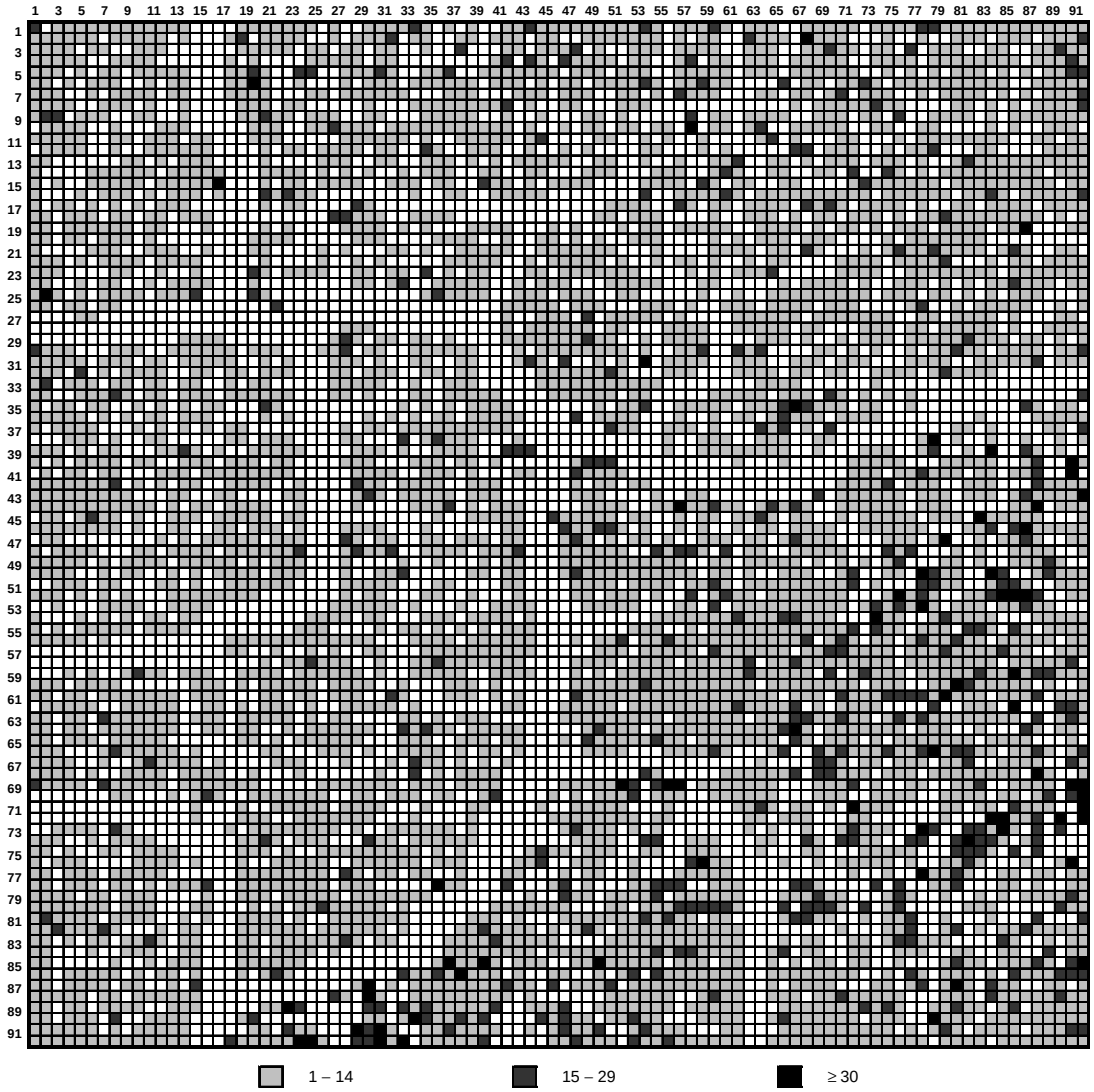


Abb. 4-4: Kohonen-Karte der klassifizierten 33.613 Reaktionen aus der Theilheimer Datenbank. Die Anzahl der eingetragenen Reaktionsbeispiele ist farblich abgestuft.

Man erkennt in Abbildung 4-4 eine Abgrenzung verschiedener Bereiche entlang der leeren Neuronenbänder. Die Euklidischen Distanzen, die während der Trainingsphase berechnet werden, bestätigen dabei die Abtrennung der Bereiche entlang dieser Linien. Im Hinblick auf die Verteilung der Reaktionsbeispiele fällt ein Bereich im unteren rechten Teil der Karte auf, in dem viele Neuronen mehr als 14 Reaktionen enthalten. Weniger häufig enthalten Neuronen im rechten oberen Bereich der Karte mindestens 15 Reaktionen. In den anderen Bereichen der Kohonen-Karte sind nur vereinzelt Neuronen mit einer höheren Anzahl an Reaktionsbeispielen zu finden. Aufgrund der großen Ausdehnung und der Vielzahl an Reaktionsbeispielen im unteren rechten Teil der Karte vermutet man dort einen Bereich, in dem bevorzugt Reaktionen eingetragen werden, bei denen nur eine einzige Bindung aufgebaut wird (siehe Abbildung 3-5). Eine genauere Analyse der Reaktionsbeispiele in der Kohonen-Karte ergibt in der Tat eine Separation der Reaktionen nach der Größe der Reaktionszentren auf der Produkt-

seite. Angesichts der Codierungsmerkmale der entwickelten Methode (siehe Kapitel 3.3.1) war solch eine primäre Einteilung der Reaktionen auch zu erwarten. Mit einer vorgegebenen maximalen Größe von sechs Bindungen im Reaktionszentrum rechnet man mit einer Aufteilung in sechs gut voneinander abgetrennte Bereiche. Die Kohonen-Karte der Abbildung 4-4 zeigt eine Einteilung in mehrere Bereiche, wobei größere Bereiche nochmals unterteilt sind. Diese Unterteilung wird durch die Bindungsordnung der Bindungen im Reaktionszentrum verursacht. Aus diesem Grund wurde die Kohonen-Karte der Abbildung 4-4 nach der Zusammensetzung der Reaktionszentren eingefärbt. Das Ergebnis ist in Abbildung 4-5 dargestellt.

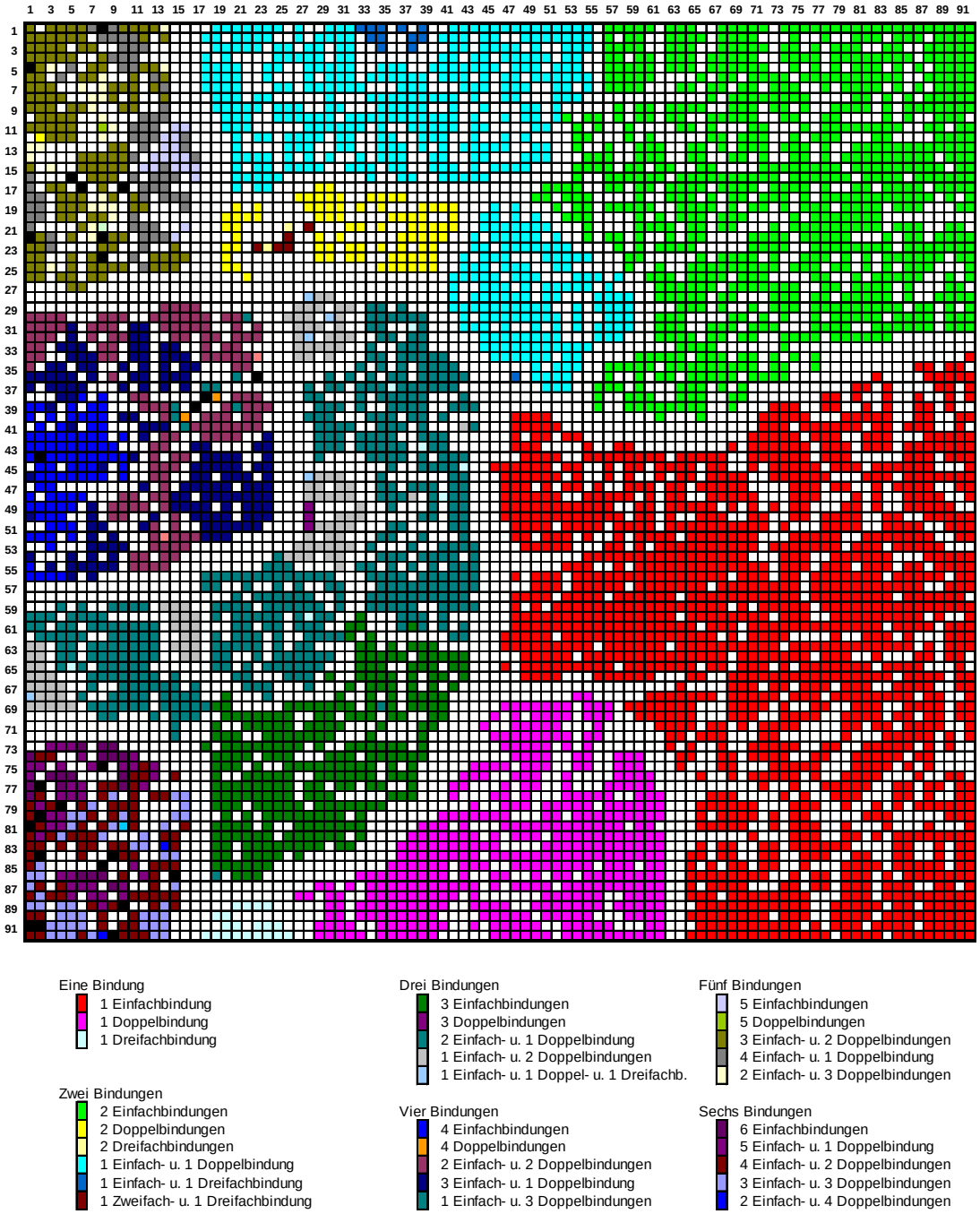


Abb. 4-5: Kohonen-Karte der Abbildung 4-4 eingefärbt nach der Zusammensetzung der Reaktionszentren.

Anhand der Abbildung 4-5 erkennt man sehr gut die Einteilung der Reaktionen nach der Größe der Reaktionszentren und den enthaltenen Bindungsordnungen. Die Größe der von verschiedenen Reaktionstypen eingenommenen Bereiche korreliert mit der Anzahl der Beispiele für jeden Reaktionstyp (vergleiche Kapitel 2.5.1). Das größte Gebiet nehmen rechts unten beispielsweise die Reaktionen ein, bei denen nur eine einzige Einfachbindung im Reaktionszentrum auf der Produktseite gebildet wurde.

4.2.1 Interpretation der klassifizierten Reaktionsdatenbank

Nach diesem generellen Blick auf die in Abbildung 4-5 dargestellte Kohonen-Karte wird im folgenden Abschnitt auf einzelne Gebiete näher eingegangen. Es werden die Bereiche näher diskutiert, in denen Reaktionen projiziert werden, die genau eine Einfach-, Zweifach- oder Dreifachbindung aufweisen. Diese Gebiete sind dem unteren rechten und unteren mittleren Teil der Karte entnommen und in den Abbildungen 4-6 und 4-9 dargestellt. In diesen Abbildungen ist für jedes Neuron das häufigste Reaktionszentrum als SMILES-Code eingetragen.

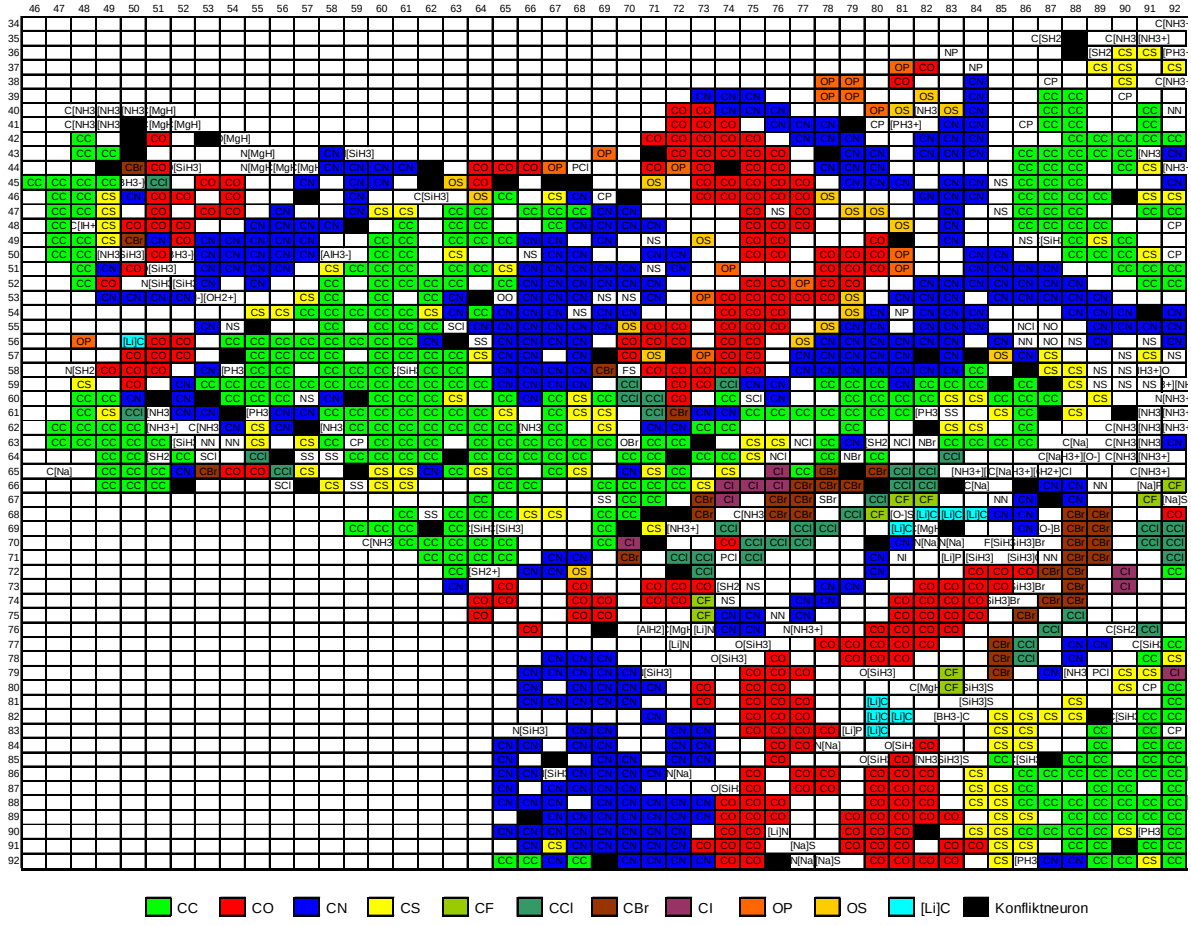


Abb. 4-6: Ausschnitt aus der Kohonen-Karte der Abbildung 4-5: Für jedes Neuron ist das häufigste Reaktionszentrum mit nur einer Einfachbindung auf der Produktseite als SMILES-Code eingetragen.

Man erkennt in Abbildung 4-6, daß in verschiedenen Bereichen der Kohonen-Karte jeweils ein Bindungstyp eingetragen wird. So findet man beispielsweise einen Bereich im mittleren unteren Teil der Karte, in dem Reaktionen eingetragen werden, bei denen Kohlenstoff-Stickstoff-Bindungen geknüpft werden. Allerdings gibt es meist keine zusammenhängenden Bereiche für je einen Bindungstyp. Im Falle der Kohlenstoff-Stickstoff-Bindungen sind weitere Bereiche im oberen Teil der Kohonen-Karte lokalisiert, die oft voneinander abgetrennt sind.

Die Aufteilung eines Bindungstyps in mehrere Teilgebiete wird verständlich, wenn man die Umgebung der geknüpften Bindung näher untersucht. Dazu wird für jeweils einen abgetrennten Bereich die strukturelle Umgebung der geknüpften Bindung dargestellt (siehe Abbildung 4-7). Dabei wird die strukturelle Umgebung durch die angegebenen Atomtypen oder funktionellen Gruppen näher charakterisiert.

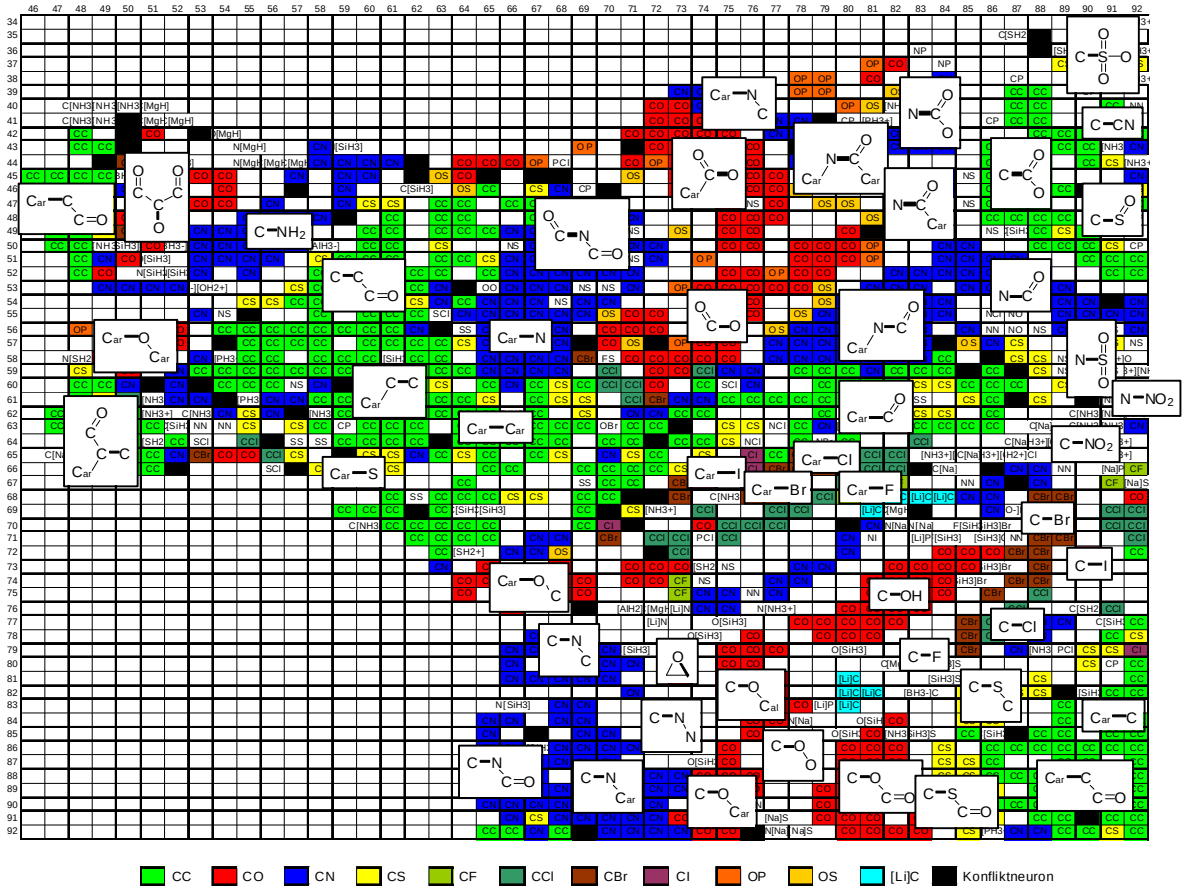


Abb. 4-7: Kohonen-Karte der Abbildung 4-6 mit eingezeichneter struktureller Umgebung des Teils des Reaktionszentrums auf der Produktseite. Die geknüpften Bindungen sind fett hervorgehoben; mit einem C wird ein aliphatisches, mit C_{ar} ein aromatisches System symbolisiert.

Da die Codierung der Reaktionen nicht auf strukturellen Eigenschaften des Reaktionszentrums, sondern auf physikochemischen Effekten der geknüpften Bindungen basiert, wurde für jeden Bereich der am häufigsten eingetragene Bindungstyp dargestellt. Dieser steht häufig stellvertretend für weitere Bindungstypen, die ähnliche physikochemische Eigenschaften

besitzen. So besitzen die in Abbildung 4-8 dargestellten funktionellen Gruppen ähnliche physikochemische Eigenschaften, obwohl sie sich in den Atomtypen bzw. funktionellen Gruppen unterscheiden.

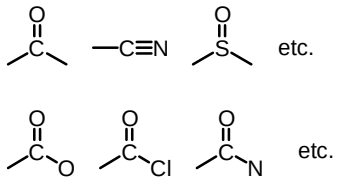


Abb. 4-8: Funktionelle Gruppen mit unterschiedlicher chemischer Struktur, aber ähnlichen physikochemischen Eigenschaften.

Die in dieser Arbeit entwickelte Klassifizierung bildet also unter Umständen strukturell verschiedene, aber physikochemisch ähnliche Bindungstypen in jeweils einem Bereich ab. In der in Abbildung 4-7 dargestellten Kohonen-Karte kann man zunächst eine Unterteilung in Bindungsknüpfungen an aromatischen oder aliphatischen Systemen erkennen. Während im oberen linken und oberen mittleren Teil der Kohonen-Karte meist Reaktionen eingetragen sind, bei denen Bindungen an aromatischen Systemen aufgebaut werden, so werden die analogen Reaktionen an aliphatischen Systemen im unteren Teil eingetragen. Aus diesem Grund sind beispielsweise Reaktionen, bei denen eine Halogen-Kohlenstoff-Bindung aufgebaut wird, in zwei voneinander getrennten Bereichen lokalisiert. Ein Bereich in der Mitte des Kohonen-Kartenausschnitts enthält Reaktionen, in denen Halogene an aromatischen Systemen beispielsweise über Sandmeyer-Reaktionen eingeführt werden. Im zweiten Bereich rechts unterhalb der Mitte findet man Reaktionen, in denen Halogene an aliphatischen Systemen beispielsweise durch Finkelstein-Reaktionen etc. synthetisiert werden. In beiden Bereichen findet zusätzlich noch eine Unterteilung nach dem eingeführten Halogen, wie C-I, C-Br etc., statt. Dieses Beispiel zeigt auch, daß aus dem Blickwinkel der Synthese eine Zusammenfassung aller Bindungstypen mit denselben Elementen, wie alle Halogen-Kohlenstoff-Bindungen an aliphatischen und aromatischen Systemen, chemisch nicht sinnvoll ist. Somit wird die Aufteilung in verschiedene, voneinander getrennte Bereiche verständlich.

Außerdem zeigt eine Analyse der Reaktionstypen mit nur einer Einfachbindung in Abbildung 4-7, daß vor allem die Kohlenstoff-Kohlenstoff-verknüpfenden Reaktionen viele Neuronen im oberen linken Teil der Kohonen-Karte belegen. Zum einen wird dies durch eine Vielzahl an Reaktionsbeispielen verursacht, andererseits auch durch eine große Variationsbreite bezüglich der auftretenden physikochemischen Effekte. In Abbildung 4-7 sind einige Bindungstypen, die Kohlenstoff-Kohlenstoff-Bindungen aufbauen, herausgestellt.

Auffallend in der Abbildung 4-7 sind außerdem Reaktionen, bei denen Bindungen zu Schwefelatomen aufgebaut werden. Im oberen Teil der Karte findet man häufiger C-S-Bindungen in Nachbarschaft zu C-C-Bindungen als in separaten Bereichen, in denen nur Reaktionen, an denen C-S-Bindungen beteiligt sind, eingetragen sind. Dies zeigt, daß es häufig zu

einem Bindungstyp, an dem zwei Kohlenstoffatome beteiligt sind, auch einen entsprechenden Typ gibt, bei dem die Reaktion an einem Kohlenstoff- und einem Schwefelatom abläuft. Diese Reaktionen sind zu den entsprechenden Reaktionen mit einer gebildeten C-C-Bindung ähnlicher als die Reaktionen mit C-S-Bindungen untereinander. Beispielsweise sind in dem Bereich im rechten oberen Eck des Kartenausschnitts Reaktionen eingetragen, bei denen sowohl eine C-C-, als auch eine C-S-Bindung geknüpft werden. Im Falle der C-C-verknüpfenden Reaktionen wird eine Kohlenstoff-Bindung zu dem Carbonylkohlenstoffatom einer Estergruppe aufgebaut, während bei den C-S-verknüpfenden Reaktionen das Kohlenstoffatom mit einer Sulfoxidgruppe reagiert.

Dagegen werden im unteren rechten Teil der Karte die Reaktionen, an denen Bindungen zu Schwefelatomen geknüpft werden, getrennt von C-C-Bindungen in eigene Bereiche eingetragen. Als Beispiele sind in Abbildung 4-7 die Alkylierungsreaktionen am Schwefelatom einer Thiocarbonsäure gezeigt, die in die Neuronen (84,86) bis (85,92) eingetragen sind, oder die Reaktionen zur Bildung eines Thioethers, die in den Neuronen (85,82) bis (86,84) lokalisiert sind.

Zusammenfassend kann man für diesen Kartenausschnitt feststellen, daß im oberen Bereich der Kohonen-Karte Reaktionen eingetragen werden, bei denen das Kohlenstoffatom der geknüpften Bindung meist direkt in einem π-System eingebunden ist, wie beispielsweise die C-N-Bindung in einem Säureamid oder in einem Anilinderivat. Dagegen sind im unteren Bereich der Karte meist Reaktionen lokalisiert, bei denen das Heteroatom der aufgebauten Bindung in Konjugation zu einem π-System steht. Beispiele sind die Alkylierungsreaktionen an den Heteroatomen von Phenol- oder Anilinderivaten.

Gut abgetrennt vom Gebiet der Reaktionen mit genau einer Einfachbindung liegt links ein Bereich, in dem Reaktionen mit genau einer Doppel- oder Dreifachbindung im Reaktionszentrum auf der Produktseite eingetragen sind. Dieser Kohonen-Kartenausschnitt ist in Abbildung 4-9 dargestellt, wobei auch hier der häufigste Reaktionstyp pro Neuron eingetragen ist.

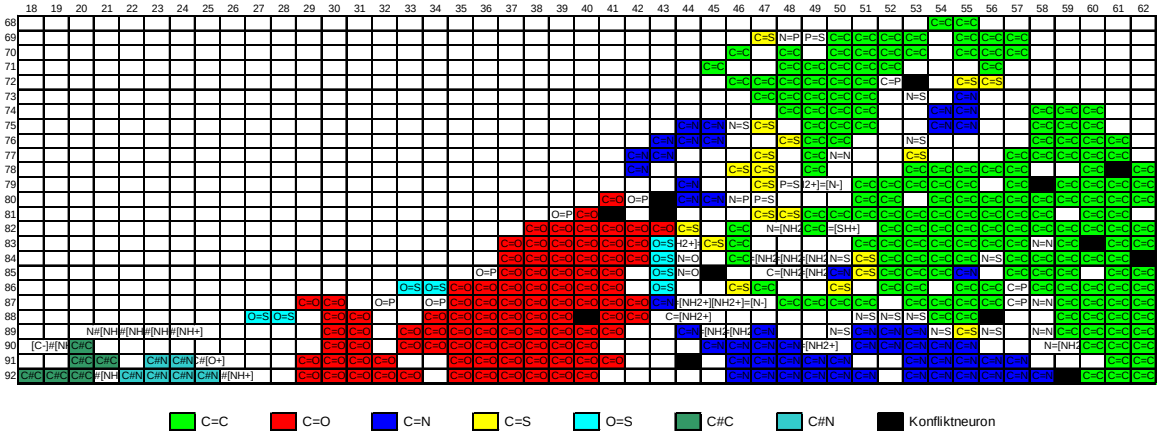


Abb. 4-9: Ausschnitt aus der Kohonen-Karte der Abbildung 4-5: Für jedes Neuron ist das häufigste Reaktionszentrum mit nur einer Mehrfachbindung auf der Produktseite als SMILES-String eingetragen.

Man erkennt, daß die Dreifachbindungen gut abgetrennt von den Zweifachbindungen projiziert werden. Innerhalb des Gebietes der Dreifachbindungen sind die C≡C-Dreifachbindungen wiederum von den C≡N-Dreifachbindungen separiert.

Im Bereich der Zweifachbindungen sind die einzelnen Reaktionstypen, also beispielsweise C=O, C=N, C=S und C=C, in getrennten Bereichen lokalisiert. Dabei nimmt die Polarität der C=C-Bindungen, über die C=S- und C=N-Bindungen bis zu den C=O-Bindungen von rechts nach links immer stärker zu. Analog zu den Reaktionen, bei denen eine Einfachbindung gebildet wird, sind auch hier Reaktionen mit dem gleichen Bindungstyp in separaten Gebieten eingetragen (siehe Abbildung 4-10). Allerdings sind wegen der Doppel- oder Dreifachbindung in diesem Fall weniger Variationsmöglichkeiten der angrenzenden funktionellen Gruppen möglich. Wie schon im Gebiet der Einfachbindungen durchsetzen die Reaktionszentren, bei denen Schwefel beteiligt ist, den Bereich der C=C-Bindungen.

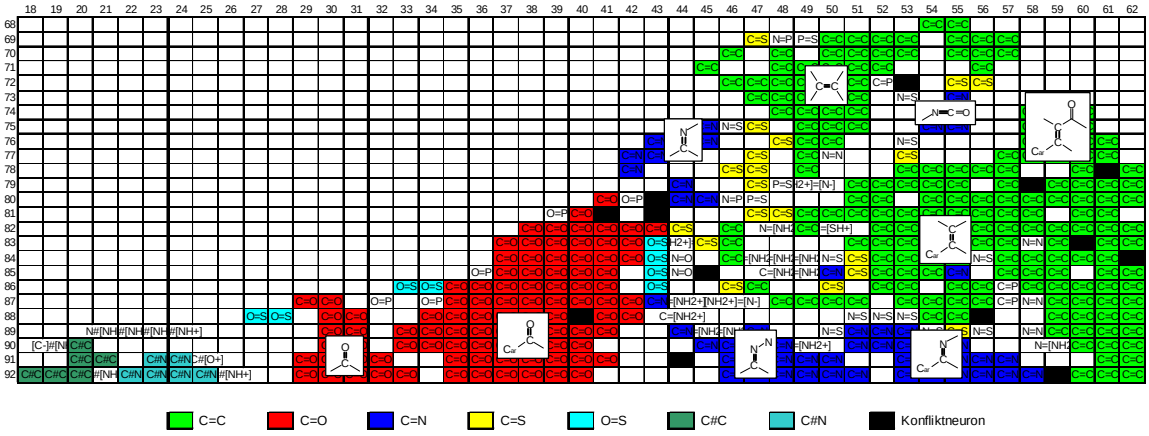


Abb. 4-10: Kohonen-Karte der Abbildung 4-9 mit eingezeichneter struktureller Umgebung des Teils des Reaktionszentrums auf der Produktseite. Die geknüpften Bindungen sind fett hervorgehoben; mit einem C wird ein aliphatisches, mit C_{ar} ein aromatisches System symbolisiert.

Wie in den vorangegangenen Abschnitten gezeigt wurde, werden die Reaktionen einer Datenbank zunächst nach der Anzahl der Bindungen im Reaktionszentrum klassifiziert. In der klassifizierten Kohonen-Karte sind diese Bereiche durch hohe Euklidische Distanzen bzw. leere Neuronenreihen voneinander abgetrennt. Innerhalb dieser Bereiche sind meist die einzelnen Reaktionszentren wiederum in Teilbereiche untergliedert. Sowohl diese Teilbereiche, als auch die einzelnen Neuronen liefern wertvolle Informationen für chemische Anwendungen. Wie am Beispiel der Reaktionstypen mit nur einer Bindung im Reaktionszentrum auf der Produktseite gezeigt wurde, können für die einzelnen Teilbereiche Bindungstypen angegeben werden. Anhand dieser Bindungstypen erscheint die Klassifizierung auch in chemischer Hinsicht überzeugend, da die Einteilung auch mit verschiedenen Reaktionsmechanismen oder Reaktionsdurchführungen erklärt werden kann.

4.2.2 Ähnliche Reaktionstypen

Chemiker teilen Reaktionen gewöhnlich nicht nach der Größe der Reaktionszentren und den enthaltenen Bindungsordnungen ein, sondern aus historischen oder mnemotechnischen Gründen nach dem Namen des Entdeckers einer Reaktion, wie Wittig- oder Michael-Reaktion[7],[65]. Manche Reaktionen werden auch nach den Edukten oder Produkten benannt, wie Glykolspaltung oder Aldolreaktion. Diese Einteilung organischer Reaktionen in Namensreaktionen ist jedem Chemiker wohl vertraut. Daher wird bei der Aufnahme einer chemischer Reaktion in eine Datenbank die entsprechende Bezeichnung der Namensreaktion oft mit abgespeichert. In der codierten Theilheimer Reaktionsdatenbank findet man 2.475 (7,4%) Namenseinträge. Aufgrund der Codierung der Reaktionen nach physikochemische Effekten gehören die in einem Neuron enthaltenen Reaktionen oft verschiedenen Reaktionstypen an, wobei meistens eine Namensreaktion überwiegt. Aus dieser physikochemischen Ähnlichkeit der gebildeten Bindungen zweier Namensreaktionen kann man wertvolle Information gewinnen, die man beispielsweise gut in der Syntheseplanung anwenden kann (siehe Kapitel 5). Im Falle der klassifizierten Theilheimer Reaktionsdatenbank ist beispielsweise in Neuron (46,34) die Wittig-Reaktion die häufigste Namensreaktion. In diesem Neuron werden zum einen die beiden Wittig-Reaktionen #3409 (siehe Abbildung 4-11) und #17112 (siehe Abbildung 4-12) eingetragen.

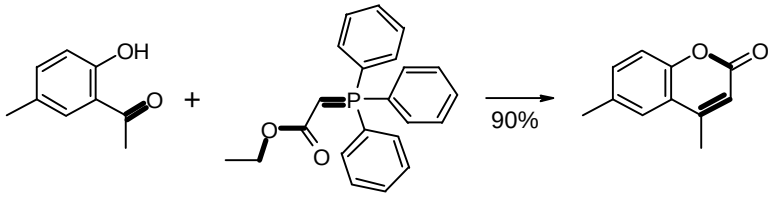


Abb. 4-11: Wittig-Reaktion #3409 (RTHE00045195) aus Neuron (46,34) der Abbildung 4-5. Die Reaktionsgleichungen sind dem ISIS Retrieval-System von MDL (Version 2.2.1) entnommen.

Bei beiden Reaktionen wird eine C-C-Doppelbindung neu geknüpft und außerdem die Estergruppe in eine Säuregruppe überführt. Deshalb findet man im Reaktionszentrum auf der Produktseite zwei Bindungen und die Reaktionen werden in die Mitte der Kohonen-Karte eingetragen.

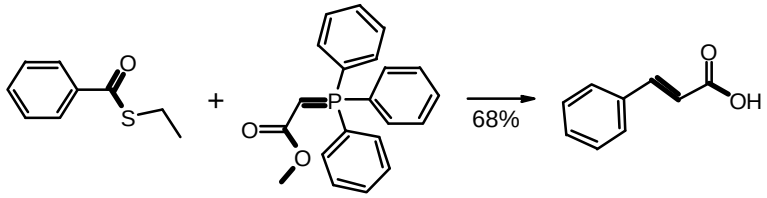


Abb. 4-12: Wittig-Reaktion #17112 (RTHE00030267) aus Neuron (46,34) der Abbildung 4-5.

Zum anderen beinhaltet dieses Neuron auch eine Horner-Synthese (#1896) und eine Favorskii Umlagerung (#20730). Wie bei der Wittig-Reaktion wird bei der Horner-Reaktion, auch als Horner-Emmons- oder Wittig-Horner-Reaktion bezeichnet, eine C=C-Doppelbindung aufgebaut; anstelle der Phosphonium-Ylide werden allerdings Phosphonsäuredialkylester eingesetzt (siehe Abbildung 4-13).

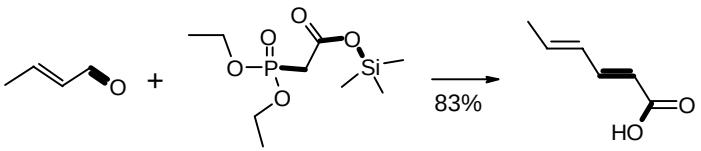


Abb. 4-13: Der zu Reaktion #3409 und #17112 ähnliche Reaktionstyp einer Horner-Reaktion: Reaktion #1896 (RTHE00045209) aus Neuron (46,34).

Bei der Favorskii-Umlagerung (siehe Abbildung 4-14) wird ebenfalls eine C=C-Doppelbindung und eine C-O-Einfachbindung wie bei den anderen Reaktionen aufgebaut, allerdings erhält man in diesem Falle nicht die Carbonsäure, sondern den Carbonsäureester, den man leicht in die entsprechende Säure überführen kann.

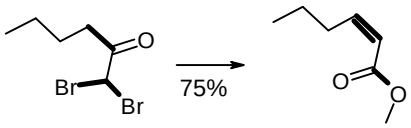


Abb. 4-14: Der zu Reaktion #3409 und #17112 ähnliche Reaktionstyp einer Favorskii-Umlagerung: Reaktion #20730 (RTHE00026722) aus Neuron (46,34).

Analoge Ähnlichkeiten zwischen verschiedenen Reaktionstypen kann man in Neuron (51,32) erkennen. In diesem Neuron sind zwei Wittig-Reaktionen #21501 und #21793 eingetragen, bei denen sowohl eine exocyclische C=C-Doppelbindung, als auch eine C-C-Ringschlußbindung aufgebaut werden (siehe Abbildung 4-15 und 4-16).

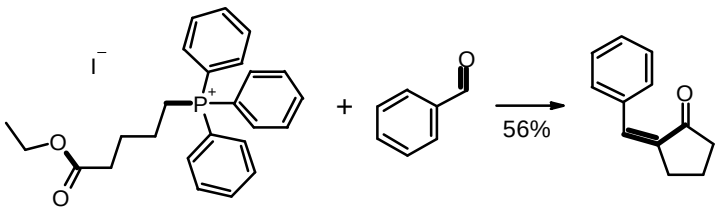


Abb. 4-15: Wittig-Reaktion #21501 (RTHE00027493) aus Neuron (51,32).

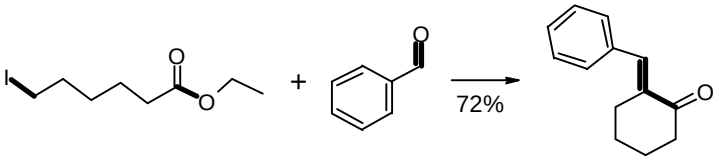


Abb. 4-16: Wittig-Reaktion #21793 (RTHE00024922) aus Neuron (51,32).

Im gleichen Neuron findet man auch eine Shapiro-Reaktion (#4732), die ebenfalls eine C=C-Doppel- und C-C-Einfachbindung ausgehend von Tosylhydrazonen aufbaut (siehe Abbildung 4-17).

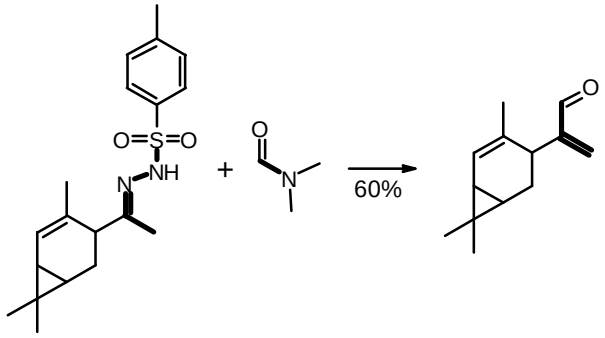


Abb. 4-17: Shapiro-Reaktion #4732 (RTHE00043558) aus Neuron (51,32).

Schließlich sei noch Neuron (55,86) angeführt, in dem die Ähnlichkeit einer intramolekularen Wittig-Reaktion (#1317) mit einer Kondensationsreaktion nach Knoevenagel (#21015) deutlich wird (siehe Abbildung 4-18).

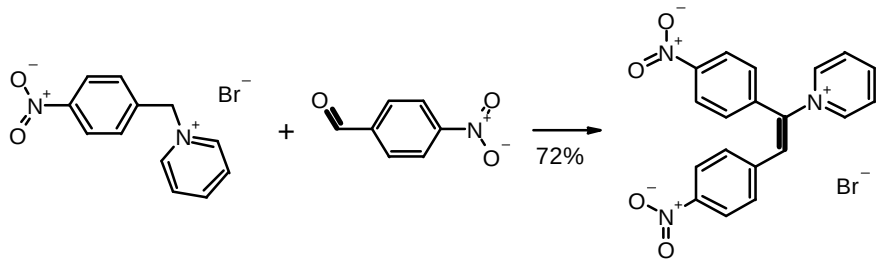


Abb. 4-18: Knoevenagel-Kondensation #21015 (RTHE00026248) aus Neuron (55,86).

Im Anhang A.3 sind für jede Namensreaktion in der codierten Theilheimer Reaktionsdatenbank die zugehörigen, ähnlichen Reaktionstypen tabellarisch aufgeführt.

Aus dieser Vielzahl an „verwandten“ Namensreaktionen werden noch zwei Reaktionstypen exemplarisch vorgestellt. Erstens der Hoffmannsche Abbau, der physikochemisch ähnliche Bindungen aufbaut wie die Gabriel-Synthese, die Smiles-Umlagerung, die Schmidt-Reaktion, die Leuckart-Reaktion, oder die Stephen-Reduktion in den Neuronen (57,48), (68,71) und (78,74). All diesen Reaktionstypen ist der Aufbau einer Kohlenstoff-Stickstoff-Bindung gemeinsam. Der zweite Reaktionstyp baut ebenfalls eine Kohlenstoff-Stickstoff-Bindung, aber ausgehend von Carbonylverbindungen, auf. Hierzu zählt die Lossen-Umlagerung, die Curtius-Umlagerung und die Schmidt-Reaktion in den Neuronen (79,16) und (83,16).

4.3 Vergleich der Theilheimer und der SPORE Datenbank

Vor allem in neuerer Zeit hat die Bedeutung organischer Reaktionen, die an fester Phase durchgeführt werden, enorm zugenommen. Festphasenreaktionen zeigen vor allem in der kombinatorischen Chemie eine Reihe von Vorteilen gegenüber den in flüssiger Phase durchgeführten Reaktionen. Es können beispielsweise Reagenzien ohne spätere Abtrennprobleme im Überschuß eingesetzt werden, die Produkte können leicht gereinigt und isoliert werden usw. Aus diesen Gründen ist es vor allem in der kombinatorischen Chemie oft wünschenswert, die Übertragbarkeit einer in einem Lösungsmittel durchgeführten Reaktion in eine an fester Phase ablaufenden Reaktion abzuschätzen. Andererseits ist man auch an einem Überblick über die an fester Phase durchführbaren Reaktionen interessiert. Beide Problemstellungen können mit der Reaktionsklassifizierung angegangen werden.

Dabei kommt die in Kapitel 4.1 beschriebene Methode des Datenbankenvergleichs zum Einsatz. Die in Kapitel 4.2 klassifizierte Theilheimer Datenbank dient dabei als Referenz beim Vergleich mit der SPORE Reaktionsdatenbank.

Zunächst soll die Verteilung der Größe der Reaktionszentren in der SPORE Reaktionsdatenbank mit der bereits in Kapitel 3.3.3 diskutierten Verteilung für die Theilheimer Reaktionsdatenbank verglichen werden. In Abbildung 4-19 ist die Verteilung der Bindungsanzahl in den Reaktionszentren auf der Produktseite für alle Reaktionen aus der SPORE Datenbank dargestellt. Während die Reaktionen aus der Theilheimer Reaktionsdatenbank mit nur einer einzigen Bindung im Reaktionszentrum auf der Produktseite einen Anteil von 38,2% ausmachen, so sind es im Falle der SPORE Reaktionsdatenbank 55,1%. Dagegen findet man in der SPORE Datenbank nur 24,9% Reaktionen, deren Reaktionszentrum aus mehr als 1 Bindung auf der Produktseite besteht, während es im Falle der Theilheimer Reaktionsdatenbank 53,3% sind. Da in der SPORE Reaktionsdatenbank bevorzugt einfache Synthesen kleiner Moleküle aufgenommen werden (siehe Kapitel 2.2.3), ist dieser Unterschied leicht verständlich.

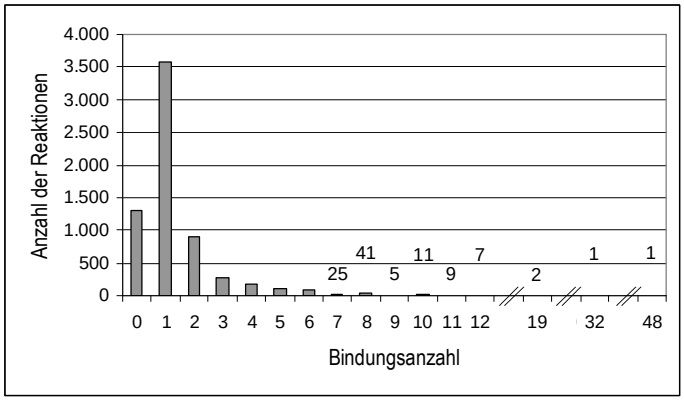


Abb. 4-19: Histogramm zur Reaktionszentrengröße. Für alle Reaktionen der SPORE Reaktionsdatenbank ist die Anzahl der am Bindungsumordnungsprozeß beteiligten Bindungen im Produktensemble ermittelt worden.

Die SPORE Reaktionsdatenbank wird nun auf dieselbe Methode wie die Theilheimer Reaktionsdatenbank codiert, d.h. Reaktionen mit maximal sechs Bindungen im Reaktionszentrum auf der Produktseite werden herangezogen. Von diesen Bindungen werden die sechs physikochemischen Effekte des Standardverfahrens berechnet, die als Eingabevektoren für ein neuronales Netz dienen. Von den insgesamt 6.502 Reaktionen haben 5.048 Reaktionen eine bis sechs Bindungen im Reaktionszentrum auf der Produktseite. PETRA kann für 320 Reaktionen keine physikochemischen Deskriptoren berechnen und bei 274 Reaktionen wirken sich fehlende Atom-Atom-Mapping-Nummern in der Datenbank aus. Somit gelangen 4.454 (68,5%) Reaktionen in den codierten Datensatz der SPORE Datenbank. Mit diesem Datensatz wird aber kein neues Netz trainiert, sondern nur in das Referenznetzwerk der Theilheimer Reaktionsdatenbank projiziert. Das Ergebnis dieser Klassifizierung ist in Abbildung 4-20 dargestellt.

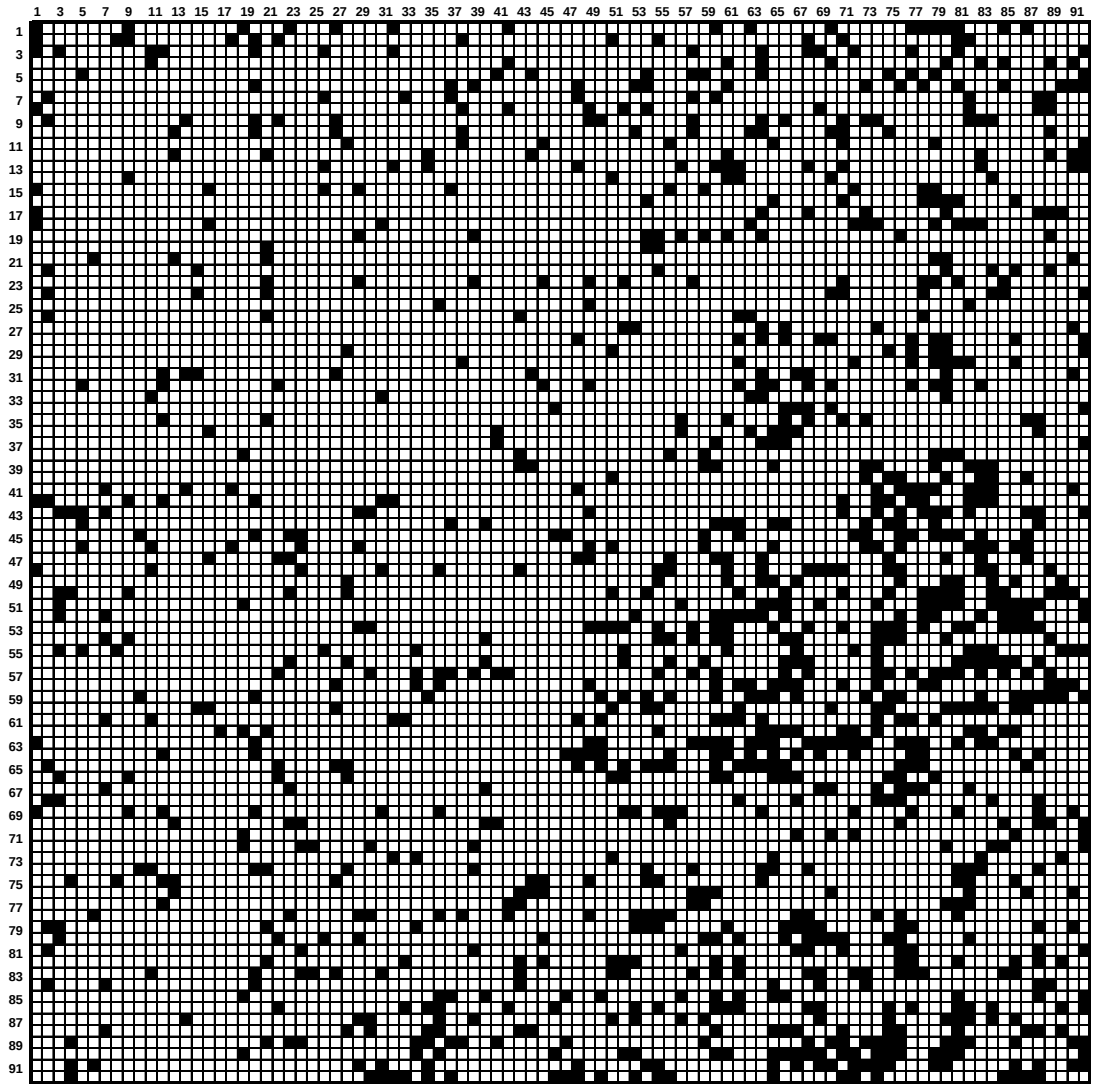


Abb. 4-20: Datensatz bestehend aus 4.454 Reaktionen aus der SPORE Reaktionsdatenbank, der in das Referenznetz der Theilheimer Reaktionsdatenbank (siehe Abbildung 4-5) projiziert wurde.

Der codierte Datensatz der SPORE Reaktionsdatenbank beträgt nur rund 13,3% bezogen auf den Datensatz aus der Theilheimer Reaktionsdatenbank. Angesichts dieses kleinen Datensatzes ist eine komplette Ausfüllung des Reaktionsraumes nicht zu erwarten. Abbildung 4-20 zeigt dies recht deutlich. Dem Histogramm der Abbildung 4-19 entsprechend liegt der größte Teil der Reaktionen im unteren rechten Bereich, in dem Reaktionen eingetragen werden, bei denen maximal eine Einfach- oder eine Mehrfachbindung im Reaktionszentrum auf der Produktseite reagiert. Allerdings ist die Verteilung der wenigen Reaktionen mit mehreren Bindungen im Reaktionszentrum überraschend: Diese füllen den Reaktionsraum in der kompletten Höhe und Breite aus. Die SPORE Reaktionsdatenbank verfügt also über eine annähernd gleich große physikochemische Variationsbreite wie die Theilheimer Reaktionsdatenbank. Somit kommt nicht nur bei einfachen Reaktionstypen, die maximal vier Bindungen im Reaktionszentrum haben, häufig eine alternative Reaktion an fester Phase in Frage, sondern auch bei Reaktionen mit größeren Reaktionszentren.

Ein detaillierter Vergleich der Theilheimer- und der SPORE Reaktionsdatenbank wird mit dem Kartenbereich durchgeführt, in dem Reaktionen eingetragen werden, bei denen nur eine Einfach- oder eine Mehrfachbindung im Reaktionszentrum auf der Produktseite gebildet wird. Dazu wird der entsprechende Ausschnitt der Kohonen-Karte in Abbildung 4-20 mit den am häufigsten je Neuron eingetragenen Reaktionszentren eingefärbt und in Abbildung 4-21 dargestellt.

Ein Vergleich der beiden Karten in den Abbildungen 4-6 und 4-21 zeigt, daß die Reaktionen aus der SPORE Datenbank diesen von der Theilheimer Reaktionsdatenbank aufgespannten Teilbereich nur teilweise ausfüllen. Die physikochemische Variationsbreite ist dabei genauso groß wie im Falle der Theilheimer Reaktionsdatenbank, da die SPORE Reaktionen das Gebiet in der gesamten Höhe und Breite ausfüllen. Für die wichtigsten Reaktionstypen, bei denen C-C-, C-O-, C-N-, C-S-, C-Cl-, C-Br-, und C-I-Bindungen aufgebaut werden, sind auch in der SPORE Reaktionsdatenbank Reaktionsbeispiele vertreten. In den Neuronen (84,68) und (81,82) sind sogar Reaktionen aus der SPORE Datenbank eingetragen, bei denen eine metallorganische Lithium-Kohlenstoff-Bindung aufgebaut wird.

Dagegen sind in dem codierten SPORE-Datensatz mit insgesamt 4.454 keine Reaktionsbeispiele enthalten, bei denen eine F-S-, N-I-, N-Cl-Bindung aufgebaut wird. Gleiches gilt auch für die Reaktionen, bei denen eine N-P-Einfachbindung gebildet wird. Diese sind in der Referenzkarte in den Neuronen (83,36) und (84,37) eingetragen worden. In dem codierten SPORE-Datensatz findet sich für diesen Reaktionstyp in den genannten Neuronen kein einziges Reaktionsbeispiel.

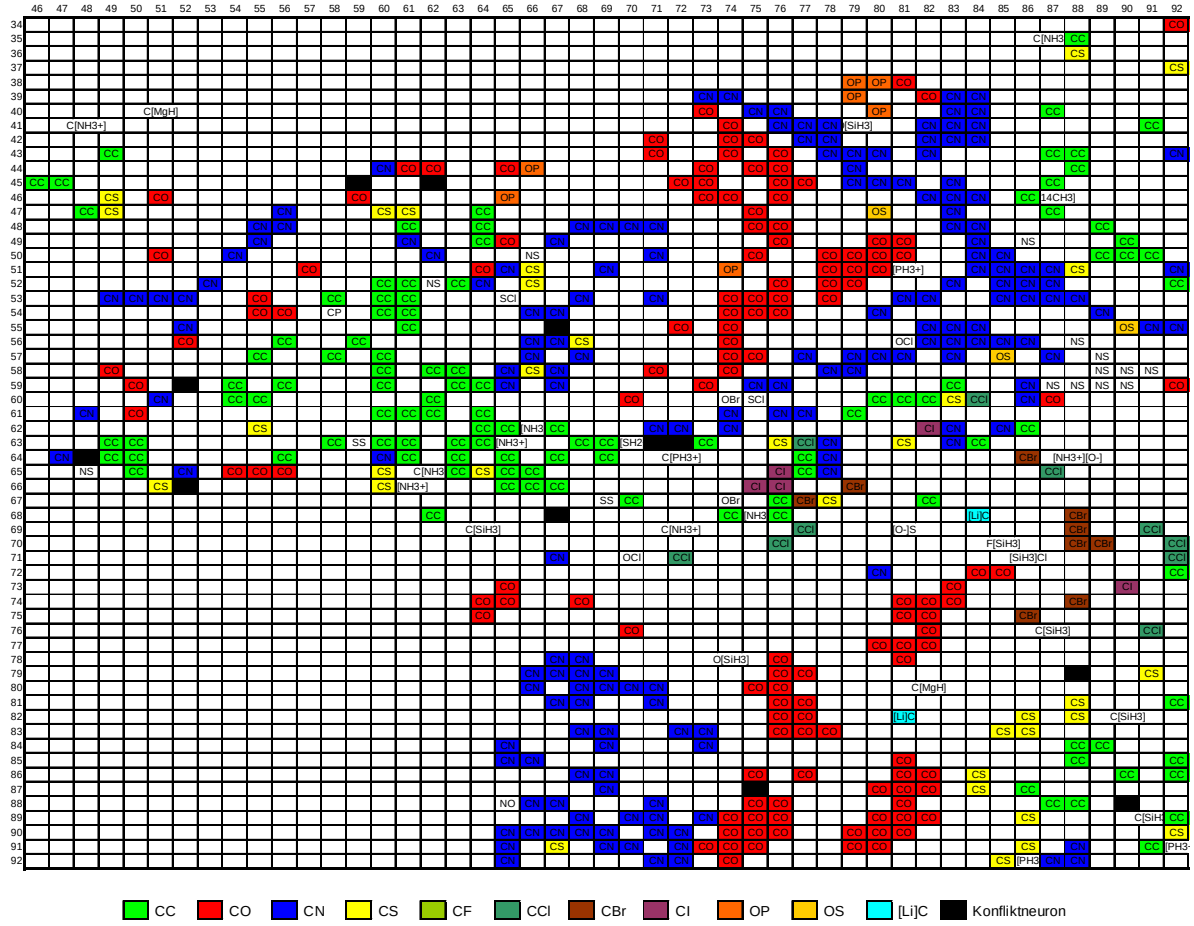


Abb. 4-21: Ausschnitt aus der Kohonen-Karte der Abbildung 4-20: Für jedes Neuron ist das häufigste Reaktionszentrum mit nur einer Einfachbindung auf der Produktseite als SMILES-Code eingetragen.

Auffallend ist außerdem das im Vergleich zur Theilheimer Reaktionsdatenbank größere Gebiet, in das Reaktionen eingetragen werden, bei denen N-S-Einfachbindungen aufgebaut werden. Das Gebiet, das sich um das Neuron (89,58) ausbreitet, enthält meistens Reaktionen, bei denen eine Sulfonamidoverbindung synthetisiert wird. Reaktion #3766 aus diesem Neuron ist stellvertretend in Abbildung 4-22 dargestellt.

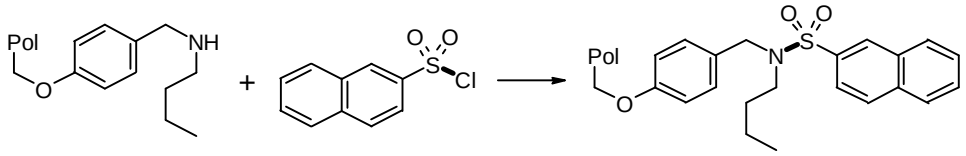


Abb. 4-22: Reaktion #3766 aus der SPORE Reaktionsdatenbank (RSPO69004172) aus Neuron (89,58). Die Abkürzung Pol steht für Polymer.

Der codierte Datensatz der SPORE Reaktionsdatenbank enthält prozentual berechnet rund doppelt so viele Reaktionsbeispiele dieses Typs wie die codierte Theilheimer Reaktionsdatenbank. Außerdem zeigen diese Reaktionsbeispiele aus der SPORE Datenbank eine größere

physikochemische Variationsbreite, so daß das größere, eingenommene Gebiet verständlich wird.

Schließlich wird noch das Gebiet, in denen Reaktionen eingetragen werden, bei denen nur eine Mehrfachbindung aufgebaut wird, verglichen. Der Kohonen-Kartenausschnitt mit dem häufigsten Reaktionstyp je Neuron ist in Abbildung 4-23 wiedergegeben.

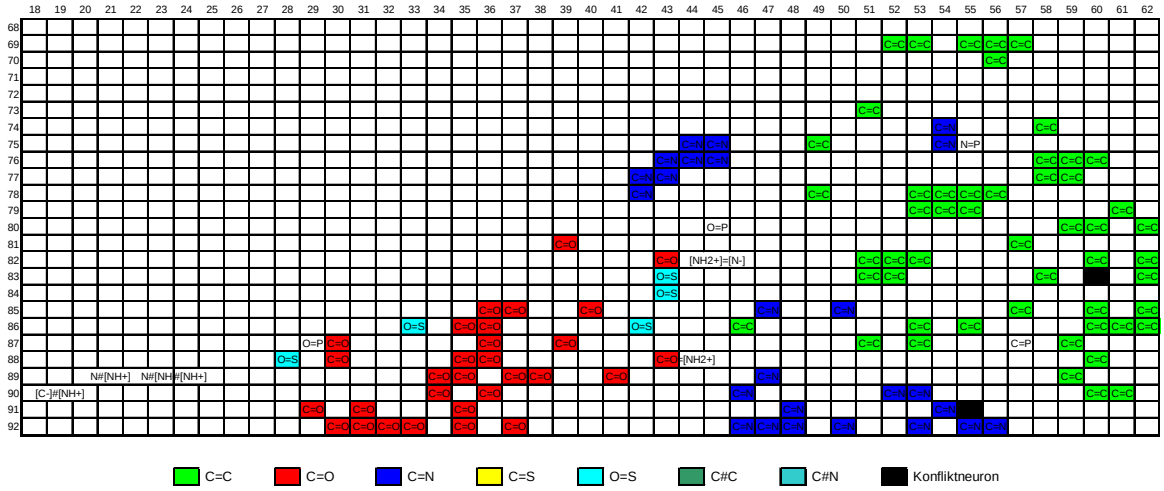


Abb. 4-23: Ausschnitt aus der Kohonen-Karte der Abbildung 4-20: Für jedes Neuron ist das häufigste Reaktionszentrum mit nur einer Mehrfachbindung auf der Produktseite als SMILES-String eingetragen.

Auch hier entspricht im wesentlichen die physikochemische Variationsbreite der codierten Reaktionen aus der SPORE Datenbank der durch die Theilheimer Reaktionsdatenbank vorgegebenen Breite. Unterschiede zeigen sich im Bereich der Doppelbindungen. Hier fällt zunächst die Abwesenheit von C=S-Bindungen auf. In der codierten SPORE Reaktionsdatenbank findet man keine einzige Reaktion, bei der eine Kohlenstoff-Schwefel-Doppelbindung aufgebaut wird. Des weiteren findet man auch keine Reaktion in den Neuronen auf der linken Seite, in denen im Referenznetzwerk C≡N-, und C≡C-Dreifachbindungen eingetragen sind.

Die aus Festphasenreaktionen bestehende SPORE Datenbank enthält nahezu alle Reaktionstypen, bei denen eine Einfach- oder Doppelbindung aufgebaut wird. Zu den wichtigsten Reaktionstypen, wie C-C, C-O, C-N, C-S, C-Hal (ohne C-F), C-Li, C=C, C=O und C=N, findet man auch in der SPORE Datenbank Reaktionsbeispiele. Die Synthese eines Moleküls nach diesen Reaktionstypen könnte also sowohl in flüssiger als auch an fester Phase erfolgen. Nur zu Reaktionen, bei denen eine C=S-Doppelbindung, eine C≡C- oder C≡N-Dreifachbindung gebildet wird, gibt es derzeit in der SPORE Reaktionsdatenbank noch keine Beispiele an fester Phase.

4.4 Zeitliche Entwicklung der ChemInform RX-Reaktionsdatenbank

Im vorangegangenen Kapitel wurde die SPORE Datenbank mit der Theilheimer Datenbank verglichen, um Unterschiede und Gemeinsamkeiten bei den Reaktionstypen einer Festphasenreaktionsdatenbank und einer Reaktionsdatenbank, die mit Reaktionen in Lösungsmitteln aufgebaut wurde, festzustellen. In diesem Kapitel wird dieselbe Vergleichsmethode eingesetzt, um die zeitliche Entwicklung von Reaktionsdatenbanken zu dokumentieren.

Für diesen Zweck bieten sich die ChemInform RX-Reaktionsdatenbanken an, da diese in mehreren Jahrgängen mit vergleichbarer Datenbankgröße vorliegt. In diesem Kapitel werden insgesamt sechs ChemInform RX-Reaktionsdatenbanken der Jahre 1992 bis 1997 mit der Theilheimer Reaktionsdatenbank als Referenzdatenbank verglichen. Die Codierung der Reaktionen erfolgt gemäß dem in Kapitel 3.3 beschriebenen Standardverfahren. Die erhaltenen Datensätze entnehme man Tabelle 4-2.

Reaktionsdatenbank	CIRX92	CIRX93	CIRX94	CIRX95	CIRX96	CIRX97
Reaktionen insgesamt	76.421	67.740	56.236	64.187	70.271	70.061
Reaktionen mit 0 oder mehr als 6 umgesetzten Bindungen	–27.355	–23.263	–19.284	–21.357	–22.326	–22.659
PETRA-Fehler	–3.498	–3.365	–2.364	–2.200	–2.198	–2.013
Datenbankfehler	–6.507	–6.617	–5.335	–6.118	–7.053	–6.588
codierte Reaktionen	39.061 (51,1%)	34.495 (50,9%)	29.253 (52,0%)	34.512 (53,8%)	38.694 (55,1%)	38.801 (55,4%)

Tab. 4-2: Die codierten ChemInform RX-Reaktionsdatenbanken. Die Fehlerquellen, die zu einer Verminderung der codierbaren Reaktionen führen, sind in Kapitel 4.2 beschrieben.

Zur Beschreibung der zeitlichen Entwicklung projiziert man jeweils eine codierte ChemInform RX-Datenbank in das trainierte Kohonen-Netz der Abbildung 4-4 und berechnet für jedes Neuron den Anteil der eingetragenen Reaktionen zur Gesamtreaktionsanzahl. Danach trägt man für alle 92x92 Neuronen die sechs Datenpunkte auf und ermittelt die entsprechende Regressionsgerade. Somit weisen Reaktionstypen, deren Reaktionsbeispiele über die Jahre annähernd konstant bleiben, eine flache Regressionsgerade auf. Andererseits sind Reaktionstypen, die in neuerer Zeit intensiv untersucht wurden, mit vielen Reaktionsbeispielen in den neueren Reaktionsdatenbanken vertreten, so daß in diesem Fall eine große Regressionsgeradensteigung für die entsprechenden Neuronen festgestellt werden kann. In Abbildung 4-24 ist die Auftragung der Datenpunkte sowie die berechnete Regressionsgerade exemplarisch für Neuron (84,39) gezeigt.

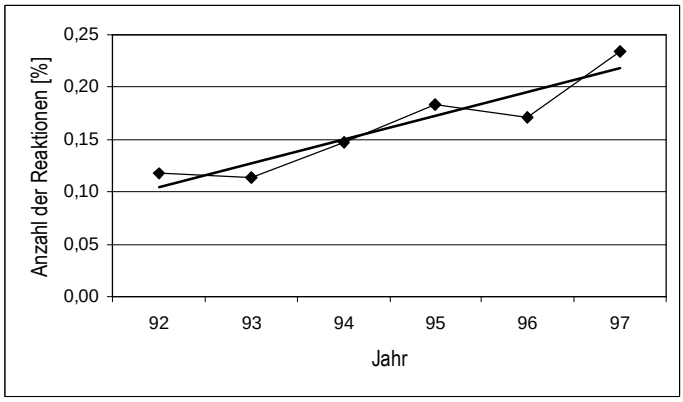
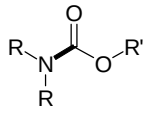


Abb. 4-24: Prozentuale Anzahl der in Neuron (84,39) eingetragenen Reaktionen in den sechs ChemInform RX-Reaktionsdatenbanken der Jahre 1992-97. Die Regressionsgerade der sechs Datenpunkte ist ebenfalls eingezeichnet.

Um die Übersichtlichkeit zu wahren, werden im folgenden nur die Kartengebiete näher diskutiert, in denen Reaktionen mit genau einer Einfachbindung im Reaktionszentrum auf der Produktseite eingetragen werden. Außerdem sind in der Abbildung 4-26 nur die besetzten Neuronen mit Symbolen gekennzeichnet, für die eine Regressionsgeradensteigung größer als 0,05 oder kleiner als –0,05 ermittelt wurde.

Für das bereits oben genannte Neuron (84,39) wurde die größte Regressionsgeradensteigung mit einem Wert von 0,023% ermittelt. In diesem Neuron sind in der Referenzkarte 41 Reaktionen aus der Theilheimer Reaktionsdatenbank eingetragen worden, wobei bei diesen eine C-N-Einfachbindung aufgebaut wird, die wiederum Bestandteil einer Carbaminsäure oder eines Carbaminsäureesters ist (siehe Abbildung 4-25).



wobei R = H, Alkyl-
R’ = H, Alkyl- o. Aryl-

Abb. 4-25: Strukturelles Merkmal der Reaktionen in Neuron (84,39): Aufbau einer Carbaminsäure oder eines Carbaminsäureesters.

Im Falle der ChemInform RX-Reaktionsdatenbanken werden in dieses Neuron ebenfalls Carbaminsäure(ester) eingetragen, die häufig aus kombinatorischen Experimenten hervorgegangen sind. Die Zunahme kombinatorischer Synthesebeispiele der letzten Jahre spiegelt sich nicht nur in diesem einzelnen Neuron (84,39) wider, sondern auch noch in weiteren Gebieten der Karte in Abbildung 4-26. In weiteren drei Gebieten sind ebenfalls Reaktionen eingetragen, bei denen eine C-N-Einfachbindung aufgebaut wird.

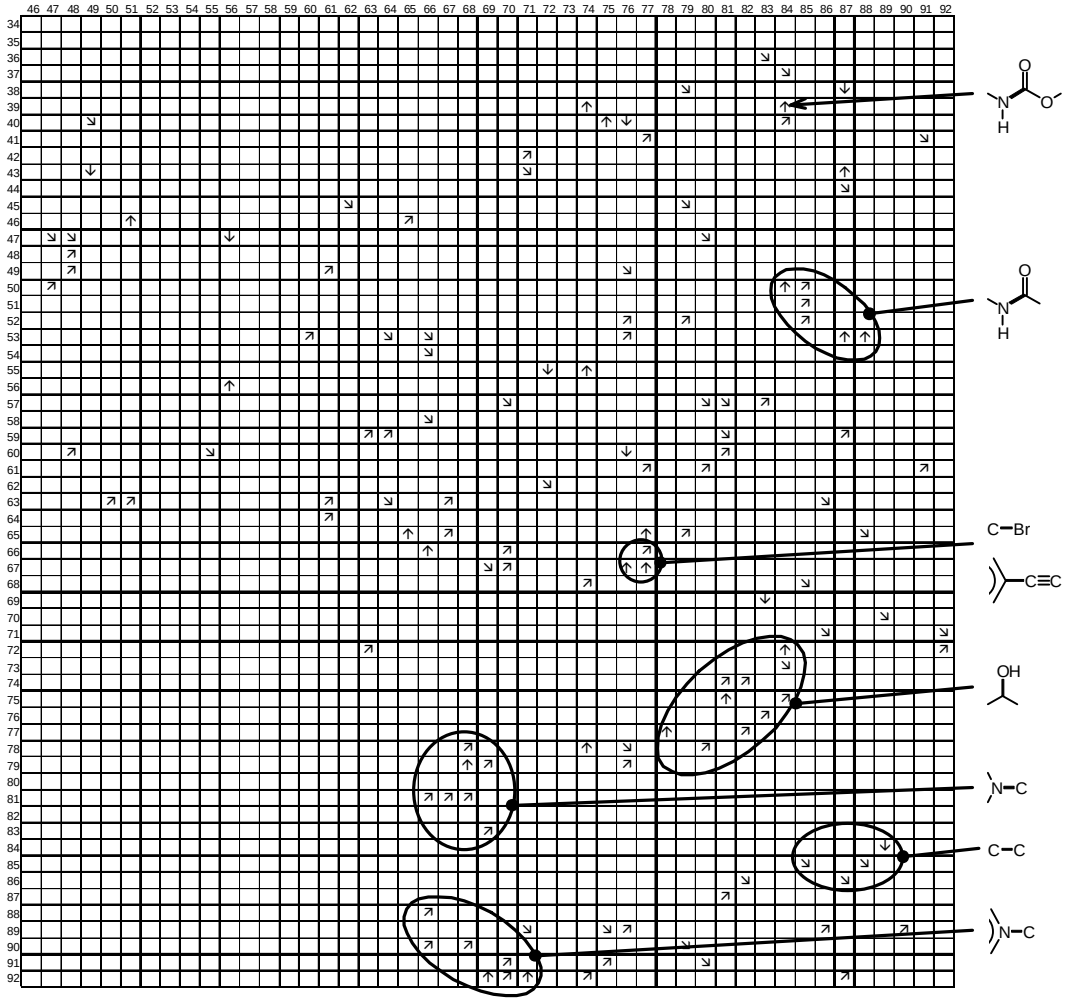


Abb. 4-26: Zeitliche Entwicklung der ChemInform RX-Reaktionsdatenbanken über die Jahre 1992–1997 für den Bereich mit nur einer gebildeten Einfachbindung im Reaktionszentrum auf der Produktseite. Direkt nach oben zeigende Pfeile ↑ symbolisieren eine Steigung der Regressionsgerade von über 1,0, schräg nach oben zeigenden Pfeile ↗ eine Geradensteigung von 0,5 bis 1,0; entsprechendes gilt für die nach unten zeigenden Pfeile ↘ und ↓ .

Da diese Gebiete weit über den Bereich der Einfachbindungen verteilt sind, ist die physikochemische Variationsbreite für diesen Reaktionstyp sehr groß: Das Stickstoffatom der aufzubauenden C-N-Einfachbindung kann im aromatischen System integriert sein (unterster Bereich), benachbart zur einer Carbonylgruppe stehen (oberer Bereich) oder am Ende einer aliphatischen Kohlenstoffkette positioniert sein (mittlerer Bereich). In weiteren zwei Bereichen der Abbildung 4-26 ist ebenfalls ein starker Anstieg der Reaktionsbeispiele zu verzeichnen. In einem Bereich ist ein Anstieg von Reaktionen zu beobachten, bei denen eine C-O-Einfachbindung gebildet wird. Zum einen handelt es sich bei diesen Reaktionen um α-Hydroxylierungen von Carbonylverbindungen, wie beispielsweise Reaktion #7924 aus CIRX97 (siehe Abbildung 4-27).

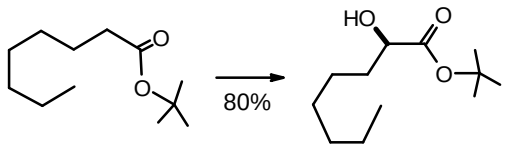


Abb. 4-27: Reaktion #7924 aus der ChemInform RX97-Datenbank (RXCI96032221) aus Neuron (78,77).

Zum anderen werden in diesem Bereich auch Reduktionsreaktionen von Ketonen zu sekundären Alkoholderivaten eingetragen, die oft stereoselektiv mit Hydrid-Übertragungsreagenzien durchgeführt werden (siehe Abbildung 4-28).

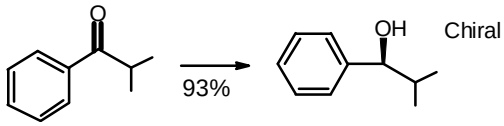


Abb. 4-28: Reaktion #5292 aus der ChemInform RX97-Datenbank (RXCI96028901) aus Neuron (83,76).

Die zahlreichen Untersuchungen zur Stereoselektivität dieses Reaktionstyps sind für die Zunahme der Reaktionen, bei denen eine C-O-Einfachbindung aufgebaut wird, maßgeblich verantwortlich.

Der fünfte Bereich mit steigender Tendenz an Reaktionsbeispielen schließlich erfaßt Reaktionen, in denen zum einen eine C-Br-Einfachbindung durch elektrophile aromatische Substitution eingeführt wird, wie es Reaktionsbeispiel #44956 aus Neuron (77,67) in Abbildung 4-29 zeigt.

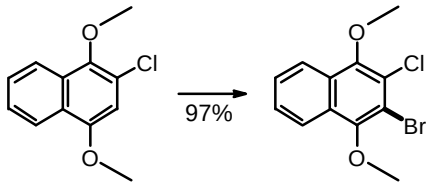


Abb. 4-29: Reaktion #44956 aus der ChemInform RX97-Datenbank (RXCI96072756) aus Neuron (77,67).

In diesem Neuron wird aber auch noch ein anderer Reaktionstyp eingetragen, dessen Reaktionszentrum auf der Produktseite physikochemisch sehr ähnlich zur elektrophilen aromatischen Substitution ist. Hierbei handelt es sich um eine Alkinylierungsreaktion, eine Substitutionsreaktion eines Alkinderivates an einem aromatischen System unter Metall-Katalyse (zum Beispiel CuI, $Pd(Ph)_4$ und NEt_3). Als Reaktionsbeispiel ist in Abbildung 4-30 Reaktion #8837 angegeben.

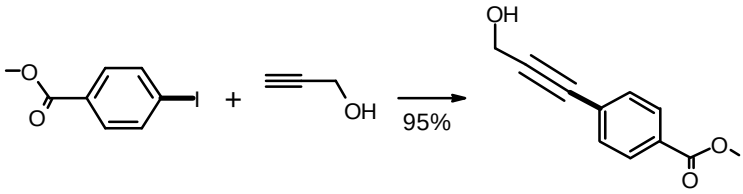


Abb. 4-30: Reaktion #8837 aus der ChemInform RX97-Datenbank (RXCI96033197) aus Neuron (76,67).

In einigen Reaktionsbeispielen aus diesem Bereich erfolgt die Alkinylierung auch an einer Doppelbindung, die mit einer Carbonylgruppe wiederum in Konjugation steht. Abbildung 4-31 zeigt ein solches Beispiel.

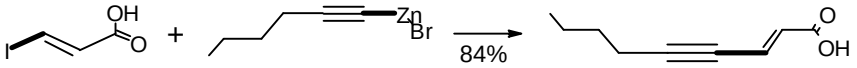


Abb. 4-31: Reaktion #8002 aus der ChemInform RX97-Datenbank (RXCI96032300) aus Neuron (77,66).

Im untersuchten Zeitraum nahm der prozentuale Anteil an Reaktionsbeispielen und damit die Bedeutung dieser Alkinylierungsreaktion ständig zu. Ein Grund liegt in der leichten Überführbarkeit der Dreifachbindung in den Produkten in verschiedene funktionelle Gruppen. Bei den Alkinylierungsreaktionen kann man auch von aromatischen Bromverbindungen ausgehen, was wiederum den prozentualen Anstieg bei den aromatischen C-Br-Einfachbindungen erklären läßt.

Schließlich wird noch ein Bereich herausgestellt, in dem in den vergangenen Jahren die Anzahl der Reaktionen stetig abgenommen hat. In diesem Bereich werden Reaktionen eingetragen, bei denen C-C-Einfachbindungen durch Alkylierungsreaktionen aufgebaut werden. Dabei werden meist Alkyllithium-, Cuprat- oder Grignardverbindungen eingesetzt. Abbildung 4-32 zeigt ein Reaktionsbeispiel, das in Neuron (89,84) eingetragen wird.

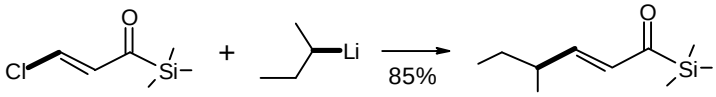


Abb. 4-32: Reaktion #46126 aus der ChemInform RX97-Datenbank (RXCI96000178) aus Neuron (89,84).

Die Bedeutung dieser Alkylierungsreaktionen ist also in den letzten Jahren stetig zurückgegangen.

Die Klassifizierung von Reaktionsdatenbanken aufeinanderfolgender Jahrgänge nach dem Referenzverfahren ermöglicht das Verfolgen der zeitlichen Entwicklung dieser Datenbanken. Bedingt durch die rapide Entwicklung auf dem Gebiet der kombinatorischen Synthese, die sich historisch bedingt anfänglich den Säureamid-Synthesen widmete, findet man in den Datenbanken einen Anstieg bei den Reaktionsbeispielen, bei denen eine C-N-Bindung gebildet wird. Die anderen beobachteten Tendenzen sind eher durch die Möglichkeiten einer Reaktion geprägt. Dazu rechnet man zum einen die stereoselektive Durchführbarkeit, wie am Beispiel der Hydroxylierungsreaktionen gezeigt. Andererseits erhöht auch eine leichte Funktionalisierung der Produkte die Bedeutung eines Reaktionstyp, wie am Beispiel der Alkinylierungsreaktionen gezeigt wurde.

4.5 Diskussion des Einsatzes der Reaktionsklassifizierung beim Datenbankenvergleich

Die Theilheimer Reaktionsdatenbank wurde in diesem Abschnitt mit der im Rahmen dieser Arbeit entwickelten Methode klassifiziert. Bedingt durch den Einsatz von Codierungsvektoren, die sowohl die Anzahl der Bindungen im Reaktionszentrum auf der Produktseite, als auch deren physikochemischen Effekte repräsentieren, werden die Reaktionen primär nach der gleichen Anzahl an Bindungen und ähnlichen physikochemischen Effekten eingeteilt. Es wurde aber auch gezeigt, daß in derselben Kohonen-Karte auch eine Einteilung in verschiedene Reaktionstypen vorliegt. Für abgetrennte Bereiche kann man meist einen Bindungstyp angeben, der sich durch die strukturelle Umgebung von anderen Bindungstypen unterscheidet.

Anhand der Wittig-Reaktionen wurde gezeigt, wie die Namensreaktionen miteinander in Beziehung stehen. Verschiedene Namensreaktionen gehen von verschiedenen Ausgangsverbindungen aus, die unter bestimmten, für diese Namensreaktion typischen, Reaktionsbedingungen umgesetzt werden. Letztlich werden aber oft dieselben Atome miteinander verknüpft, so daß für den Aufbau einer Bindung in der Regel verschiedene Namensreaktionen durchgeführt werden können. Die Namensreaktionen, die zu denselben Bindungsverknüpfungen führen und darüber hinaus auch noch in physikochemischen Eigenschaften dieser aufgebauten Bindungen weitgehend übereinstimmen, werden mit der klassifizierten Kohonen-Karte aufgezeigt. Somit stellt dieses Verfahren ein neuartiges Indexierungsverfahren dar, das nicht nur die Konnektivität der aufgebauten Atome berücksichtigt, sondern auch die physikochemischen Effekte der geknüpften Bindungen einbezieht.

Will man die in verschiedenen Datenbanken enthaltenen Reaktionen miteinander vergleichen, so stellt die Beschränkung auf das Wesentliche einer Reaktion, nämlich das Reaktionszentrum einer Reaktion, einen guten Ansatz dar. Ein Vergleich der Reaktionen ist auch dann möglich, falls diese aus verschiedenen Reaktionsdatenbanken, wie Festphasenreaktionsdatenbanken oder Datenbanken über Reaktionen an flüssiger Phase, stammen. Für den Vergleich ist es dann unerheblich, ob die Moleküle bei der Reaktion an einer festen Phase gebunden sind oder nicht.

Ein Vergleich der Theilheimer- mit der SPORE Reaktionsdatenbank basierend auf der Reaktionsklassifizierung führt zu dem Ergebnis, daß inzwischen an fester Phase viele verschiedene Reaktionstypen realisierbar sind, auch wenn ursprünglich diese Synthesen für die Knüpfung von Kohlenstoff-Stickstoff-Bindungen ausgerichtet waren. Wesentliche Unterschiede ergaben sich nur bei den Mehrfachbindungen.

Im letzten Kapitel zum Vergleich von Reaktionsdatenbanken wurde die zeitliche Entwicklung der ChemInform RX-Reaktionsdatenbank beschrieben. Es zeigte sich, daß bedingt durch die kombinatorische Synthese einige Reaktionstypen, bei denen zum Beispiel eine Kohlen-

stoff-Stickstoff-Bindung geknüpft wird, in den letzten Jahren stark an Bedeutung gewonnen haben.

5 Praktische Anwendung: Syntheseplanung

5.1 Computergestützte Syntheseplanung

Corey betrat vor mehr als 30 Jahren bei der Planung organischer Synthesen Neuland: Er entwickelte das Konzept der Retrosyntheseplanung[66]. Von einem Zielmolekül wird mittels formulierter Regeln und Schritte, die nach- oder nebeneinander vollzogen werden, ein Syntheseplan ausgearbeitet. Auf Corey gehen auch die heute noch verwendeten Begriffe in der Syntheseplanung zurück. Jede formulierte chemische Umsetzung kann aus zwei Blickwinkeln betrachtet werden: Die in Laboratorien durchführbare *Reaktion* von Edukten zu Produkten bezeichnet man als die *synthetische* Richtung, während die umgekehrte Richtung als *retrosynthetische* oder *antithetische* bezeichnet wird. Die Formulierung retrosynthetischer Schritte endet bei der Angabe von *Synthons*, das sind meist geladene Fragmente, die eine gedankliche Vorstufe zu den eingesetzten Reagenzien darstellen. Das Ergebnis einer Retrosyntheseplanung wird in Form eines *Synthesebaumes* dargestellt, wobei das Syntheseziel die Wurzel eines auf dem Kopf stehenden Baumes darstellt. Die Zwischenprodukte der Reaktionen werden graphentheoretisch betrachtet von den Knoten des Baumes repräsentiert, die Ausgangsmaterialien sind die Endpunkte von Ästen und die Reaktionen die Kanten. Die Planung einer Synthese unter Anwendung von Regeln und Heuristiken läßt sich leicht in einen Algorithmus übertragen, der Bestandteil eines Computerprogramms sein kann. Dem ersten computergestützten Syntheseplanungsprogramm von Corey und Wipke im Jahre 1969 mit dem Namen OCSS[67] folgten beispielsweise LHASA[68], SECS[69], CICLOPS[70], EROS[71], SYNCHEM[72], SYNGEN[73] und AIPHOS[74].

5.2 Das WODCA Programmsystem

Im Jahre 1988 begann man im Arbeitskreis von Gasteiger ein separates Programmsystem für die Syntheseplanung zu entwickeln. Aus dem Programmsystem EROS5, das damals sowohl für die computergestützte Syntheseplanung als auch für die Reaktionsvorhersage konzipiert war, ging ein eigenständiges Programm nur für die Syntheseplanung hervor, das WODCA genannt wurde[75][76]. WODCA ist ein Akronym für den Ausdruck „***W**orkbench for the **O**rganization of **D**ata for **C**hemical **A**pplications*“. In der Zwischenzeit wurde WODCA zu einem Programmsystem für die interaktive Syntheseplanung erweitert, das unter zahlreichen UNIX-Betriebssystemen (Sun Solaris, SGI IRIX, SuSE Linux) lauffähig ist. Der Programmbenutzer wird bei der Ausarbeitung eines Syntheseplans durch eine Vielzahl an Methoden unterstützt, wobei hauptsächlich zwei Methoden eingesetzt werden, die oft einander abwechselnd angewendet werden. Zum einen kann mit Hilfe von Ähnlichkeitssuchen für eine Zielverbindung festgestellt werden, ob kommerziell erhältliche Ausgangsmaterialien

verfügbar sind, die mit häufig leicht durchführbaren Reaktionen – wie Hydrolysen, Oxidationen etc. – in die Zielverbindung überführt werden können. Falls die Ähnlichkeitssuche keine geeigneten Ausgangsmaterialien hervorbringt, kann die zweite Methode angewendet werden. Dabei können für die Zielverbindung strategische Bindungen bestimmt und bewertet werden. Als strategisch werden in WODCA nur solche Bindungen gekennzeichnet, die in Syntheserichtung aufgrund ihrer physikochemischen Eigenschaften leicht aufzubauen sind. Zerlegt man die Zielverbindung durch Bruch einer ausgewählten strategischen Bindung, so werden anschließend die freien Valenzen durch geeignete Substituenten abgesättigt. Mit den neu erzeugten Synthesevorstufen kann man wiederum eine Ähnlichkeitssuche durchführen. Hat man für jede Synthesevorstufe eine kommerziell erhältliche Verbindung gefunden, so ist der Synthesebaum vollständig aufgebaut.

Die gegenwärtige Programmversion 4.00 basiert auf einer logisch-zentrierten Funktionsweise, in der wissensbasierte Heuristiken eingebunden sind. Diese Heuristiken werden in einer um chemiespezifische Inhalte erweiterten Tcl/Tk-Programmiersprache geschrieben und getrennt vom Programmkern in einer Wissensbasis gehalten. WODCA unterscheidet grundsätzlich den Aufbau einer Kohlenstoff-Kohlenstoff- und einer Kohlenstoff-Heteroatom-Bindung. Beim retrosynthetischen Bruch einer Kohlenstoff-Kohlenstoff-Bindung werden Regeln und Heuristiken angewandt, die hauptsächlich auf physikochemischen Effekten basieren. Für spezielle Reaktionstypen existiert darüber hinaus auch eine Wissensbasis, die jederzeit um weitere Regeln erweitert werden kann. Ein großer Nachteil ist jedoch, daß die Heuristiken einzeln konzipiert und programmiert werden müssen. Vor kurzem wurden beispielsweise Regeln und Heuristiken in WODCA aufgenommen, die die retrosynthetische Planung substituierter Aromaten ermöglichen[77]. Andererseits fehlen beispielsweise in WODCA noch Heuristiken, die pericyclische Reaktionen erfassen. Für Cyclohexen-Derivate können derzeit keine retrosynthetischen Wege vorgeschlagen werden, die in Syntheserichtung über Diels-Alder-Reaktionen laufen würden.

Beim Bruch einer Kohlenstoff-Heteroatom-Bindung kommt dagegen hauptsächlich eine Transform-Bibliothek zum Einsatz. In dieser werden für jeden Reaktionstyp strukturelle Charakteristika abgelegt, die angeben, welche funktionellen Gruppen für den Reaktionstyp vorausgesetzt werden und welche Bindungen umgesetzt werden. Dieses Konzept einer manuell erstellten Wissensbasis ist sehr zeitaufwendig, da Strategien oder Transforms für eine Reihe von Reaktionstypen aufgestellt und gepflegt werden müssen. Andererseits ist eine große Transform-Bibliothek keine Garantie für ein erfolgreiches Syntheseplanungsprogramm, wie es die Einstellung des CASP-Projekts mit der damals größten Wissensbasis von über 6.000 Transforms zeigte[78]. Eine manuell erstellte Wissensbasis ist darüber hinaus meist auf ein Reaktionsmedium ausgerichtet. Gegenwärtig ist die Wissensbasis in WODCA für Laborsynthesen optimiert. Die direkte Übertragung einer retrosynthetischen Syntheseplanung auf

Reaktionen an fester Phase oder auf großtechnisch durchgeführte Reaktionen ist meist nicht von Erfolg gekrönt.

Aus den genannten Gründen sollte man den Schwerpunkt der computergestützten Syntheseplanung auf die Entwicklung von Alternativen zu einer manuellen Wissensbasis legen. Auch in diesem Fall bietet sich die in Reaktionsdatenbanken gespeicherte Information an, die sich aus Millionen von Einzelreaktionen inklusive den dazugehörigen chemischen Fakten zusammensetzt.

Diese Methode wird beispielsweise in einem wissensbasierten System zur Syntheseplanung umgesetzt. Dieses von Nakayama entwickelte Konzept namens KASP (***K**nowledge **A**cquisition system for **S**ynthesis **P**lanning*)[79] basiert auf Transforms, die aus einer Reaktionsdatenbank mit ca. 30.000 Einträgen gewonnen wurden.

In dieser Arbeit wird eine Methode vorgestellt, die eine Wissensbasis durch das Klassifizieren organischer Reaktionen generiert. Mit Hilfe dieser Wissensbasis kann man bei der Syntheseplanung nach strategischen Bindungen in Targetmolekülen suchen lassen und diese Bindungen hinsichtlich ihrer Eignung bewerten lassen. Darüber hinaus kann man auch auf das chemische Faktenwissen, in Form von Ausbeuten, Reaktionsbedingungen etc., zurückgreifen.

5.3 Syntheseplanung mittels Reaktionsklassifizierung

In Kapitel 2.2.1 wurde dargelegt, daß die Theilheimer Reaktionsdatenbank synthetisch wichtige Reaktionen in hoher Ausbeute enthält. Die in dieser Datenbank enthaltene Information eignet sich also sehr gut als Wissensbasis für die Syntheseplanung. Wie in Kapitel 3.2.1 erläutert wurde, kann bei der Syntheseplanung nur der Teil des Reaktionszentrums zur Reaktionscodierung herangezogen werden, der ausschließlich die neu gebildeten Bindungen auf der Produktseite beschreibt. Auf diese Weise wurden bereits die Reaktionen der Theilheimer Reaktionsdatenbank in Kapitel 4.2 klassifiziert, wobei das in Kapitel 3.3 erwähnte Standardverfahren zur Codierung eingesetzt wurde. Die resultierende Kohonen-Karte ist in Abbildung 4-5 dargestellt.

Die klassifizierte Kohonen-Karte kann in der computergestützten Syntheseplanung eingesetzt werden, um in einem Zielmolekül nach Bindungen zu suchen, die das Molekül in kleinere, eventuell käufliche Ausgangsverbindungen zerlegen. Man bezeichnet solche Bindungen als strategisch, wenn beim Aufbau dieser Bindungen möglichst viele der folgenden Aspekte erfüllt sind:

- Quantitative Umsetzung der Ausgangsverbindungen zum Zielmolekül
- Leichte experimentelle Durchführbarkeit der Reaktion
- Zerlegung des Zielmoleküls in gleich große Vorstufen etc.

Aus diesem Grund reicht es meistens nicht aus, strategische Bindungen nur zu bestimmen, sondern diese müssen auch anhand der oben genannten Bedingungen bewertet werden.

Zur Suche nach strategischen Bindungen in einem Targetmolekül wird zunächst für alle Nicht-Wasserstoff-Bindungen die jeweils ähnlichste Reaktion bestimmt. Bindungen zu Wasserstoffatomen können nicht gesucht und beurteilt werden, da sie in den Reaktionsdatenbanken nicht abgespeichert wurden. Von jeder zu untersuchenden Bindung werden dieselben sechs physikochemischen Deskriptoren berechnet, die auch bei der Codierung der Reaktionsdatenbank verwendet wurden. Das Zielmolekül wird also als Produkt der zu untersuchenden Reaktion angesehen, dessen Ausgangsverbindungen zunächst noch unbekannt sind. Die ähnlichste Reaktion zu dieser Reaktion, für die ein Retrosyntheseschritt gesucht wird, ist diejenige Reaktion aus dem Trainingsdatensatz, die physikochemisch betrachtet das ähnlichste Reaktionszentrum auf der Produktseite aufweist. Dazu wird diese retrosynthetische Reaktion in das trainierte Kohonen-Netz projiziert und das Gewinnerneuron ermittelt. Die in diesem Neuron eingetragenen Reaktionen sind dann zu der retrosynthetischen Reaktion am ähnlichsten. Da im Trainingsdatensatz maximal 6 Bindungen im Reaktionszentrum auf der Produktseite zulässig sind, können alle Einzelbindungen und sämtliche Bindungskombinationen bis maximal zum 6er Tupel systematisch durchlaufen werden, um strategische Bindungen in dem Targetmolekül aufzufinden.

Die zu untersuchende Bindung wird nach dem modifizierten USMILES-Code (siehe Kapitel 3.2.3.2) angeordnet, das auch zur Ausrichtung der geknüpften Bindungen des Trainingsdatensatzes eingesetzt wurde. Im Falle einer symmetrischen Bindung erfolgt die Ausrichtung hier allerdings nicht über die σ-Elektronegativität, obwohl alle Reaktionen des Trainingsdatensatzes mit einem symmetrischen Teil des Reaktionszentrums so ausgerichtet wurden. Vielmehr müssen bei der Anfrage einer symmetrischen Bindung beide „Ausrichtungsmöglichkeiten“ beachtet werden. Im Falle einer C-C-Bindung existieren beispielsweise zwei Möglichkeiten zum Aufbau der Bindung (siehe Abbildung 5-1). Deshalb werden die zu beiden Bindungsausrichtungen jeweils ähnlichsten Reaktionen gesucht und bewertet. Aus diesem Grund können bei symmetrischen Reaktionszentren zwei Gewinnerneuronen angegeben sein.

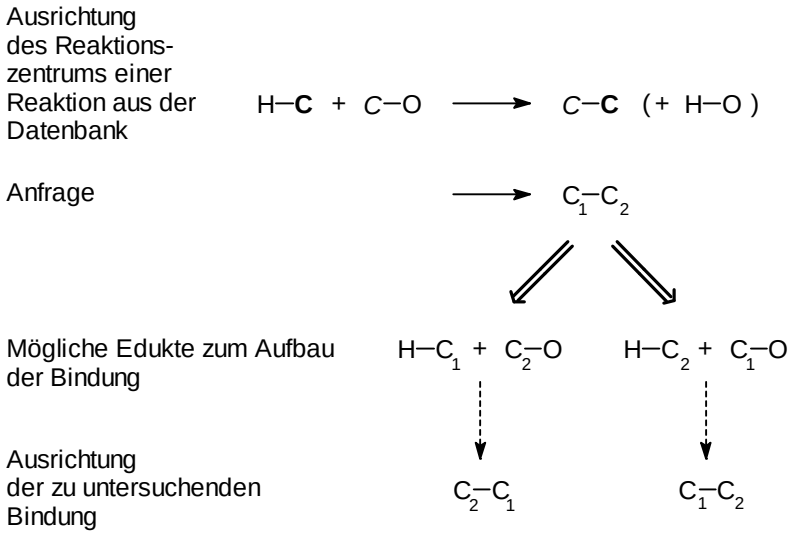


Abb. 5-1: Für eine zu untersuchende, symmetrische Bindung werden zwei Ausrichtungen berücksichtigt.

Generell kann jede Bindung einer retrosynthetischen Reaktion, zu der mindestens eine ähnliche Reaktion gefunden wird, als strategisch betrachtet werden. Wird zu einer Bindung keine Gewinnerreaktion ermittelt, so zeigt die Bindung in der vorgegebenen Wissensbasis keine strategische Bedeutung.

Eine Bewertung der gefundenen strategischen Bindungen erfolgt aus der Anzahl der ähnlichsten Reaktionen sowie den zugehörigen chemischen Angaben. Häufig sind Reaktionstypen, die aufgrund ihrer Reaktionsbedingungen leicht durchführbar sind und nahezu quantitativ die entsprechenden Produkte ohne nennenswerte Nebenprodukte liefern, in der Theilheimer Reaktionsdatenbank mit mehreren Reaktionsbeispielen vertreten. Diese Reaktionsbeispiele werden aufgrund ihrer physikochemischen Ähnlichkeit meist im selben Neuron eingetragen, so daß die strategische Bedeutung einer Bindung um so größer ist je mehr Gewinnerreaktionen zu der retrosynthetischen Reaktion gefunden werden. Andererseits erscheint eine Bindung nicht besonders strategisch, wenn die ähnlichsten Reaktionen nur geringe Ausbeuten aufweisen, oder unter sehr speziellen Reaktionsbedingungen durchgeführt werden müssen. Das Reaktionszentrum der ähnlichsten Reaktion wird anschließend auf die Anfragereaktion übertragen, so daß die retrosynthetische Reaktion nach genau demselben Reaktionstyp ablaufen würde. Dazu muß auch die Umgebung des Reaktionszentrums auf der Eduktseite der ähnlichsten Reaktion angepaßt werden, um reale Ausgangsverbindungen für die Anfragereaktion zu erhalten.

Diese Methode wird anhand der Synthese von zwei Zielmolekülen, nämlich eines Pyrazolderivats und 2-Oxo-5-phenyl-cyclopentan-carbonsäuremethylester, vorgestellt.

5.4 Syntheseplanungsbeispiel: Pyrazole

5.4.1 Bekannte Herstellungsverfahren

Für den Aufbau von 1*H*-Pyrazolderivaten findet man in der Literatur häufig zwei Synthesewege[80]. Der eine Syntheseweg verläuft über mehrere heterolytische Bindungsbrüche und -knüpfungen, der andere über eine pericyclische Reaktion. Der heterolytische Bindungsbruch von 1*H*-Pyrazol führt zunächst zu einem Hydrazono-enol-Derivat, das man in der ketotautomeren Form einsetzen würde. Dieses wiederum synthetisiert man aus der entsprechenden 1,3-Diketoverbindung und dem Hydrazinderivat (siehe Abbildung 5-2).

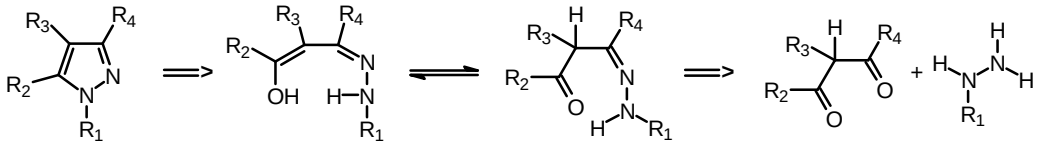


Abb. 5-2: Wichtigste Retrosynthesemöglichkeit für Pyrazolderivate.

Diese als Knorr-Pyrazol-Synthese in der Literatur bekannte Aufbaureaktion für 1*H*-Pyrazole geht also von 1,3-Diketoverbindungen und Hydrazinderivaten aus und verläuft über zwei Kondensationsreaktionen.

Der andere Darstellungsweg für Pyrazole verläuft über eine 1,3-dipolare Cycloaddition von Diazoalkanen an aktivierten Acetylenderivate. Allerdings sind mit dieser Methode N-substituierte Pyrazole nicht direkt herstellbar, da während der Synthese eine Verschiebung des Substituenten vom 5-C-Atom zu dem 1-N-Atom stattfinden muß, was nur für ein Wasserstoffatom leicht möglich ist (siehe Abbildung 5-3).

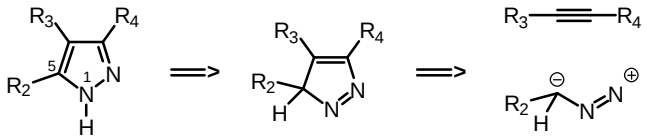
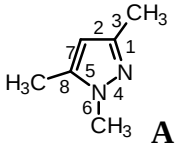


Abb. 5-3: Zweitwichtigste Retrosynthesemöglichkeit für Pyrazolderivate.

5.4.2 Syntheseplanung basierend auf der Reaktionsklassifizierung

Als Targetmolekül **A** wurde das mit insgesamt drei Methylgruppen substituierte Pyrazolderivat ausgewählt, dessen Ringsystem nur von schwachen induktiven Effekten beeinflusst wird (siehe Abbildung 5-4). In seiner elektronischen Natur nimmt es stellvertretend für

verschiedene Pyrazolderivate eine Mittelposition ein, die durch Substituenten beeinflußt werden, welche ein breites Spektrum an induktiven bzw. mesomeren Effekten aufweisen können.



A

Abb. 5-4: Pyrazolderivat **A**, das stellvertretend für alle N-substituierten Pyrazolderivate für eine Syntheseplanungsstudie eingesetzt wird. Die angegebenen Ziffern sind die Bindungsnummern.

Für dieses Pyrazolderivat **A** wird eine Suche nach strategischen Bindungen basierend auf der klassifizierten Theilheimer Reaktionsdatenbank durchgeführt. Für 22 Bindungskombinationen werden insgesamt 82 ähnliche Reaktionen gefunden, die aus 24 Neuronen stammen (siehe Tabelle 5-1). Da die Reaktionen in einem Neuron zueinander sehr ähnlich sind, genügt es in der Regel, einige wenige Beispiele aus jedem Neuron zu analysieren, um die verschiedenen vorgeschlagenen Reaktionswege beurteilen zu können. Die Bindungskombinationen sind sortiert nach der Anzahl der gefundenen ähnlichen Reaktionen in Tabelle 5-1 zusammengestellt. Die Bindungskombinationen, die im folgenden näher erläutert werden, sind in der Tabelle grau hervorgehoben.

Die meisten Gewinnerreaktionen findet man zu der Bindung mit der Bindungsnummer 8. Der retrosynthetische Bindungsbruch der Bindung 8 würde aber nur die Methylgruppe abspalten, das Pyrazolsystem selbst kann durch Methylierung des Pyrazolsystems beispielsweise mit Butyllithium und Dimethylsulfat nicht aufgebaut werden. Gleiches gilt für Bindung 3, für die 12 ähnliche Reaktionen gefunden werden. An dritter Stelle folgt mit zehn ähnlichen Reaktionsbeispielen eine 3er Bindungskombination, bei der die Bindungen mit den Nummern 1, 5 und 7 geknüpft werden. Eine Änderung der Bindungsordnung an diesen drei Bindungen hat die Spaltung des Fünfrings zur Folge, so daß der diesen Reaktionen zugrunde liegende Reaktionstyp zum Aufbau des Pyrazolsystems eingesetzt werden kann. Eine genaue Analyse dieser zehn Reaktionen in Neuron (2,64) zeigt, daß diese meist nach der Knorr-Pyrazol-Synthesemethode aufgebaut werden. Die ähnlichste Reaktion zu der Anfragereaktion ist Reaktion #38785, die in Abbildung 5-5 dargestellt ist.

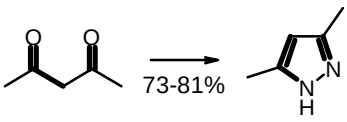
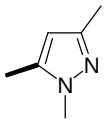
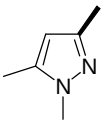
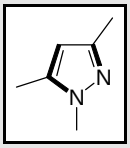
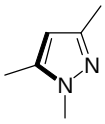
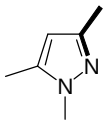
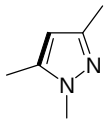
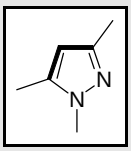
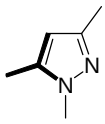
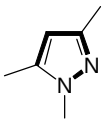
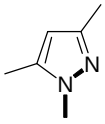
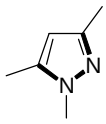
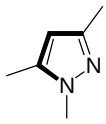
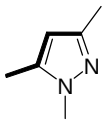
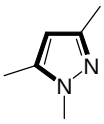
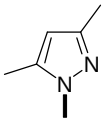
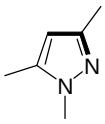
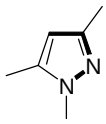
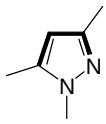
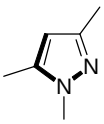
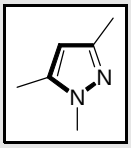
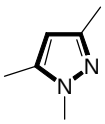
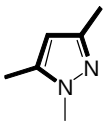


Abb. 5-5: Reaktion #38785 aus der Theilheimer Reaktionsdatenbank (RTHE00007837) aus Neuron (2,64).

Weitere vier der insgesamt zehn Gewinnerreaktionen (#26525, #28195, #29859 und #44938) bauen ebenfalls aus einer 1,3-Dicarbonylverbindung und einer Hydrazinverbindung Pyrazolderivate auf. Drei Reaktionen (#26063, #28329 und #30695) gehen von Mono- oder Diacetalverbindungen anstelle einer 1,3-Dicarbonylverbindung aus (siehe Abbildung 5-6).

(86,90) u. (46,45)	(46,45)	(2,64)	(45,26)	(51,2)	(58,84) u. (62,79)
					
#6330 (63%), #8046 (81%), #13861 (76%), #13862 (k.A.), #17388 (83%), #25448 (60-80%), #38053 (34%), #38070 (75-80%), #38865 (100%), #41471 (90%), #43425 (73%), #44131 (63%), #44543 (55-61%), #45402 (70-72%), #45836 (90%), #45989 (85%), #46169 (66%), #46271 (91%)	#17388 (83%), #25448 (60-80%), #38070 (75-80%), #38865 (100%), #41471 (90%), #43425 (73%), #44543 (55-61%), #45402 (70-72%), #45836 (90%), #45989 (85%), #46169 (66%), #46271 (91%)	#11761 (75-85%), #26063 (76%), #26525 (82%), #28195 (80%), #28329 (100%), #29859 (k.A.), #30695 (69%), #34638 (72%), #38785 (73-81%), #44938 (k.A.)	#440 (83%), #2875 (95%), #6120 (99%), #10177 (78%), #19241 (77%), #37936 (k.A.), #39610 (82%)	#2589 (68%), #5857 (98%), #10161 (91%), #18613 (84%), #25637 (92%)	#4383 (83%), #31402 (59%), #45665 (79%)
(48,28)	(26,56)	(1,31)	(57,22)	(30,57)	(8,60)
					
#6773 (82%), #14925 (75%), #22229 (96%)	#4018 (84%), #15346 (100%), #21136 (k.A.)	#4192 (73%), #22201 (71%), #39593 (k.A.)	#29305 (88%), #29306 (k.A.)	#1372 (55%), #18835 (95%)	#30977 (70%), #42954 (k.A.)
(28,58)	(18,41)	(68,91)	(53,4)	(31,57)	(3,66)
					
#6674 (55-80%), #12099 (90%)	#15739 (66%), #29679 (82%)	#21455 (67%)	#42288 (93%)	#18338 (59%)	#13536 (60%)
(8,66)	(3,35)	(1,26)	(6,2)		
					
#195 (95%)	#26610 (95%)	#18511 (k.A.)	#22573 (90%)		

Tab. 5-1: Ergebnis der Suche nach strategischen Bindungen für das Pyrazolderivat **A**: Für die angegebenen 22 Bindungskombinationen sind die ähnlichsten Reaktionen aus der Theilheimer Reaktionsdatenbank angegeben, die dem Gewinnerneuron entnommen sind. Über der Strukturformel, die markiert die strategischen Bindungen zeigt, ist das Gewinnerneuron angegeben, darunter sind die Gewinnerreaktionen aus der Theilheimer Reaktionsdatenbank mit der Ausbeute in Klammern angeführt.

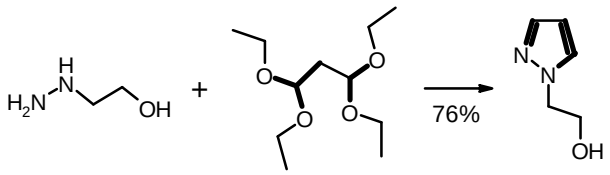


Abb. 5-6: Reaktion #26063 aus der Theilheimer Reaktionsdatenbank (RTHE00022639) aus Neuron (2,64).

Im Falle der Reaktion #11761 wird Penta-3-in-2-on eingesetzt, das mit Phenylhydrazin umgesetzt wird und ebenfalls ein Pyrazolderivat liefert (siehe Abbildung 5-7).

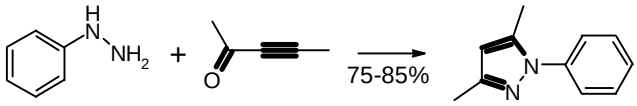


Abb. 5-7: Reaktion #11761 aus der Theilheimer Reaktionsdatenbank (RTHE00034833) aus Neuron (2,64).

Schließlich zeigt Reaktion #34638 noch eine Variation dieser Synthesemethode: Hier wird α-Epichlorhydrin mit Phenylhydrazin umgesetzt (siehe Abbildung 5-8).

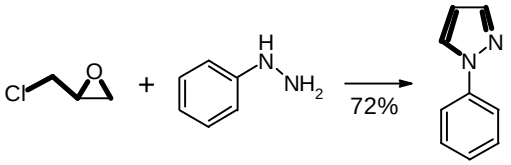


Abb. 5-8: Reaktion #34638 aus der Theilheimer Reaktionsdatenbank (RTHE00012065) aus Neuron (2,64).

Für die zweite in der Literatur angegebene Synthesemöglichkeit der Pyrazole findet man an 20. Stelle in der Bewertungsabfolge nur eine einzige Reaktion (siehe Abbildung 5-9). Die Reaktion #26610 wird bei dem untersuchten 4er-Bindungsbruch 2-4-5-7 in Neuron (3,35) als ähnlichste Reaktion ausgegeben. In diesem Falle wird aus Ethin und Diazomethan 1*H*-Pyrazol synthetisiert. Angesichts der wenigen Reaktionsbeispiele eignet sich dieser Syntheseweg allerdings weniger zum Aufbau unterschiedlich substituierter Pyrazolderivate.

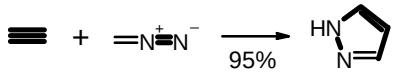


Abb. 5-9: Reaktion #26610 aus der Theilheimer Reaktionsdatenbank (RTHE00018366) aus Neuron (3,35).

Die anderen vorgeschlagenen Bindungskombinationen enthalten häufig alternative Synthesewege zu den oft in der Literatur genannten Herstellungsverfahren. Reaktion #22229 in Neuron (48,28) baut den Pyrazolring durch Ringverengung auf (siehe Abbildung 5-10). Allerdings erfordert in diesem Fall die Synthese der Ausgangsstoffe für eine solche Umlagerungsreaktion eine eigene Syntheseplanungsstudie.

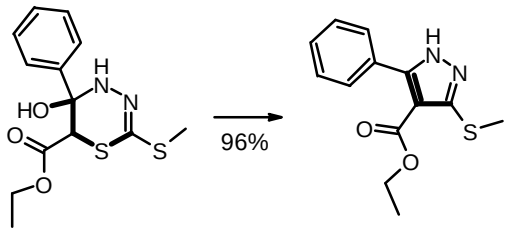


Abb. 5-10: Reaktion #22229 aus der Theilheimer Reaktionsdatenbank (RTHE00026483) aus Neuron (48,28).

Die anderen Reaktionen in den Neuronen behandeln entweder die Synthese von offenkettigen stickstoffhaltigen Verbindungen, wie Alkylidenhydrazinen, oder die Synthese von anderen stickstoffhaltigen Heterocyclen wie Indole und Indazole. Die sieben ähnlichen Reaktionen in Neuron (45,26) behandeln beispielsweise die Synthese von Indol-Derivaten, so daß sie für die Synthese von Pyrazolderivaten weniger geeignet erscheinen.

Das trainierte Netz schlägt also zum Aufbau der Pyrazolderivate genau den in der Literatur genannten Syntheseweg vor. Gleichzeitig werden auch Variationsmöglichkeiten bei den Edukten aufgezeigt, wie der Einsatz von Ketalen, Alkyl-in-on-Derivaten oder α-Chlorepoxiden anstelle der 1,3-Dicarbonylverbindung. Für das Targetmolekül schlägt das trainierte Netz somit die in Abbildung 5-11 dargestellten Ausgangsverbindungen und Reaktionswege vor.

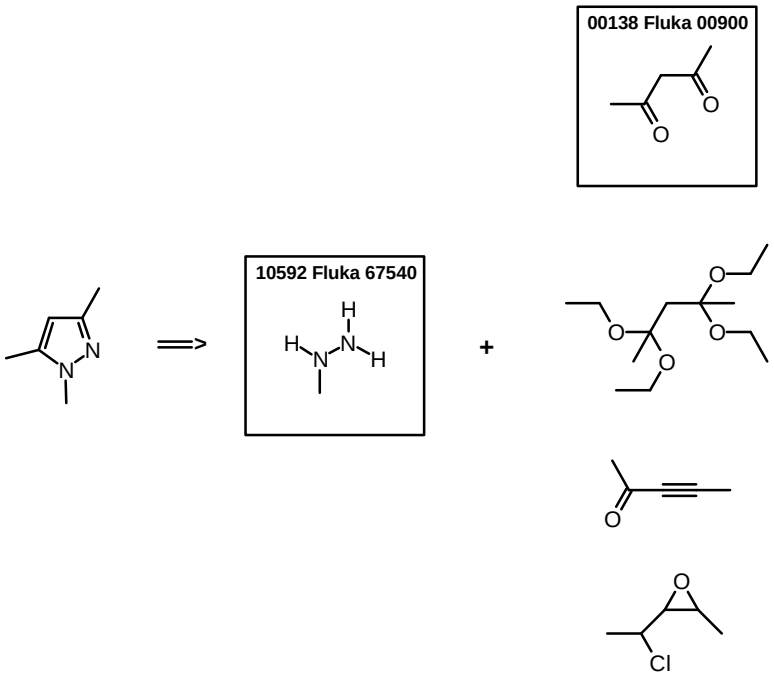
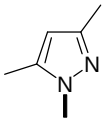
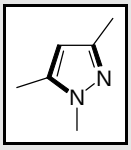
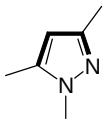
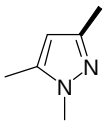
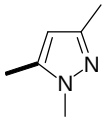
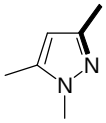
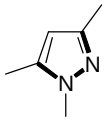


Abb. 5-11: Die vom trainierten Netz vorgeschlagenen Ausgangsverbindungen für die Synthese des Pyrazolderivats **A**.

Diese Suche und Bewertung strategischer Bindungen basierend auf der klassifizierten Theilheimer Reaktionsdatenbank wird nun derjenigen in der klassifizierten SPORE Reaktionsdatenbank gegenübergestellt. Hierzu wird auf das in Kapitel 4.3 beschriebene neuronale Netz mit Reaktionen aus der SPORE Reaktionsdatenbank zurückgegriffen.

Im Unterschied zur klassifizierten Theilheimer Reaktionsdatenbank werden die ähnlichsten Reaktionen nicht nur dem Gewinnerneuron entnommen, sondern auch der ersten Nachbarschaftssphäre, da viele Neuronen in dem klassifizierten Datensatz aus der SPORE Datenbank keine einzige Reaktion enthalten. Im Unterschied zur klassifizierten Theilheimer Reaktionsdatenbank werden hier nur 7 Bindungskombinationen mit insgesamt 15 ähnlichen Reaktionen gefunden, die aus 8 Neuronen stammen. Die Bindungskombinationen sind wiederum sortiert nach der Anzahl der gefundenen ähnlichen Reaktionen in Tabelle 5-2 zusammengestellt.

$(68,90)^1$  #2906 (k.A.), #2907 (k.A.), #2908 (k.A.), #2909 (k.A.), #2910 (k.A.), #5306 (83%)	$(1,63)^1$  #1459 (k.A.), #1460 (k.A.), #2260 (100%)	$(2,65)^1$ u. $(3,66)^G$  #2694 (100%), #2696 (100%)	$(46,45)^G$  #4283 (k.A.)
$(46,45)^G$  #4283 (k.A.)	$(51,2)^G$  #4816 (71%)	$(30,57)^G$  #3349 (k.A.)	

Tab. 5-2: Ergebnis der Suche nach strategischen Bindungen für das Pyrazolderivat **A**: Für die angegebenen 7 Bindungskombinationen sind die ähnlichsten Reaktionen aus der SPORE Reaktionsdatenbank angegeben, die dem Gewinnerneuron$^{(G)}$ und der ersten Nachbarschaftssphäre$^{(1)}$ entnommen sind.

Das Ergebnis dieser Suche nach strategischen Bindungen liefert trotz der geringeren Anzahl an strategischen Bindungskombinationen ähnliche Ergebnisse wie im Falle der Theilheimer Reaktionsdatenbank. Die mit Festphasenreaktionen aufgebaute Wissensbasis bewertet die exocyclische Kohlenstoff-Stickstoff-Bindung am höchsten, während in der Theilheimer Reaktionsdatenbank für diese Bindung nur eine einzige ähnliche Reaktion gefunden wird (siehe Tabelle 5-1). Danach folgt die 3er Bindungskombination mit den Bindungsnummern 1, 5 und 7, die den Pyrazolring in zwei Fragmente teilt. In Syntheserichtung entspricht dieser Bindungsbruch wieder der Knorr-Pyrazol-Synthese, diesmal allerdings an fester Phase ausgeführt. Die ähnlichste Reaktion aus der SPORE Datenbank, die in Neuron (1,63) eingetragen ist, zeigt Abbildung 5-12.

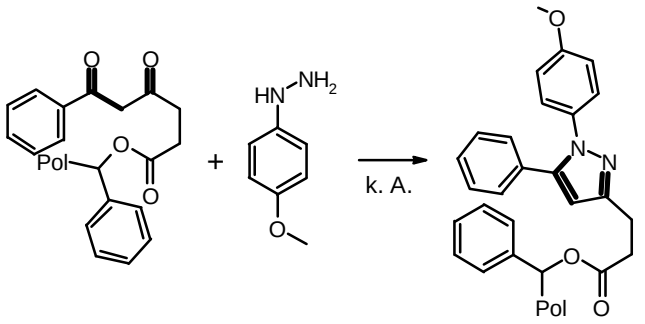


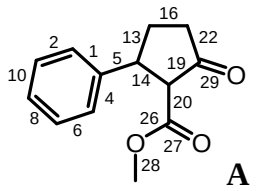
Abb. 5-12: Reaktion #1459 aus der SPORE Reaktionsdatenbank (RSPO69001499) aus Neuron (1,63).

Die anderen Reaktionen sind für die Pyrazolsynthese weniger geeignet, da Imin-, Pyridin- oder Indazol-Derivate aufgebaut werden.

Die Suche und Bewertung von strategischen Bindungen im Pyrazolderivat **A** zeigt, daß die Synthese ausgehend von einer 1,3-Dicarbonylverbindung und einem Hydrazinderivat sowohl an fester Phase als auch in flüssigem Medium durchgeführt werden kann. Der entsprechende Bindungsbruch erhält in beiden Fällen eine hohe Bewertung. Sieht man von den Reaktionen mit der höchsten Bewertung einmal ab, die für den Aufbau des Ringsystems nicht verwendet werden können, so werden die Knorr-Pyrazol-Synthesen am höchsten bewertet. Im flüssigen Medium weist das trainierte Netz auch noch auf Alternativen bei den 1,3-Dicarbonylverbindungen als Edukte hin.

5.5 Syntheseplanungsbeispiel: 2-Oxo-5-phenyl-cyclopentan-carbonsäuremethylester

2-Oxo-5-phenyl-cyclopentan-carbonsäuremethylester **A** wurde von Prof. Troschütz vom Pharmazeutischen Institut der Universität Erlangen-Nürnberg als Zielmolekül für eine WODCA Studie vorgeschlagen. Für dieses Molekül wurde daraufhin eine detaillierte Syntheseplanungsstudie ausgearbeitet[81].

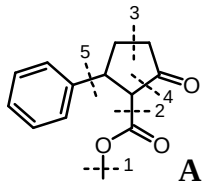


A

Abb. 5-13: 2-Oxo-5-phenyl-cyclopentan-carbonsäuremethylester **A** als Zielmolekül für eine Suche und Bewertung von strategischen Bindungen. Die Zahlen sind die Bindungsnummern.

5.5.1 Bestimmung und Bewertung strategischer Bindungen basierend auf der Theilheimer Reaktionsdatenbank

Basierend auf der klassifizierten Theilheimer Reaktionsdatenbank werden in diesem Molekül **A** strategische Bindungen ermittelt und diese bewertet. Es werden für alle diejenigen Bindungen nach den ähnlichsten Reaktionen gesucht, an denen keine Wasserstoffatome beteiligt sind und die nicht in carbocyclischen aromatischen Systemen eingebunden sind. Bindungen zu Wasserstoffatomen werden bei der Suche ausgeschlossen, da man aus den Reaktionsdatenbanken nur Information zum Aufbau von Bindungen ableiten könnte, an denen spezielle Wasserstoffatome beteiligt sind. Zur Bewertung der strategischen Bindungen werden diese nach der Häufigkeit der gefundenen Reaktionsbeispiele sortiert. Die ersten fünf Bindungen mit den meisten Reaktionsbeispielen sind in Abbildung 5-14 wiedergegeben.



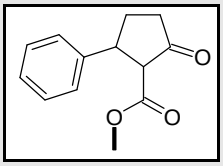
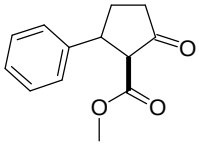
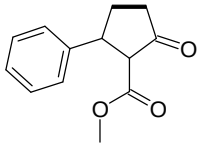
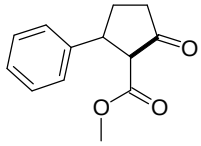
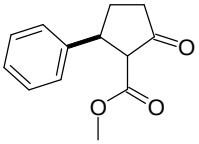
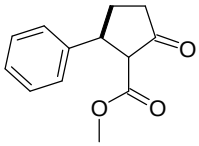
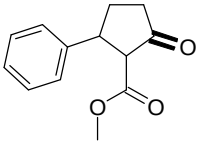
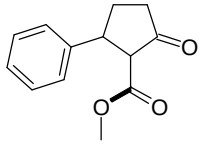
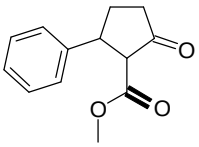
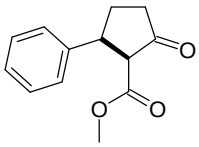
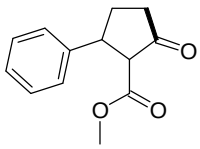
A

Abb. 5-14: Die ersten fünf strategischen Bindungen des Targetmoleküls **A** mit den höchsten Bewertungen.

Eine detaillierte Aufstellung der ähnlichsten Reaktionen der elf nicht-aromatischen C-C- und C-O-Bindungen des Targets sind in Tabelle 5-3 wiedergegeben.

Die meisten ähnlichen Reaktionen findet man zu der Anfragereaktion, die die Alkoxybindung des Esters spaltet. Die 52 gefundenen Reaktionen verdeutlichen, daß es sich bei der Alkylierung eines Carboxylats um eine häufig durchgeführte Reaktion handelt, die zumeist in hoher Ausbeute abläuft. Andererseits kann durch eine solche Reaktion das Grundgerüst dieses Targets nicht aufgebaut werden, so daß nach besser geeigneten Bindungsbrüchen gesucht werden muß. Die Retrosynthese der 2-Oxo-5-phenyl-cyclopentan-carbonsäure **B** wird daher nicht weiterverfolgt. Eine Zusammenstellung des gesamten Syntheseplans ist in Abbildung 5-22 abgebildet.

In der Bewertung an zweiter Stelle mit 13 Reaktionsbeispielen folgt die Knüpfung der Estergruppe an den Cyclopentanonring. Allen diesen Beispielen liegt eine nucleophile Substitution an einer Carbonylgruppe mit tetraedrischer Zwischenverbindung zugrunde. Die 13 Reaktionsbeispiele setzen dazu beispielsweise Oxalsäuredialkylester **E** (#23633, #45710 und

(79,90)  insg. 52 Reaktionen (hier 1-20): #447 (98%), #2992 (94%), #4118 (74%), #4207 (k.A.), #4394 (90%), #7556 (92%), #7687 (95%), #9849 (86%), #12377 (91%), #13279 (70%), #13499 (k.A.), #15355 (90%), #17066 (84%), #18979 (k.A.), #20390 (k.A.), #20391 (65%), #20482 (97%), #21215 (92%), #21743 (71%), #22614 (82%), ...	(88,41)  #10541 (95%), #18179 (71%), #23632 (k.A.), #23633(87-92%), #24023 (90%), #31167 (k.A.), #35298 (68-69%), #42606 (84%), #44096 (66%), #44556 (k.A.), #45710 (60%), #46046 (74%), #46719 (70%)	(70,69) u. (68,63)  #2975 (90%), #6510 (95%), #14591 (90%), #17433 (75-85%), #24000 (60%), #26114 (100%), #28522 (78%), #30537 (84%), #36289 (90%)	(87,47)  #4428 (80%), #4429 (k.A.), #18967 (k.A.), #18968 (k.A.), #19452 (82%), #41689 (91%), #43458 (k.A.), #43462 (66%)
(46,45) u. (92,85)  #688 (82%), #1751 (64%), #5566 (85%), #14824 (k.A.), #33155 (70%), #41070 (79%), #41071 (67%)	(70,69) u. (60,61)  #2975 (90%), #17637 (97%), #20728 (75-85%), #24360 (80%), #35252 (78%)	(32,92)  #15967 (85%), #20497 (85%), #20498 (75%), #27306 (94%), #35303 (81%)	(76,54)  #17426 (k.A.), #42609 (k.A.), #43251 (92%), #46045 (83%), #46630 (90-94%)
(41,84)  #3183 (k.A.), #12452 (k.A.), #17790 (k.A.), #26932 (k.A.), #44859 (91-93%)	(60,53) u. (61,63)  #4278 (86%), #37418 (k.A.)	(87,43)  #34573 (69%)	

Tab. 5-3: Ergebnis der Suche nach strategischen Bindungen für das Zielmolekül **A**: Für die angegebenen 11 Bindungskombinationen sind die ähnlichsten Reaktionen aus der Theilheimer Reaktionsdatenbank angegeben, die dem jeweiligen Gewinnerneuron entnommen sind.

#46046), Kohlensäuredialkylester **D** (#31167, #42606) oder Acyl-dialkyl-phosphonsäuretriester **F** (#18179) mit einer C-H-aciden Verbindung um (siehe Abbildung 5-15).

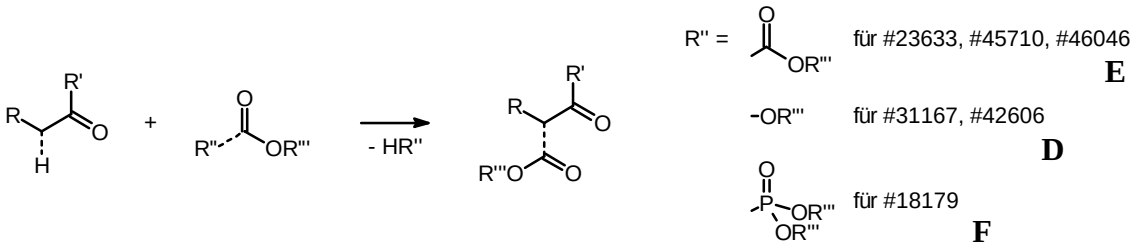
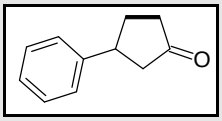
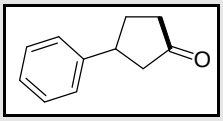
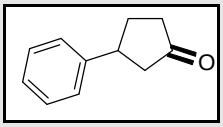
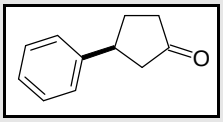
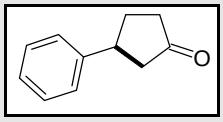
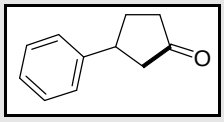
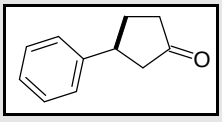


Abb. 5-15: Die ähnlichsten Reaktionsbeispiele aus der Theilheimer Reaktionsdatenbank zur Knüpfung der Bindung 20.

Eine Übertragung dieses Reaktionstyps auf die Anfragereaktion ergibt einen weiteren Retrosyntheseschritt in dem Syntheseplan der Abbildung 5-22. Während die Esterderivate **D**, **E** und **F** meist als käufliche Verbindung nicht weiter in synthetisierbare Vorstufen zerlegt werden müssen, muß für 3-Phenyl-cyclopentanon **C** (siehe Abbildung 5-22) der Retrosyntheseweg weiter verfolgt werden. Dazu wird für dieses neue Ausgangsmolekül wiederum eine Suche und Bewertung der strategischen Bindungen durchgeführt. Man beschränkt sich auch hier auf alle Bindungen, an denen keine Wasserstoffatome beteiligt sind und die nicht in carbocyclischen aromatischen Systemen eingebunden sind. Das Ergebnis dieser Suche ist in Tabelle 5-4 gezeigt.

(70,69) u. (68,63)	(88,44)	(29,92)	(46,45) u. (92,85)
			
#2975 (90%), #6510 (95%), #14591 (90%), #17433 (75-85%), #24000 (60%), #26114 (100%), #28522 (78%), #30537 (84%), #36289 (90%)	#29584 (93%), #32001 (73%), #35851 (79%), #38752 (95%), #42280 (83%), #43165 (72%), #44843 (60%), #44849 (k.A.), #46428 (70%)	#2943 (94-98%), #10595 (38%), #10596 (38%), #16261 (95%), #16262 (86%), #19113 (k.A.), #32761 (60%), #33349 (k.A.)	#688 (82%), #1751 (64%), #5566 (85%), #14824 (k.A.), #33155 (70%), #41070 (79%), #41071 (67%)
(60,63) u. (66,64)	(87,44)	(70,69) u. (60,61)	
			
#8650 (100%), #14851 (50%), #24031 (67%), #28015 (80%), #37697 (94%), #39369 (89%)	#1212 (73%), #28004 (56-63%), #32003 (75%), #32746 (91%), #40762 (71%), #43460 (82-87%)	#2975 (90%), #17637 (97%), #20728 (75-85%), #24360 (80%), #35252 (78%)	

Tab. 5-4: Ergebnis der Suche nach strategischen Bindungen für das Zwischenprodukt **C**: Für die angegebenen 7 Bindungskombinationen sind die ähnlichsten Reaktionen aus der Theilheimer Reaktionsdatenbank angegeben, die dem jeweiligen Gewinnerneuron entnommen sind.

Im Gegensatz zur vorangegangenen Untersuchung sind bei diesem Targetmolekül **C** einige Bewertungen der strategischen Bindungen sehr ähnlich, da für die Bindungen häufig gleich viele Reaktionsbeispiele gefunden werden. Aus diesem Grund kann man hier keine bevorzugte strategische Bindung angeben, sondern sollte zumindest die ersten vier Bindungen genauer analysieren.

In der Bewertungsreihenfolge der strategischen Bindungen stehen an erster Stelle Reaktionen, die die Bindung Nummer 16 (siehe Abbildung 5-13) im Fünfring behandeln. Die ähnlichsten Reaktionen bauen allerdings diese Bindung nicht auf, sondern hydrieren meist nur eine Doppelbindung im Ring zu einer Einfachbindung, wie beispielsweise in Reaktion #6510 (siehe Abbildung 5-16).

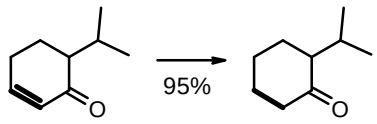


Abb. 5-16: Reaktion #6510 aus der Theilheimer Reaktionsdatenbank (RTHE00038570) aus Neuron (68,63).

Danach folgen in der Bewertung Reaktionen, die alle aus einer Dicarbonsäure ein Cycloalkanon-Derivat aufbauen. Als Beispiel für diesen Reaktionstyp wurde Reaktion #46428 in Abbildung 5-17 wiedergegeben.

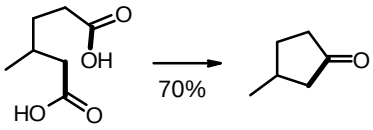


Abb. 5-17: Reaktion #46428 aus der Theilheimer Reaktionsdatenbank (RTHE00001546) aus Neuron (88,44).

Obwohl die Ausbeute für diese Reaktion nur 70% beträgt, läuft die Reaktion #38752, die ebenfalls zu einem Cyclopentanon-Derivat führt, nämlich 3,3,4-Trimethylcyclopentanon, mit 95% Ausbeute ab. Somit stellt dieser Reaktionstyp eine gute Synthesemöglichkeit für das Targetmolekül **C** dar. Übertragen auf das vorliegende Targetmolekül würde der strategische Bindungsbruch zur 3-Phenyladipinsäure **I** führen. Dieser Retrosyntheseweg ist ebenfalls in dem Syntheseplan der Abbildung 5-22 aufgenommen und wird später weiterverfolgt.

An dritter Stelle in Tabelle 5-4 folgen Reaktionen, die entweder durch Oxidation von sekundären Alkoholen oder durch Spaltung einer Doppelbindung eine Carbonylgruppe erzeugen. Da diese Reaktionen nur die Funktionalität verändern, das Targetmolekül aber nicht zerlegen können, wird nach besseren strategischen Bindungen gesucht.

An vierter Stelle in der Bewertung folgt ein retrosynthetischer Bindungsbruch, der das Targetmolekül in zentraler Lage in zwei annähernd gleich große Fragmente zerlegt. In der Reaktion #5566, die in Neuron (92,85) eingetragen wird, wird eine Alkylierungsreaktion mit anschließender Reduktion ausgehend von einer Carbonylgruppe und Brombenzol durchgeführt. Unter Einsatz von Lithium setzen sich beide Edukte zunächst zu einer Hydroxyverbindung um, die anschließend in flüssigem Ammoniak mit Lithium reduziert wird (siehe Abbildung 5-18)[82].

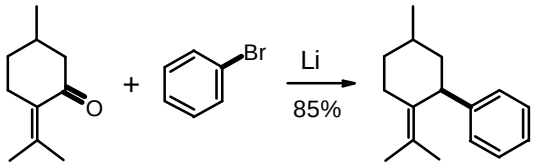


Abb. 5-18: Reaktion #5566 aus der Theilheimer Reaktionsdatenbank (RTHE00041031) aus Neuron (92,85).

Dieser metallorganische Reaktionstyp stellt somit eine gute Synthesemöglichkeit für das Molekül **C** dar, zumal die Ausbeuten dieser ähnlichsten Reaktion von 85% relativ hoch ausfällt. Die Retrosynthese von Molekül **C** basierend auf dieser ähnlichen Reaktion #5566 führt zu Brombenzol **G** und Cyclopentadion **H** (siehe Abbildung 5-22).

Der nächste in Tabelle 5-4 angegebenen Bindungsbruch der Bindung 14 ist für die Retrosynthese des Targetmoleküls weniger von Bedeutung. Die als ähnlichste Beispiele ausgegeben Reaktionen in den Neuronen (60,63) und (66,64) stellen Hydrierungsreaktionen dar, die eine C-C-Einfachbindung im Produkt durch Hydrierung einer Doppelbindung erzeugen.

Aufgrund der physikochemischen Effekte wird der Bruch an Bindung 19 anders bewertet als die bereits diskutierte strategische Bindung 22 (siehe Abbildung 5-13). Bei Letztgenannter zeigen die induktiven und mesomeren Effekte des Phenylrings aufgrund der zusätzlichen, homologen CH_2-Gruppe kleinere Auswirkungen, so daß die unterschiedliche Bewertung der beiden Bindungen 19 und 22 und die damit einhergehenden unterschiedlichen ähnlichsten Reaktionsbeispiele verständlich werden. Für die strategische Bindung 19 werden nicht nur Reaktionen vorgeschlagen, die wie für Bindung 22 aus Dicarbonsäuren ein Cycloalkanon bilden, sondern auch Reaktionen, die beispielsweise durch Ringverengung Cycloalkanone erzeugen.

Schließlich wird aus Neuron (60,61) beispielsweise Reaktion #20728 (siehe Abbildung 5-19) zum Aufbau der strategischen Bindung 13 vorgeschlagen, die als Cyclopropan-Ringsynthese aus 1,3-Dibrom-1-phenyl-propan und Zinkpulver zum Aufbau des Targets **C** wenig geeignet erscheint.

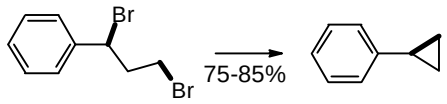
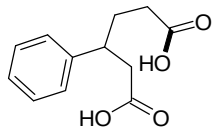
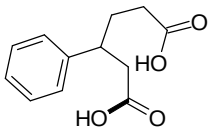
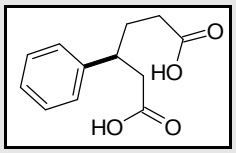
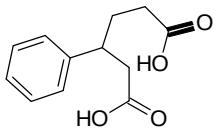
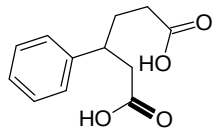
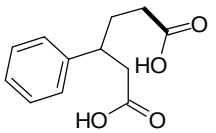
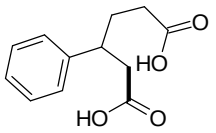
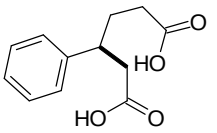
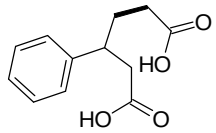
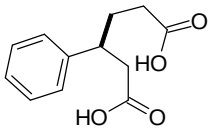


Abb. 5-19: Reaktion #20728 aus der Theilheimer Reaktionsdatenbank (RTHE00028950) aus Neuron (60,61).

Da 3-Phenyladipinsäure **I** nicht als kommerzielles Produkt verfügbar ist, wird für dieses Target abermals eine Syntheseplanung durchgeführt. Insgesamt werden 78 strategische Bindungskombinationen gefunden. Davon sind in Tabelle 5-5 auszugsweise nur die ersten zehn Bindungskombinationen nach Reaktionsbeispielen sortiert gezeigt.

Man erkennt, daß sechs strategische Bindungen paarweise gleich behandelt werden, da der mesomere und induktive Effekt der Phenylgruppe nur die unmittelbar angrenzenden Bindungen beeinflußt. Für die strategischen Bindungen in Nachbarschaft zur Phenylgruppe (in Tabelle 5-5 an Position 8 und 10) werden unterschiedliche ähnliche Reaktionen gefunden, während für die endständigen Bindungen der aliphatischen Kette (in Tabelle 5-5 an Position 6 und 7) jeweils dieselben ähnlichen Reaktionen gefunden werden.

An erster und zweiter Stelle in der Bewertung stehen Reaktionen, bei denen die beiden Hydroxgruppe der Säureeinheit gebildet werden. In den ähnlichen Reaktionen werden dazu hauptsächlich die entsprechenden Ester gespalten. An dritter Stelle in der Bewertung folgen

(76,52)  insg. 62 Reaktionen (hier 1-20): #4057 (99%), #4294 (82%), #6991 (93%), #9460 (100%), #11168 (92%), #13864 (k.A.), #14663 (70%), #15048 (k.A.), #16407 (67-85%), #17565 (62%), #19734 (67%), #20224 (k.A.), #20824 (k.A.), #21739 (82%), #22432 (95%), #23644 (k.A.), #24768 (k.A.), #25943 (k.A.), #26142 (84%), #26656 (k.A.), ...	(76,52)  insg. 62 Reaktionen (hier 1-20): #4057 (99%), #4294 (82%), #6991 (93%), #9460 (100%), #11168 (92%), #13864 (k.A.), #14663 (70%), #15048 (k.A.), #16407 (67-85%), #17565 (62%), #19734 (67%), #20224 (k.A.), #20824 (k.A.), #21739 (82%), #22432 (95%), #23644 (k.A.), #24768 (k.A.), #25943 (k.A.), #26142 (84%), #26656 (k.A.), ...	(92,85) u. (46,45)  insg. 41 Reaktionen (hier 1-20): #2241 (73%), #2666 (k.A.), #2667 (95%), #2669 (96%), #5789 (93%), #5790 (k.A.), #6283 (92%), #6284 (k.A.), #7046 (79%), #7052 (72%), #7954 (88%), #7955 (k.A.), #9109 (88-97%), #9541 (95%), #11248 (60%), #12338 (70%), #17388 (83%), #23996 (95%), #25448 (60-80%), #28201 (78%)..	(40,85)  insg. 28 Reaktionen (hier 1-20): #495 (75%), #1318 (82%), #3125 (k.A.), #5251 (k.A.), #6523 (k.A.), #7276 (100%), #10421 (90%), #13367 (90%), #15415 (100%), #15421 (70%), #17313 (98-99%), #21217 (94%), #21641 (k.A.), #21846 (98%), #25711 (82%), #26388 (87%), #26906 (90%), #31138 (85%), #31712 (93%), #32209 (k.A.), ...
(40,85)  insg. 28 Reaktionen (hier 1-20): #495 (75%), #1318 (82%), #3125 (k.A.), #5251 (k.A.), #6523 (k.A.), #7276 (100%), #10421 (90%), #13367 (90%), #15415 (100%), #15421 (70%), #17313 (98-99%), #21217 (94%), #21641 (k.A.), #21846 (98%), #25711 (82%), #26388 (87%), #26906 (90%), #31138 (85%), #31712 (93%), #32209 (k.A.), ...	(87,39) u. (46,45)  #3049 (80-90%), #8874 (52%), #10343 (66%), #19810 (78%), #21789 (90%), #22966 (98%), #24740 (k.A.), #31050 (100%), #32763 (63%), #33950 (94%), #36033 (85%), #39587 (76%), #42630 (k.A.), #45378 (k.A.), #45686 (40-73%), #46007 (87%), #46521 (80%)	(87,39) u. (46,45)  #3049 (80-90%), #8874 (52%), #10343 (66%), #19810 (78%), #21789 (90%), #22966 (98%), #24740 (k.A.), #31050 (100%), #32763 (63%), #33950 (94%), #36033 (85%), #39587 (76%), #42630 (k.A.), #45378 (k.A.), #45686 (40-73%), #46007 (87%), #46521 (80%)	(68,63) u. (61,63)  #1593 (81%), #3200 (90%), #3485 (60%), #4639 (80%), #4827 (k.A.), #8393 (93%), #10005 (83%), #14738 (k.A.), #15786 (95%), #19871 (79%), #20772 (k.A.), #28379 (89%), #36087 (77%), #43436 (70%), #45236 (47%)
(67,62)  #501 (91%), #2270 (100%), #3582 (50%), #10002 (95%), #10003 (89%), #10634 (85%), #15402 (95%), #16830 (90%), #34000 (k.A.), #34494 (51%), #35215 (45-55%), #44087 (55%), #44091 (65%), #45246 (k.A.), #45247 (60-65%)	(60,61)  #43188 (92%), #46277 (k.A.)		

Tab. 5-5: Ergebnis der Suche nach strategischen Bindungen für das Zwischenprodukt **I**: Für die angegebenen zehn Bindungskombinationen sind die ähnlichsten Reaktionen aus der Theilheimer Reaktionsdatenbank angegeben, die dem jeweiligen Gewinnerneuron entnommen sind.

41 Reaktionsbeispiele, die die Bildung zwischen dem Phenylring und der aliphatischen Kette aufbauen. In den vorangegangenen Studien fand man für diese Bindung nur sieben ähnliche Reaktionsbeispiele (siehe Tabellen 5-3 oder 5-4), da der Aufbau einer Bindung zwischen dem

Phenylring und einem cyclischen Aliphaten gesucht wurde. Die Öffnung des Ringes führt nun zu 41 Reaktionsbeispielen, bei denen es sich hauptsächlich um metallorganische Reaktionen, Chlormethylierungsreaktionen und Friedel-Crafts-Reaktionen handelt. Chlormethylierungsreaktionen sind wegen der entstehenden Funktionalität in β-Position zum Phenylring für die Anfragereaktion nicht geeignet. Bei der in Abbildung 5-20 dargestellten Reaktion #2667 handelt es sich um eine Grignard-Reaktion, die ausgehend von Brombenzol und Magnesium sowie einem primären Alkohol 1,3-Diphenyl-1-propen bildet.

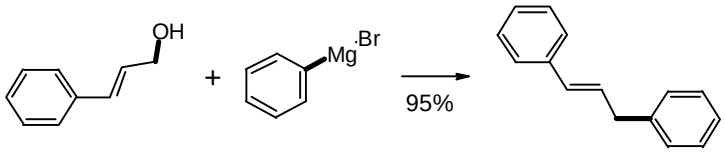


Abb. 5-20: Reaktion #2667 aus der Theilheimer Reaktionsdatenbank (RTHE00045157) aus Neuron (92,85).

Übertragen auf das Targetmolekül würde der retrosynthetische Bindungsbruch zu den in Abbildung 5-21 gezeigten Ausgangsverbindungen führen.

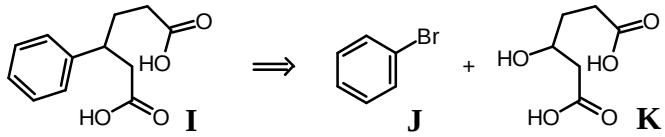


Abb. 5-21: Übertragung der ähnlichsten Reaktion #2667 auf die Synthese der Ausgangsverbindung **I**.

Allerdings muß man bei der Umsetzung dieser beiden Ausgangsverbindungen **J** und **K** in synthetischer Richtung beachten, daß das Molekül **K** leicht deprotoniert werden kann und die entsprechende Grignardverbindung **J'** zerstört. Daher müßten die Moleküle **J'** und **K** im Mengenverhältnis 3:1 eingesetzt werden, um zunächst die beiden Carbonsäuren in die Carboxylate zu überführen. Diese sind dann nicht mehr elektrophil, so daß die Reaktion von **J'** mit einer aktivierten Form der Alkoholgruppe **K'** zu der Ausgangsverbindung **I** reagiert.

Wie oben erwähnt, findet man unter den 41 ähnlichen Reaktionsbeispielen auch Friedel-Crafts-Reaktionen. Bei den Friedel-Crafts-Reaktionen stellt sich allerdings ein ähnliches Problem wie bei der metallorganischen Reaktion, nämlich das einer zu Nebenreaktionen neigenden Umsetzung. Im synthetischen Schritt der Retroreaktion von Molekül **I** könnte die Ausgangsverbindung **K** mit Benzol und einer Lewis-Säure über eine Alkylierungsreaktion zum gewünschten Target **I** reagieren oder auch über eine unerwünschte Friedel-Crafts-Acylierungsreaktion verlaufen.

Die restlichen, ermittelten strategischen Bindungen in der Bewertungsabfolge des Targetmoleküls **I** (siehe Tabelle 5-5) werden häufig wieder durch Oxidationsreaktionen, wie im Falle der Carbonylgruppen, oder Hydrierungsreaktionen, wie im Falle der C-C-Bindungen, eingeführt. Andere Reaktionstypen, wie eine nucleophile Substitutionsreaktion einer C-H-aciden Verbindung an einer Kohlenstoff-Halogen-Bindung (Reaktion #3200), sind ebenfalls nicht selektiv.

Zusammenfassend werden in Abbildung 5-22 die in diesem Kapitel diskutierten Retrosynthesewege nochmals dargestellt.

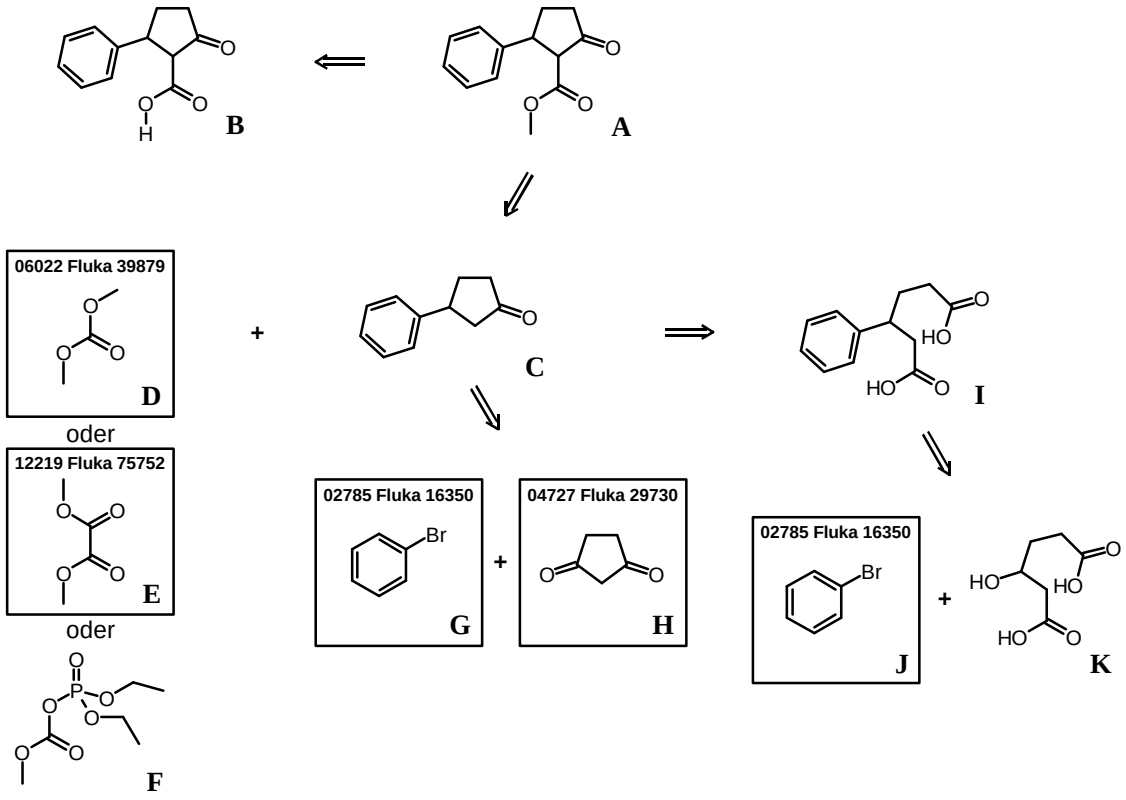
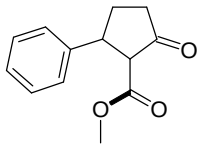
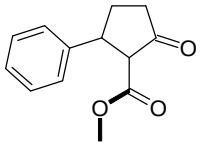
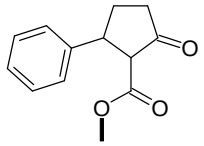
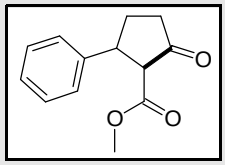
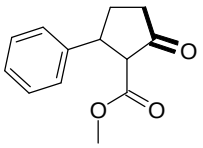
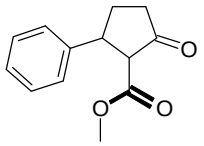
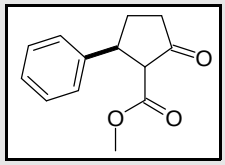
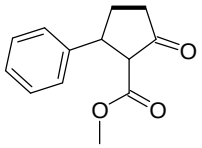
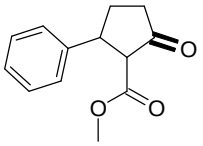
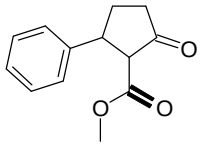
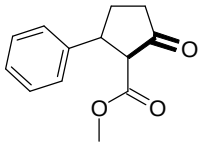
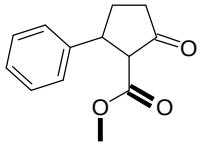


Abb. 5-22: Retrosynthetische Analyse von 2-Oxo-5-phenyl-cyclopentan-carbonsäuremethylester **A**.

5.5.2 Bestimmung und Bewertung strategischer Bindungen basierend auf der SPORE Reaktionsdatenbank

Die Suche nach strategischen Bindungen im 2-Oxo-5-phenyl-cyclopentan-carbonsäuremethylester **A** (siehe Abbildung 5-13) basierend auf der klassifizierten SPORE Reaktionsdatenbank führt zu den in Tabelle 5-6 wiedergegebenen ähnlichen Reaktionen.

Auch hier werden wieder die Reaktionen im Gewinnerneuron und die Reaktionen in der ersten Nachbarschaftssphäre als ähnliche Reaktionen angesehen. Obwohl alle Kombinationen bis maximal 6 Bindungen durchlaufen werden, werden nur für 12 Bindungskombinationen ähnliche Reaktionen gefunden. Insgesamt werden 118 Reaktionen ausgegeben, die aus 21 Neuronen stammen. Die Reaktionen mit den häufigsten Reaktionsbeispielen behandeln alle den Aufbau der Estergruppe (siehe die ersten drei Zellen in Tabelle 5-6). Anschließend folgt ein Reaktionstyp, der zu einer Vereinfachung des Ringsystems führt. Dabei handelt es sich

(76,54)^{G}, (75,53)^{1}, (75,54)^{1}, (76,53)^{1}  insg. 54 Reaktionen (hier 1-20): #568 (32%), #707 (k.A.), #887 (40%), #952 (k.A.), #975 (100%), #1110 (55%), #1111 (55%), #1114 (58%), #1115 (55%), #1118 (57%), #1119 (51%), #1260 (k.A.), #1262 (k.A.), #1389 (9%), #1489 (81%), #1616 (52%), #2159 (k.A.), #2431 (k.A.), #2580 (k.A.), #2587 (k.A.), ...	(68,34)^{G}, (67,34)^{1}, (68,35)^{1}  insg. 34 Reaktionen (hier 1-20): #23 (55%), #24 (90%), #299 (k.A.), #356 (36%), #357 (33%), #436 (94%), #437 (98%), #438 (77%), #439 (86%), #443 (67%), #576 (85%), #577 (64%), #579 (k.A.), #670 (k.A.), #671 (k.A.), #807 (88%), #1017 (90-95%), #1054 (k.A.), #1063 (100%), #1386 (23%), ...	(79,90)^{G}, (79,91)^{1}, (80,89)^{1}, (80,90)^{1}, (80,91)^{1}  #280 (90%), #281 (93%), #282 (97%), #283 (73%), #284 (60%), #285 (0,1%), #286 (91%), #2470 (90%), #4148 (52%), #4408 (63%), #4885 (65%), #4886 (69%), #5663 (k.A.), #5800 (k.A.), #5801 (k.A.), #5806 (k.A.), #6335 (k.A.), #6483 (k.A.)	(87,46)^{1}  #56 (46%), #57 (15%)
(20,6)^{1}  #4580 (87%), #4581 (58%)	(27,10)^{1}  #651 (68%), #654 (48%)	(92,86)^{1}  #3512 (k.A.)	(67,64)^{1}  #4146 (k.A.)
(32,92)^{G}  #1718 (81%)	(40,85)^{1}  #649 (81%)	(20,9)^{1}  #2324 (k.A.)	(21,12)^{1}  #2078 (48%)

Tab. 5-6: Ergebnis der Suche nach strategischen Bindungen für das Zielmolekül **A**: Für die angegebenen 12 Bindungskombinationen sind die ähnlichsten Reaktionen aus der SPORE Reaktionsdatenbank angegeben, die dem Gewinnerneuron$^{(G)}$ und der ersten Nachbarschaftssphäre$^{(1)}$ entnommen sind.

um Reaktion #56, in der eine intramolekulare nucleophile Substitutionsreaktion ausgeführt wird, die zu einem Cyclohexanon-Derivat führt (siehe Abbildung 5-23).

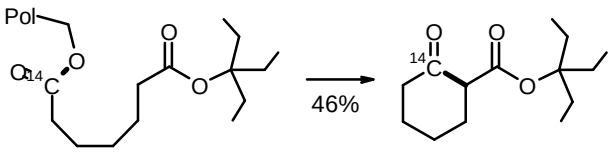


Abb. 5-23: Reaktion #56 aus der SPORE Reaktionsdatenbank (RSPO69000061) aus Neuron (87,46).

Allerdings beträgt die Ausbeute für diese Reaktion nur 46%. Für Reaktion #57, bei der das Eduktmolekül eine zusätzliche Ethylgruppe trägt und auf derselben Weise reagiert wie Reaktion #56, wird die Ausbeute sogar mit nur 15% angegeben. Für die angefragte Reaktion ist daher in Syntheserichtung auch keine hohe Ausbeute zu erwarten. Danach folgen in der Bewertung der strategischen Bindungen zwei Bindungskombinationen, bei denen die ähnlich-

sten Reaktionen ausgehend von einem Enamin oder einem Imin eine Carbonylgruppe aufbauen. Da in den Eduktmolekülen die Funktionalität schon enthalten ist, sind diese Reaktionstypen weniger geeignet, um das Targetmolekül aufzubauen. Die nächste Reaktion #3512 aus einem Neuron in der ersten Nachbarschaftssphäre behandelt die Knüpfung des Phenylrings mit dem Cyclopentanonring. Allerdings handelt es sich bei der Reaktion aus der SPORE Reaktionsdatenbank um eine intramolekulare Cyclisierungsreaktion, bei der der Ringschluß gegenüber der intermolekularen Anfragereaktion erheblich erleichtert wird.

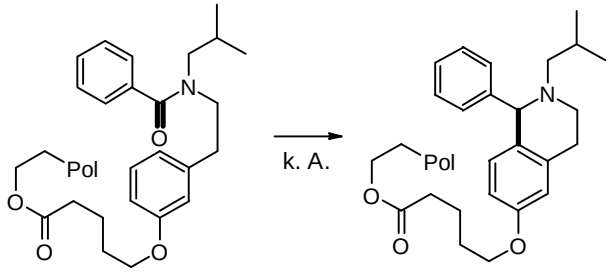


Abb. 5-24: Reaktion #3512 aus der SPORE Reaktionsdatenbank (RSPO69004028) aus Neuron (92,86).

Für Bindung Nummer 16 (siehe Abbildung 5-13) findet man als ähnlichste Reaktion #4146, bei der die C-C-Einfachbindung durch Hydrierung einer Doppelbindung aufgebaut wird. Verfolgt man diesen Reaktionsweg weiter, so schlägt das trainierte Netz zum Aufbau der Doppelbindung im Cyclopentenon-System mit den ähnlichen Reaktionen #6150 und #6154 eine Wittig-Reaktion vor.

Die restlichen in Tabelle 5-5 aufgezeigten Reaktionen behandeln die Einführung einer Carbonylgruppe, eventuell mit weiteren Bindungen. Dazu werden ausschließlich Oxidationsreaktionen oder Reaktionen von Enaminen eingesetzt, so daß auch diese Reaktionen nicht zu einer Vereinfachung des Targetmoleküls führen.

Die Suche nach strategischen Bindungen basierend auf der klassifizierten SPORE Reaktionsdatenbank ergibt also für dieses Molekül keinen geeignet erscheinenden Syntheseweg. Es werden hauptsächlich Reaktionen gefunden, die die Estergruppierung oder die Carbonylgruppe durch Oxidation eines sekundären Alkohols aufbauen. Aus diesem Grund sollte das Zielmolekül **A** bevorzugt über die in Kapitel 5.5.1 diskutierten (Retro-)Synthesewege, d.h. in einem Lösungsmittel, aufgebaut werden. Möchte man trotzdem das Molekül an fester Phase synthetisieren, so müßten zunächst neue Synthesewege zum Aufbau der strategischen Bindungen gefunden werden, wobei man momentan in der Literatur über Festphasensynthesen nicht viel Information finden wird.

5.6 Diskussion des Einsatzes der Reaktionsklassifizierung in der Syntheseplanung

Abschließend sollen die Vor- und Nachteile des auf der Reaktionsklassifizierung beruhenden Verfahrens der Syntheseplanung diskutiert werden.

- Schnelle und vollständige Suche nach Synthesestrategien basierend auf physikochemischen Effekten
  Die Suche nach Synthesestrategien zum Aufbau eines Targetmoleküls liefert innerhalb weniger Sekunden eine Reihe von Reaktionsbeispielen. Im Targetmolekül werden dabei alle Kombinationen aus bis zu 6 Bindungen systematisch durchlaufen. Die Suche nach der geeignetsten Synthesestrategie ist auf der Basis der codierten Reaktionsdatenbank somit vollständig.
- Implizite Klassifizierung der Trefferliste
  Reaktionen innerhalb eines Neurons weisen sehr ähnliche physikochemische Deskriptoren auf und basieren zumeist auf demselben Reaktionstyp. Daher ist es nicht notwendig, alle Reaktionen einer Trefferliste zu analysieren, sondern lediglich einige wenige Beispiele aus jedem Gewinnerneuron.
- Flexible und erweiterbare Wissensbasis
  Die Wissensbasis ist jederzeit austausch- und erweiterbar. Somit können nicht nur naßchemische Synthesestrategien für die präparative Chemie vorgeschlagen werden. Alternativ zu den ebenfalls vorgestellten Reaktionen an fester Phase können Reaktionswege auch für großtechnische Synthesen oder metallorganische Synthesen vorhergesagt werden. Voraussetzung dafür ist allerdings das Vorhandensein einer Datenbank mit vielen verschiedenen Reaktionsbeispielen zum Generieren einer Wissensbasis.
- Vorschläge zu den Reaktionsbedingungen werden gegeben
  Falls zu den Reaktionsbeispielen einer Datenbank zusätzliche Angaben, wie eingesetzte Lösungsmittel und Katalysatoren abgespeichert wurden, kann auf diese chemische Information bei der Syntheseplanung zurückgegriffen werden. Für die übertragene Reaktion, die das Targetmolekül aufbaut, kann dann ein geeignetes Lösungsmittel und eventuell ein benötigter Katalysator vorgeschlagen werden.
- Keine Einarbeitungszeit in Retrieval-Systeme erforderlich
  Die Suche in Reaktionsdatenbanken nach Reaktionen, die der geplanten Synthese einer Zielverbindung ähnlich sind, ist direkt im Syntheseplanungsprogramm realisierbar. Somit entfällt eine zeitintensive Einarbeitungszeit in die verschiedenen Retrieval-Systeme verschiedener Anbieter und das korrekte Formulieren einer Suchanfrage.

Neben dieser Reihe von Vorteilen zeigt die Syntheseplanung basierend auf der Reaktionsklassifizierung aber auch folgende Nachteile:

- Datenbank mit einer Vielzahl an Reaktionen muß vorhanden sein
  Um retrosynthetische Informationen ableiten zu können, muß man auf möglichst viele Reaktionen zurückgreifen können. Da heutzutage bereits Millionen von Reaktionen in kommerziell erhältlichen Datenbanken zugänglich sind, ist diese Grundvoraussetzung meistens erfüllt. Darüber hinaus besitzen viele Chemieunternehmen auch firmeninterne Datenbanken, in denen sämtliche chemische Information, die in den Labors erarbeitet wurden, über die Jahre hinweg angesammelt wird. Diese Datenbanken stellen ebenfalls geeignete Quellen zum Aufbau einer Wissensbasis dar.
- Satz an Deskriptoren muß noch um weitere Effekte ergänzt werden
  Bisher werden sterische Einflüsse bei der Codierung der Reaktionen nicht berücksichtigt. Große raumerfüllende Gruppen in Nachbarschaft der reagierenden Atome oder Bindungen können beispielsweise einen großen Einfluß auf die Kinetik eines vorgeschlagenen Syntheseweges nehmen. Daher sollte in Zukunft eine Größe, die sterische Effekte beschreiben kann, zur Codierung der Reaktionen eingesetzt werden.
  Die Synthese von Naturstoffen oder Pharmazeutika erfolgt fast ausschließlich stereoselektiv. Um stereochemische Vorgänge bei organischen Reaktionen berücksichtigen zu können, wie beispielsweise Retention oder Inversion, sollte bei der Codierung der Reaktionen zukünftig auch die stereochemische Information erfaßt werden.

5.7 Anschluß an das Syntheseplanungsprogramm WODCA

Wie in den vorangegangenen zwei Anwendungsbeispielen gezeigt wurde, liefert die Suche und Bewertung strategischer Bindungen basierend auf der Reaktionsklassifizierung für die Syntheseplanung wertvolle Ergebnisse.

In WODCA sind zur Zeit verschiedene Methoden zur Generierung von Synthesevorstufen implementiert. Der Benutzer kann zwischen Strategien für aliphatische Bindungen, für Bindungen zu Benzolringen, für Kohlenstoff-Heteroatom-Bindungen und Strategien zum Aufbau polycyclischer Bindungen wählen. Zusätzlich zu diesen vier Methoden ist nun eine Suche und Bewertung strategischer Bindungen basierend auf der Reaktionsklassifizierung als weitere Strategie vorstellbar. Im Gegensatz zu den vier anderen Methoden kann diese neue Methode auf alle Bindungstypen angewendet werden. Der Benutzer muß sich nicht mehr für eine Strategie entscheiden, sondern kann die universell anwendbare neue Methode zur gleichzeitigen Bewertung strategischer aliphatischer, polycyclischer etc. Bindungen einsetzen.

Die entwickelte Methode zur Suche und Bewertung strategischer Bindungen wurde bereits teilweise in das Programmsystem WODCA aufgenommen, so daß diese Methode in naher Zukunft den Benutzern dieses Syntheseplanungsprogramms zur Verfügung stehen wird.

6 Praktische Anwendung: Reaktionsvorhersage

6.1 Computergestützte Reaktionsvorhersage

Bei der computergestützten Reaktionsvorhersage werden Programmsysteme zur Vorhersage des Reaktionsproduktes und eventuell entstehender Nebenprodukte sowie der Reaktivität bzw. Kinetik eingesetzt. Im Gegensatz zur Syntheseplanung werden hier die Edukte sowie die Reaktionsbedingungen vorgegeben. Um aus den Edukten mögliche Produkte zu erzeugen, kommen beispielsweise Reaktionsgeneratoren zum Einsatz (siehe Kapitel 2.1.3), die die Bindungsumordnungsprozesse festlegen. Bei diesem Ansatz wird somit auf das in einer Reaktionsdatenbank gespeicherte Wissen verzichtet, mit dem Vorteil, daß neue Reaktionen, die noch nicht in Datenbanken gespeichert sind, entdeckt werden können. Andererseits hat dieser Ansatz auch den gravierenden Nachteil, daß Reaktionen gebildet werden können, die sich experimentell nicht verwirklichen lassen. In Abbildung 6-1 wird dies am Beispiel von Ethanol gezeigt.

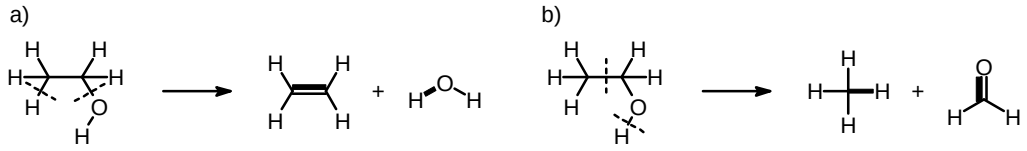


Abb. 6-1: Zwei mögliche Produkte bei Anwendung des Reaktionsgenerators 2.2:
a) die Abspaltung von Wasser aus Ethanol führt zur Bildung von Ethen;
b) die Spaltung der Kohlenstoff-Kohlenstoff-Bindung und der Sauerstoff-Wasserstoff-Bindung stellt eine neue Reaktion dar, die experimentell bisher nicht durchgeführt worden ist.

Um zwischen chemisch durchführbaren und theoretischen Reaktionen zu differenzieren, werden physikochemische Effekte von Atomen oder Bindungen, die am Elektronenumordnungsprozeß beteiligt sind, berechnet und ausgewertet. Zur Erzeugung von Produkten nach vorgegebenen Reaktionsregeln müssen diese Effekte innerhalb definierter Grenzen liegen, damit die Reaktionen auch in der Praxis durchführbar sind. Das im Arbeitskreis von Gasteiger entwickelte Syntheseplanungsprogramm EROS basiert auf diesem Ansatz.

Ein anderes Programm, genannt SOPHIA (***S**ystem for **o**rganic reaction **p**rediction by **h**euristic **a**pproach*), beruht auf einer anderen Methode zur Erzeugung von Reaktionen, das frei von chemischen Reaktionstypen (wie nucleophile Substitutionsreaktion etc.) oder Namensreaktionen Produkte zu willkürlichen vorgegebenen Edukten und Reaktionsbedingungen erzeugt[83]. Dazu greift SOPHIA auf eine Wissensbasis zu, in der Reaktionen nicht als Transformationen von Substrukturen auf der Edukt- und Produktseite enthalten sind, sondern die zuvor aus Reaktionsdatenbanken nach strukturellen Eigenschaften des Reaktionszentrums abstrahiert wurden.

Reaktionsvorhersagesysteme spielen heutzutage eine wichtige Rolle in der organischen Synthese, besonders im Bereich der chemischen Prozeßentwicklung und der kombinatorischen Synthese. In den folgenden Kapiteln wird das EROS Programmsystem näher vorgestellt, und ein Vorhersagemodell basierend auf der Reaktionsklassifizierung aufgestellt, mit dem man die Bildung unsymmetrisch substituierter, regioisomerer Pyrazole abschätzen kann.

6.2 Das EROS Programmsystem

Vor über 20 Jahren wurde im Arbeitskreis von Gasteiger die erste Programmversion eines Systems zur Reaktionsvorhersage entwickelt, das den Namen EROS trägt[84]. EROS ist eine Abkürzung für den Ausdruck „***E**laboration of **R**eactions for **O**rganic **S**ynthesis*“. Seit dieser Zeit wurde das Programmsystem immer weiter entwickelt; die letzte fertiggestellte Programmversion trägt den Namen EROS7[85]. Diese Programmversion zeichnet sich gegenüber der Vorgängerversion vor allem dadurch aus, daß sie auf einer molekülorbitalorientierten chemischen Datenstruktur[86] beruht, die es erlaubt, nicht nur organische Reaktionen, sondern auch anorganische und metallorganische Reaktionen vorherzusagen. Dieses als Expertensystem konzipierte System enthält eine abgekapselte Wissensbasis, in der Reaktionsregeln und eventuell kinetische Gleichungen angegeben werden. Eine solche Regeldatei muß für den Aufbau eines Reaktionsnetzwerkes entweder vorhanden sein, oder in den Programmiersprachen C++ oder Tcl neu implementiert werden.

Die Reaktivität einer Reaktion kann in EROS7 entweder gesetzt oder berechnet werden. Liegt zu jeder Reaktion des Netzwerkes eine Geschwindigkeitskonstante vor, so können unter Zuhilfenahme von drei zur Auswahl stehenden Methoden die Differentialgleichungen integriert werden und man erhält so in Abhängigkeit von der Zeit die Konzentrationen aller Substanzen. Falls nicht alle Geschwindigkeitskonstanten bekannt sind, so kann EROS die Geschwindigkeitskonstanten aus den physikochemischen Eigenschaften der Edukte bzw. Produkte mit einer mathematischen Funktion oder dem Aufruf neuronaler Netze abschätzen. Als neuronale Netztypen stehen Backpropagation-, Kohonen- und Counterpropagation-Netze zur Verfügung[87]. Beispielsweise konnten mit abgeleiteten mathematischen Gleichungen die Reaktivität von Hydrolysereaktionen verschiedener Amide und Benzoylphenylharnstoff-Derivate abgeschätzt werden[88]. Falls Geschwindigkeitskonstanten weder vorliegen, noch berechnet oder abgeschätzt werden können, so können zumindest noch über die Reaktionswahrscheinlichkeiten die Endkonzentrationen der Verbindungen berechnet werden.

Ein Ziel dieser Arbeit war es, Einsatzmöglichkeiten für neuronale Netze in der Reaktionsvorhersage aufzuzeigen, die mit Reaktionen aus Reaktionsdatenbanken trainiert wurden.

6.3 Bekannte Vorhersagemodelle zur Regioselektivität bei der Synthese von Pyrazolen

Bei Pyrazolen tritt Regioisomerie immer dann auf, wenn ein monosubstituiertes Hydrazin mit einer unsymmetrisch substituierten 1,3-Dicarbonylverbindung reagiert. Wie in Abbildung 6-2 dargestellt, können bei der Reaktion einer Hydrazinverbindung (mit R_1 ungleich H) und einer 1,3-Dicarbonylverbindung (mit R_2 ungleich R_4) zwei regioisomere Pyrazole gebildet werden. Oft wird eines der zwei Regioisomere bevorzugt gebildet. Welches der beiden Isomere entsteht, kann man beispielsweise experimentell durch eine massenspektroskopische Analyse aufklären[89]. Eine andere Methode beruht auf der Messung der vicinalen ^{15}N-H Kopplungskonstante nach Umsetzung von Hydrazinverbindungen des Typs Ph-NH-$^{15}NH_2$ [90]. Wünschenswert wäre ein Modell oder eine Methode, die bereits im Vorfeld des Experiments das bevorzugt gebildete Regioisomere vorhersagen kann.

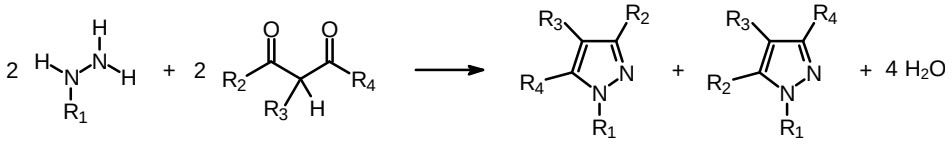


Abb. 6-2: Die Umsetzung monosubstituierter Hydrazine mit unsymmetrisch substituierten 1,3-Diketonen führt zu zwei Regioisomeren.

6.3.1 Reaktionsmechanismus

Als erstes sollte man versuchen, anhand des Reaktionsmechanismus eine Abschätzung der Regioselektivität zu erreichen. Leider ist der genaue Mechanismus dieser für die Pyrazolsynthese wichtigsten Reaktion bisher nur teilweise bekannt[91].

Selivanov et al. konnte zwar unter Anwendung von ^{1}H-NMR-Untersuchungen zeigen, daß nach einem schnellen Additionsschritt der Hydrazinkomponente an die 1,3-Dicarbonylverbindung die Zwischenverbindung sofort cyclisiert, bevor zwei sukzessive, unmittelbar aufeinander folgende Wasserabspaltungsreaktionen zum Pyrazolderivat führen[92]. Eine Hydrazon-Zwischenverbindung, die nach dem ersten Additionsschritt und einer anschließenden Wasserabspaltung entstehen würde, konnte nicht festgestellt werden. Der Reaktionsmechanismus[93] von Acetylaceton mit Hydrazin ist in Abbildung 6-3 dargestellt.

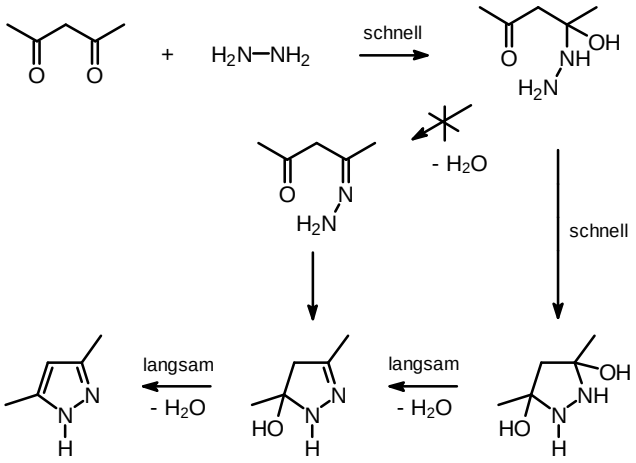


Abb. 6-3: Der Mechanismus zur Reaktion von Acetylaceton mit Hydrazin, der zur Bildung von 3,5-Dimethyl-1*H*-pyrazol führt.

Unklar ist aber zur Zeit noch der genaue Mechanismus dieser einleitenden Additionsreaktion. Theoretisch könnte er auf zwölf verschiedenen mechanistischen Reaktionswegen erfolgen, wenn man neben der unsymmetrischen 1,3-Dicarbonylverbindung auch noch die zwei tautomeren Enole berücksichtigt, die mit der unsymmetrischen Hydrazinkomponente reagieren können (siehe Abbildung 6-4).

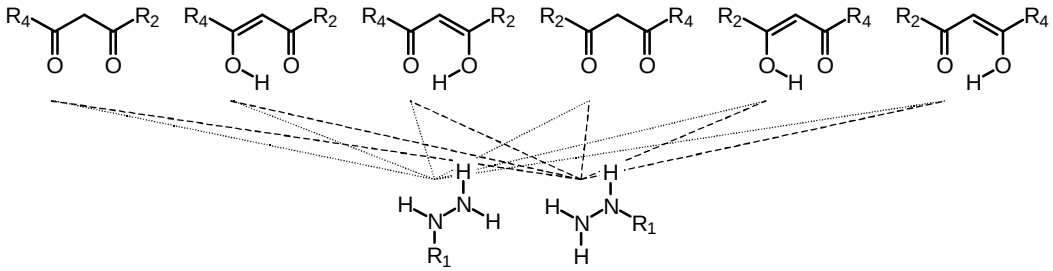


Abb. 6-4: Der erste Additionsschritt eines unsymmetrischen Hydrazinderivats an eine unsymmetrische 1,3-Dicarbonylverbindung kann auf zwölf verschiedenen Reaktionswegen eingeleitet werden.

Im weiteren Verlauf dieser Reaktion könnten wiederum tautomere Formen die Vielfalt an mechanistischen Reaktionspfaden erhöhen. Solange die tatsächlich reagierenden Species noch unbekannt sind, ist eine Abschätzung des Regioisomerieverhältnisses anhand des Reaktionsmechanismus nicht erfolgversprechend.

6.3.2 Nucleophilie und Elektrophilie

Es bedarf also eines anderen Verfahrens zur Abschätzung der Reaktivität der zu regioisomeren Pyrazolderivate führenden Reaktionen. Ein anderes Modell stützt sich auf die Betrachtung der Nucleophilie der Hydrazinkomponente bzw. der Elektrophilie der 1,3-Dicarbonylverbindung. Den Lehrbüchern der organischen Chemie zufolge greift das Hydrazinderivat

mit der nucleophileren Gruppe am elektrophileren Carbonylkohlenstoffatom an[94]. Folglich reagiert das weniger nucleophile Amin-Fragment mit der weniger elektrophilen Carbonylgruppe. Das Regioisomerenverhältnis wird also nach dieser stark vereinfachten Überlegung allein durch die Nucleophilie- bzw. Elektrophilie-Stärke bestimmt. Mittels einer Abschätzung der Nucleophilie bzw. Elektrophilie sollte man nach diesem Verfahren das bevorzugt gebildete Regioisomere vorhersagen können. Die Nucleophilie ist allerdings eine Größe, die man in der Regel kinetisch mißt. Man leitet sie durch den Vergleich verschiedener Reaktionsgeschwindigkeiten ab. Auf der Basis von linearen Freie-Enthalpie-Beziehungen kann man dann für Reagenzien unabhängige Nucleophilie- bzw. Elektrophilie-Werte angeben[95]. Eine Korrelation der Nucleophilie mit der thermodynamischen Basizität, die sich für jede Verbindung berechnen läßt, kann man nur für einige wenige Reagenzien angeben; in der Regel erhält man aber keine Korrelation dieser beiden Größen. Solange man keine umfassende Nucleophilie-Elektrophilie-Theorie ausgearbeitet hat, kann man die Nucleophilie- bzw. Elektrophilie-Stärke nur für diejenigen Reagenzien angeben, für die kinetische Daten vorliegen. Somit kann nach diesem Modell auch nur für diejenigen Produkte das Regioisomerieverhältnis abgeschätzt werden, für die kinetische Daten der entsprechenden Edukte vorliegen.

6.4 Vorhersage der Regioselektivität mittels Reaktionsklassifizierung

Das im Rahmen dieser Arbeit entwickelte Verfahren zur Abschätzung der Regioselektivität beruht auf dem induktiven Lernen, d.h. dem Lernen anhand von Beispielen. Die in Reaktionsdatenbanken gespeicherten Reaktionen zum Aufbau eines Pyrazols aus einer 1,3-Dicarbonylverbindung und einem Hydrazinderivat dienen dabei zum Aufbau einer Wissensbasis. Basierend auf dieser Information wird für eine Hydrazinverbindung und eine 1,3-Dicarbonylverbindung das bevorzugt gebildete Regioisomer vorhergesagt.

6.4.1 Bildung eines Trainingsdatensatzes

Als Trainingsdatensatz für ein Kohonen-Netz wurden Reaktionen aus der Beilstein CrossFire Datenbank ausgewählt. Diese bietet zur Zeit die umfassendste Sammlung an elektronisch erfaßten Reaktionen in der Organischen Chemie. Da auch diese Datenbank häufig nicht-stöchiometrische Reaktionsgleichungen enthält (siehe Kapitel 2.2.5), wurde in der Anfrage auf der Eduktseite die Hydrazinkomponente nicht vorgegeben (siehe Abbildung 6-5). Die Anfrage wird aber so formuliert, daß eine 1,3-Dicarbonylverbindung mit mindestens einem Wasserstoffatom in 2-Position als Substruktur auf der Eduktseite zu einem Pyrazolderivat reagiert, wobei zwei Kohlenstoff-Stickstoff-Bindungen aufgebaut werden. Somit kann auf die explizite Angabe der Hydrazinkomponente verzichtet werden. Als weiteres Aus-

wahlkriterium wird in der Anfrage die Reaktionsausbeute aufgenommen. Es sollen nur solche Reaktion als Treffer ausgegeben werden, die eine Ausbeute von größer 50% aufweisen. Somit wird sichergestellt, daß keine Reaktion in den experimentellen Datensatz aufgenommen wird, bei der das andere Regioisomere bevorzugt gebildet werden könnte.

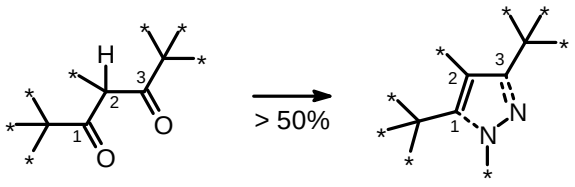


Abb. 6-5: Formulierte Anfrage an die Reaktionsdatenbank: Die gestrichelte Linie deutet an, daß die Bindungen geknüpft werden müssen; die Ziffern an den Elementsymbolen stehen für Atom-Atom-Mapping-Nummern; der Stern symbolisiert eine freie Valenz.

Die nach ca. 40 Sekunden abgeschlossene Suche in der CrossFire*plus*Reactions Datenbank ergab 755 Treffer. Davon haben 12 Reaktionen identische Produktensemble, so daß 6 Reaktionen von der Trefferliste ausgeschlossen werden. Bei drei Reaktionen sind beide Regioisomere in der Reaktionsgleichung angegeben. Die Edukte der verbleibenden 746 Reaktionen werden einer Symmetrieprüfung unterzogen. Nur diejenigen Edukte mit einer unsymmetrischen 1,3-Dicarbonylverbindung und einem monosubstituierten Hydrazin können zu regioisomeren Produkten reagieren. Diese Forderung erfüllen 313 Reaktionen, die den ersten Teil des Trainingsdatensatzes bilden. Dabei ging man davon aus, daß für jede dieser Reaktionen auch das angegebene Regioisomere bevorzugt gebildet wird. Eine Überprüfung, ob in der Originalveröffentlichung die Konstitution des Pyrazolderivats experimentell bestimmt wurde, oder nur theoretisch postuliert wurde, erfolgte angesichts der großen Treffermenge nicht.

Für den zweiten Teil des Trainingsdatensatzes wurde zu jeder der 313 Reaktionen diejenige Reaktion erzeugt, die zu dem entsprechenden regioisomeren Produkt führt. Somit besteht der Trainingsdatensatz aus insgesamt 626 Reaktionen, wobei die ersten 313 Reaktionen die experimentell erhaltenen Regioisomere wiedergeben, während die restlichen 313 Reaktionen zu den generierten, regioisomeren Produkten führen.

6.4.2 Codierung des Trainingsdatensatzes

Da man bei der Reaktionsvorhersage von den Edukten ausgeht und das Produkt oder die möglichen Produkte noch nicht kennt, verwendet man hier zur Codierung der Reaktionen nur physikochemische Deskriptoren auf der Eduktseite.

Zur Codierung der Reaktionszentren auf der Eduktseite der 626 Reaktionen wurde im Wesentlichen das Standardverfahren herangezogen (siehe Kapitel 3.3). Nur die Anzahl und Anordnung der zu codierenden Bindungen erfolgt hier nicht nach dem Standardverfahren, sondern nach einer anderen Methode, die ebenfalls mit dem CORA Programmsystem (siehe

Kapitel 8.3) realisiert werden kann. Im Reaktionszentrum einer Pyrazolsynthese sind auf der Seite der Edukte die beiden Bindungen zu den Carbonylgruppen sowie eine Einfachbindung enthalten, auf der Seite der Produkte die zu den Stickstoffatomen neu geknüpften Bindungen sowie eine Doppelbindung (siehe Abbildung 6-6). Somit würde man mit dem Standardverfahren, das lediglich die reagierenden Bindungen erfaßt, nur drei Bindungen auf der Seite der Edukte codieren können.

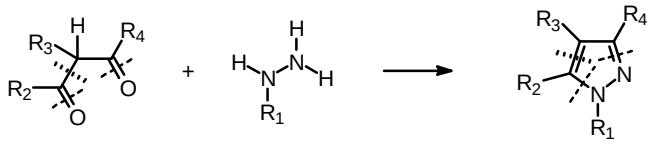


Abb. 6-6: Das Reaktionszentrum einer Pyrazolsynthese. Die unterschiedlich gestrichelten Bindungen markieren Bindungen die gebrochen oder geknüpft werden (----) sowie Bindungen bei denen die Bindungsordnung geändert wird (·····).

Das in Abbildung 6-6 dargestellte Reaktionszentrum gehört allerdings zu der Gesamtreaktionsgleichung einer Pyrazolsynthese. Diese Synthese setzt sich aus einer Reihe von Einzelreaktionen zusammen, bei denen mehr als die drei reagierenden Bindungen des Reaktionszentrums auf den Syntheseweg Einfluß nehmen können (siehe Kapitel 6.3.1). Daher sollten auch alle Einfluß nehmenden, physikochemischen Effekte der am Reaktionsgeschehen beteiligten Bindungen in den Codierungsvektor einer jeden Reaktion eingehen. Zur Ermittlung und eindeutigen Ausrichtung dieser vorgegebenen Bindungen in jeder der 626 Reaktionen wird eine Reaktionssubstruktursuche eingesetzt. Die formulierte Anfrage der Reaktionssubstruktursuche auf der Edukt- und Produktseite ist in Abbildung 6-7 wiedergegeben.

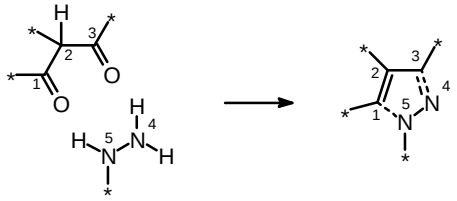


Abb. 6-7: Formulierte Anfrage der Reaktionssubstruktursuche zur Ausrichtung der Atome und Bindungen in den Edukten einer Pyrazol bildenden Reaktion. Die Atom-Atom-Mapping-Nummern sind mit Ziffern neben den Atomen dargestellt.

Durch die Angabe der Atom-Atom-Mapping-Nummern in der Anfrage werden die Atome des Pyrazolsystems auf der Produktseite eindeutig festgelegt. Nach Übertragung dieser Nummern auf die Eduktensemble ist auch die Anordnung der Atome der 1,3-Dicarbonylverbindungen sowie der Hydrazinverbindung eindeutig festgelegt. Für die beiden Eduktensemble, die zu regioisomeren Produkten führen, ist die sogenannte Atom-Atom-Mapping-Nummer an den beiden Carbonylgruppen der 1,3-Dicarbonylverbindung vertauscht und man kann zwischen beiden Eduktensemble differenzieren (siehe Abbildung 6-8). Aus der für Regioisomere

unterschiedlichen Anordnung der Atome des Reaktionszentrums resultiert letztlich auch eine unterschiedliche Anordnung der Bindungen, von denen physikochemische Deskriptoren berechnet werden.

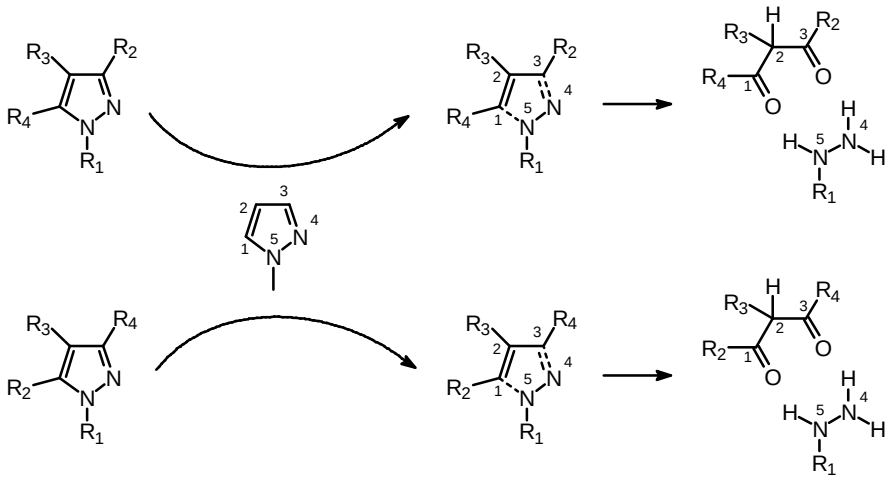


Abb. 6-8: Übertragung der Atom-Atom-Mapping-Nummern auf die beiden regioisomeren Produkte und Überführung dieser Nummern auf die Atome des Eduktensembles.

Nach der eindeutigen Ausrichtung der Bindungen kann man nach dem Standardverfahren mit der Berechnung der physikochemischen Deskriptoren fortfahren. Die Beschränkung auf elektronische Eigenschaften ohne Berücksichtigung sterischer Effekte ist bei diesem Reaktionstyp vertretbar, da sowohl der einleitende Additionsschritt als auch der Cyclisierungsschritt an einer sp^2-hybridisierten Carbonylgruppe stattfindet. Diese kann im Vergleich zu einem sp^3-hybridisierten Kohlenstoffatom nicht so stark abgeschirmt sein, so daß in erster Linie der Einfluß der elektronischen Eigenschaften die Verteilung der Regioisomere bestimmen wird. Da die meisten Reaktionen unter ähnlichen Reaktionsbedingungen durchgeführt werden, ist auch der Einfluß des Lösungsmittels, des Katalysators etc. auf nahezu alle Edukte gleich und bleibt bei der Codierung unberücksichtigt.

Als physikochemische Eigenschaften werden die sechs Effekte aus dem Standarddeskriptorensatz herangezogen (siehe Kapitel 3.3.1), die von insgesamt fünf Bindungen in den Edukten berechnet werden. Diese fünf Bindungen werden aufgrund reaktionsmechanistischer Überlegungen ausgewählt (siehe Abbildungen 6-3 und 6-4). In der Reihenfolge der eindeutigen Anordnung der Bindungen sind das die Bindungen zwischen den Atomen O ==> C1, C1 ==> C2, C2 ==> C3, C3 ==> O und H_2N ==> NH. Die Pfeile beschreiben die Richtung einer Bindung in der Form Startatom ==> Endatom.

Die Berechnung der PETRA-Deskriptoren ist für alle 626 Edukte erfolgreich, da es sich bei den Eduktmolekülen durchwegs um einfache Verbindungen handelt, die keine für die organische Chemie außergewöhnlichen chemischen Elemente enthalten. Der Trainingsdatensatz kann also in 626 Eingabevektoren für ein Kohonen-Netz transformiert werden.

Erwähnenswert ist hier, daß sich die Codierungsvektoren der Regioisomere nur in der Abfolge der Deskriptorelemente unterscheiden. Zwei Reaktionen, die zu regioisomeren Produkten führen (siehe Abbildungen 6-9 und 6-10) sollen dies verdeutlichen.

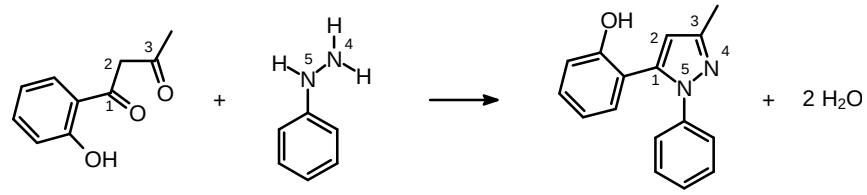


Abb. 6-9: Reaktion Nr. 3 des Trainingsdatensatzes (Beilstein #721302).

Die Codierung der Reaktion Nummer 3 führt zu einem 36-dimensionalen Vektor (siehe Tabelle 6-1), der für alle fünf ausgewählten Bindungen die entsprechenden sechs physikochemischen Effekte enthält. Da nach dem Standardverfahren maximal sechs Bindungen codiert werden, aber nur fünf Bindungen in den Edukten ausgewählt wurden, ist im Codierungsvektor ab Position 31 der Offsetwert von –5,0 eingetragen.

Bindung	b_o	$\Delta\chi_{AB,\sigma}$	$\Delta\chi_{AB,\pi}$	$\Delta q_{AB,tot}$	D^-_{AB}	D^+_{AB}
O ==> C1	0,000000	0,971632	0,237485	-0,489000	0,000000	0,000000
C1 ==> C2	-1,000000	0,493633	0,921042	0,094500	0,000000	0,093049
C2 ==> C3	-1,000000	-0,609533	-0,935813	-0,123600	0,153114	0,067570
C3 ==> O	0,000000	-0,882032	-0,103214	0,594900	0,479620	0,462720
H_2N ==> NH	-1,000000	-0,128733	-0,146914	-0,116800	0,000000	0,000000
	-5,000000	-5,000000	-5,000000	-5,000000	-5,000000	-5,000000

Tab. 6-1: Codierungsvektor für Reaktion Nr. 3 des Trainingsdatensatzes. Die Symbole der Deskriptoren sind in Kapitel 3.3.1 erklärt.

Die korrespondierende Reaktion Nummer 316 (siehe Abbildung 6-10) führt zu einem Codierungsvektor (siehe Tabelle 6-2), der dieselben absoluten Zahlenwerte wie der der Reaktion Nummer 3 aufweist, nur teilweise an unterschiedlicher Position.

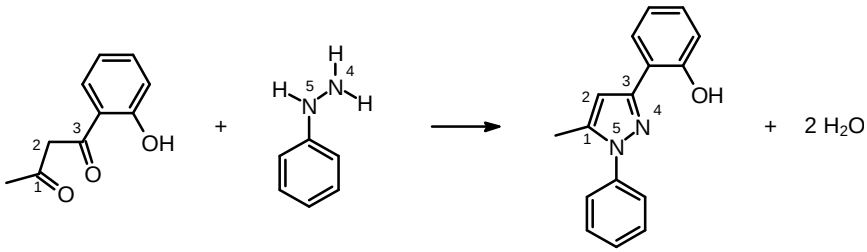


Abb. 6-10: Reaktion Nr. 316 des Trainingsdatensatzes (erzeugte Reaktion aus Beilstein #721302).

Bindung	b_o	$\Delta\chi_{AB,\sigma}$	$\Delta\chi_{AB,\pi}$	$\Delta q_{AB,tot}$	D^-_{AB}	D^+_{AB}
O ==> C1	0.000000	0.882032	0.103214	-0.594900	0.000000	0.000000
C1 ==> C2	-1.000000	0.609533	0.935813	0.123600	0.396600	0.393020
C2 ==> C3	-1.000000	-0.493633	-0.921042	-0.094500	0.143912	0.072171
C3 ==> O	0.000000	-0.971632	-0.237485	0.489000	0.000000	0.155660
H_2N ==> NH	-1.000000	-0.128733	-0.146914	-0.116800	0.000000	0.000000
	-5.000000	-5.000000	-5.000000	-5.000000	-5.000000	-5.000000

Tab. 6-2: Codierungsvektor für Reaktion Nr. 316 des Trainingsdatensatzes. Die Symbole der Deskriptoren sind in Kapitel 3.3.1 erklärt.

Anhand dieses Beispieles wird auch die Bedeutung einer einheitlichen Ausrichtung der Reaktionszentren (siehe Kapitel 3.2.3.2) deutlich.

6.4.3 Klassifizierung des Trainingsdatensatzes

Nachdem die 626 Reaktionen codiert sind, können sie einem Kohonen-Netz als Trainingsdatensatz zur Klassifizierung übergeben werden. Zuvor muß man noch die Netzparameter bestimmen. Die Festlegung der Dimension eines neuronalen Netzes ist von zentraler Bedeutung für die Vorhersagequalität. Wählt man die Dimension zu klein, so treten besonders häufig Konfliktneuronen auf und man kann für ein Testobjekt keine konkrete Klasse mehr angeben. Andererseits kann man die Netzgröße auch nicht beliebig groß wählen, da in diesem Fall viele leere Neuronen auftreten. Um eine optimale Netzwerkgröße auszuwählen, werden die Konfliktneuronen und leeren Neuronen in Abhängigkeit von der Netzwerkgröße ermittelt. Außerdem wird auch der Anteil derjenigen Neuronen an der Gesamtneuronenanzahl ermittelt, in denen Reaktionen eingetragen wurden, die zu experimentell erhaltenen Regioisomeren führen. Analog wird als vierte Größe der Anteil derjenigen Neuronen an der Gesamtneuronenanzahl ermittelt, in denen Reaktionen, die zu generierten Regioisomeren führen, projiziert wurden. Obwohl das Kohonen-Netz ohne Angabe einer Klassenzugehörigkeit trainiert wird (siehe Kapitel 2.5.2.3) werden hier für jede der 626 Reaktionen die entsprechende Klassenzugehörigkeit dem Eingabevektor angehängt, um die vier Größen schnell ermitteln zu können. Die ersten 313 Reaktionen, die zu experimentell bestätigten Regioisomeren führen, werden dabei der Klasse 1 zugeordnet, die restlichen Reaktionen, die zu den generierten Regioisomeren führen, der Klasse 2.

Der Kurvenverlauf dieser vier Größen ist in Abbildung 6-11 dargestellt. Man erkennt zum einen, daß sich der Anteil der leeren Neuronen asymptotisch einem Maximalwert annähert, während sich die Anzahl der Konfliktneuronen einem Minimalwert angleicht.

Die zwei Kurven der Klassenanteile zeigen eine Maximalwertfunktion. In ihrem Verlauf bilden sie mit der Kurve der Konfliktneuronen jeweils einen Schnittpunkt. Als optimale Netz-

werkgröße wurde diejenige Dimension ausgewählt, bei der als erstes beide Klassenanteile über dem Anteil der Konfliktneuronen liegen, nämlich bei einer Größe von 13 x 13.

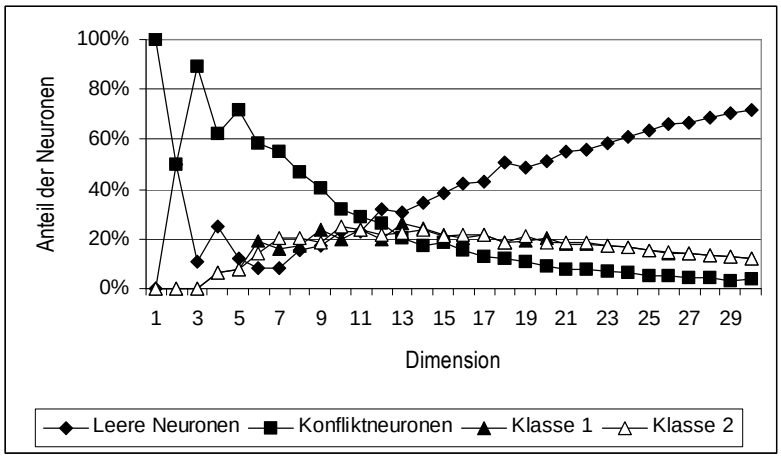


Abb. 6-11: Relativer Anteil der Anzahl an Konfliktneuronen, leeren Neuronen, Neuronen der Kasse 1 und Neuronen der Klasse 2 in Abhängigkeit von der Dimension des Netzes zur Ermittlung der optimalen Netzgröße. Die Dimension ist dabei die Anzahl der Neuronen in x- und y-Richtung, so daß quadratische Kohonen-Karten resultieren.

Die anderen Netzparameter zur Konfiguration des Netzes entnehme man Tabelle 6-3.

Netzparameter	Wert	Kommando
Netzwerkgröße	13 x 13	create 36 **13 13**
Netzwerktopologie	quadratisch-planar	set top_type r
Dimension der Neuronen	36	create **36** 13 13
Anzahl der maximal durchlaufenen Zyklen	40.000	train 40000
Lernrate zu Beginn h (t=0)	0,8	set par 4 **0.8**
Lernfaktor a	0,95	set dnc1 0.2 **0.95**
Spannweiten zu Beginn s_x (t=0), s_y (t=0)	4	set par **4** 0.8
Änderung der Spannweiten Ds_x, Ds_y	0,2	set dnc1 **0.2** 0.95
Zyklen konstanter Trainingsparameter t_s	400	set dnc **400** srdS

Tab. 6-3: Netz- und Trainingsparameter für die Klassifizierung der 626 Reaktionen zur Pyrazolsynthese. Die entsprechenden Kommandos sind für die KMAP Version 3.0 angegeben.

Das Trainieren der Kohonen-Karte ist nach 37.600 Zyklen beendet, weil die Lernrate und die Spannweite den als Standardwert eingestellten Grenzwert von je 0,1 unterschreiten.

Das Ergebnis der trainierten, quadratisch-planaren Kohonen-Karte mit der Dimension 13x13 ist in Abbildung 6-12 wiedergegeben.

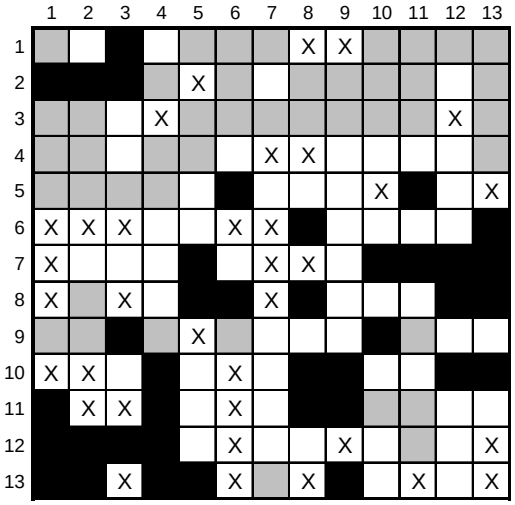


Abb. 6-12: Kohonen-Karte nach Trainieren mit 626 Reaktionen, die zu regioisomeren Produkten führen. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen werden mit einem X symbolisiert.

Wie man aufgrund der gleichen Reaktionsanzahl der zwei Klassen erwartet, belegen beide Klassen in der Kohonen-Karte etwa gleich viele Neuronen. Im oberen Teil der Karte sind vor allem die Reaktionen mit den experimentell erhaltenen Produkten eingetragen, während man im unteren Teil hauptsächlich Reaktionen mit den generierten Produkten antrifft. Beide Bereiche sind mit einem relativ hohen Anteil an Konfliktneuronen durchsetzt. Bei der ausgewählten Kartengröße von 13x13 liegen insgesamt 35 Konfliktsituationen vor.

Andererseits sind die beiden Bereiche auf der oberen und unteren Seite der Karte auch durch kleine Bereiche der jeweils anderen Klasse durchsetzt. So findet man beispielsweise im Gebiet der Reaktionen, die zu experimentell bestätigten Produkte führen, in den Neuronen (1,2), (2,2), (3,2) und (3,1) Reaktionen, die zu generierten Regioisomeren führen. Entsprechendes findet man auch in dem anderen Bereich der Karte, nämlich in den Neuronen (10,11), (11,11) und (11,12), in denen experimentell bestätigte Regioisomere eingetragen werden, obwohl in der Umgebung überwiegend Reaktionen zu finden sind, die zu generierten Regioisomeren führen.

Die Analyse der Schichten des trainierten Kohonen-Netzes gibt Aufschluß über den Einfluß der physikochemischen Effekte bei der Klassifizierung. Die 36 Eingabevektoren bedingen ein Kohonen-Netz mit 36 Schichten. Jeder Schicht ist damit auch ein physikochemischer Effekt einer Bindung zugeordnet. Der erste physikochemische Effekt wird durch Schicht 1 repräsentiert, der zweite Effekt durch Schicht 2 usw. Von diesen Schichten werden im folgenden einige näher diskutiert.

Großen Einfluß auf die Klassifizierung nimmt beispielsweise die Differenz der σ-Elektronegativitäten der Bindung zwischen den Atomen C1 und C2. Dieser Effekt wird mit der ach-

ten Schicht des Kohonen-Netzes erfaßt, die in Abbildung 6-13a dargestellt ist. Man erkennt eine Abstufung der Werte von der rechten oberen Seite zur linken und unteren Seite. Eine ähnliche Differenzierung zeigt auch die Gesamtladungsdifferenz zwischen denselben Atomen C1 und C2 (siehe Abbildung 6-13b). Während im oberen und rechten Teil des Netzes beide Eigenschaften hoch miteinander korreliert sind, so zeigen sich im mittleren Teil des Netzes einige Unterschiede. Dort wird durch die Gesamtladungsdifferenz ein weiterer Bereich mit höheren Werten herausgestellt, der durch die σ-Elektronegativität nicht abgetrennt wird. Die Gesamtladungsdifferenz zwischen den Atomen 1 und 2 spiegelt vor allem die Polarität der Bindung wieder.

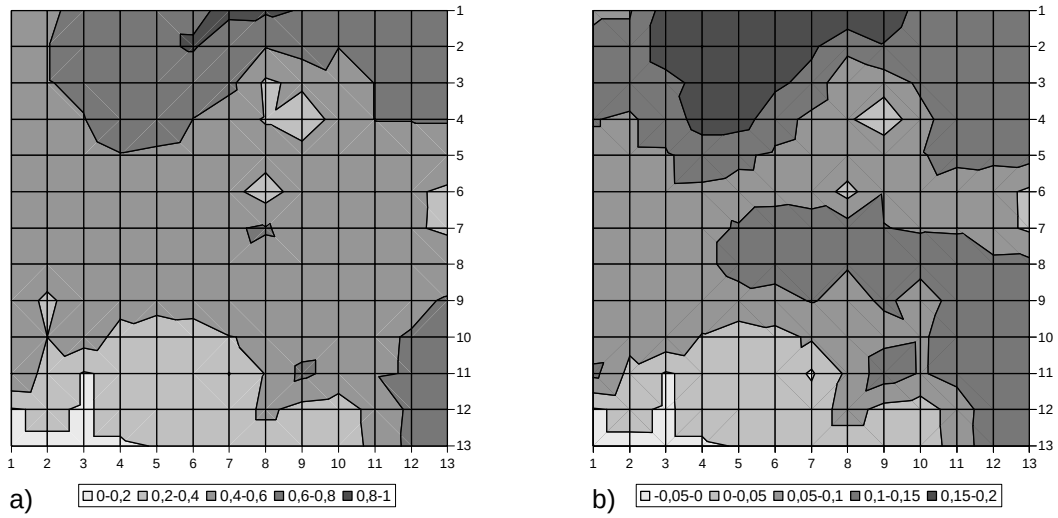


Abb. 6-13: a) Darstellung der Schicht 8 des Kohonen-Netzes: Einfluß der σ-Elektronegativitätsdifferenz zwischen den Atomen C1 und C2;
b) Darstellung der Schicht 10 des Kohonen-Netzes: Einfluß der Gesamtladungsdifferenz zwischen den Atomen C1 und C2.

Ein Maß für die Stabilisierung einer negativen Ladung an Atom C2 nach heterolytischem Bruch der Bindung zwischen C2 und C3 liefert die Delokalisationsstabilisierung (siehe Abbildung 6-14). Das dabei entstehende negativ geladene Fragment wird durch die benachbarte Carbonylgruppe stabilisiert und eventuell zusätzlich durch den Substituenten R_3.

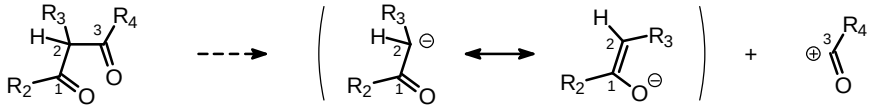


Abb. 6-14: Heterolytischer Bindungsbruch zwischen den Kohlenstoffatomen C2 und C3 zur Berechnung der Delokalisationsstabilisierung.

Die Delokalisationsstabilisierung einer negativen Ladung an C2 wird in der Schicht 17 repräsentiert. Diese Schicht ist in Abbildung 6-15a wiedergegeben.

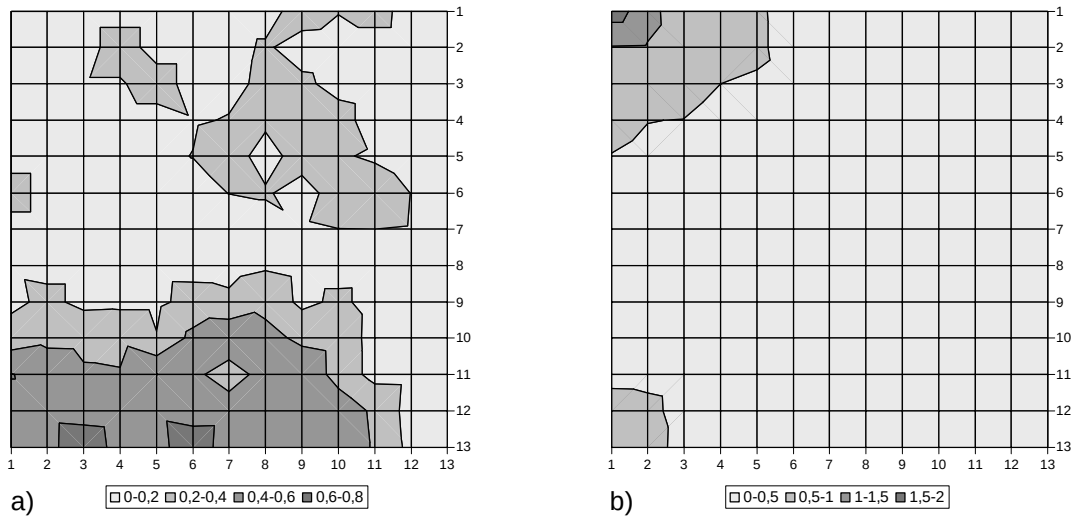


Abb. 6-15: a) Darstellung der Schicht 17 des Kohonen-Netzes: Einfluß der Delokalisationsstabilisierung einer negativen Ladung an C2 nach heterolytischem Bruch der Bindung C2 - C3;
b) Darstellung der Schicht 24 des Kohonen-Netzes: Einfluß der Delokalisationsstabilisierung einer positiven Ladung an C3 nach heterolytischem Bruch der Bindung C3 = O.

Diese Schicht zeigt eine Abstufung von der linken unteren zur rechten oberen Seite, wobei im linken unteren Teil des Netzes die höchste Stabilisierung auftritt. In diesem Bereich der Karte werden überwiegend die erzeugten Regioisomere eingetragen. Demnach scheint eine hohe Delokalisationsstabilisierung einer negativen Ladung an C2 das andere, experimentell nur schwer synthetisierbare Regioisomere zu begünstigen, obwohl im Verlauf der Reaktion aus der Einfachbindung zwischen den Atomen C2 und C3 eine delokalisierte, aromatische Heterocyclenbindung ausgebildet wird.

Die in Abbildung 6-15b dargestellte Schicht 24 bringt den Einfluß der Delokalisationsstabilisierung einer positiven Ladung an Atom C3 zum Ausdruck, nachdem ein heterolytischer Bruch der Bindung zwischen dem Kohlenstoffatom C3 und dem Sauerstoffatom erfolgte. Diese Eigenschaft differenziert hauptsächlich den oberen linken und ein wenig den unteren linken Bereich des Netzes vom restlichen Teil. In Neuron (1,1) wird Reaktion #169 eingetragen, die in Nachbarschaft zum C-Atom in der 1,3-Dicarbonylverbindung das größte delokalisierte System im Datensatz besitzt. Die Substituenten am C3-Atom in den Eduktmolekülen der in Neuron (1,13) eingetragenen Reaktionen tragen alle einen p-Fluorphenylrest, der die positive Ladung durch mesomere oder stark induktive Effekte effektiv stabilisieren kann. Eine hohe Delokalisationsstabilisierung an diesem Atom begünstigt somit die Bildung der experimentell herstellbaren Regioisomere.

Der Einfluß zweier physikochemischer Bindungseigenschaften im Hydrazinderivat ist in Abbildung 6-16 gezeigt. Auf der linken Seite ist Schicht 27 dargestellt, die die Differenz der π-Elektronegativität zwischen den Atomen N4 und N5 repräsentiert. Auf der rechten Seite ist

Schicht 28 des Netzes wiedergegeben, die die Gesamtladungsdifferenz zwischen den Atomen N4 und N5 widerspiegelt. Die Codierungsvektoren von Reaktionen, die zu regioisomeren Produkten führen, sind jedoch ab dem Element 25 identisch (siehe Tabellen 6-1 und 6-2), da ab dieser Position die sechs physikochemischen Effekte des Hydrazinderivats wiedergegeben werden. Daher können diese physikochemischen Deskriptoren im Hydrazinderivat nicht zur Abtrennung von Reaktionen, die zu experimentellen oder generierten Regioisomeren führen, herangezogen werden. Die Schichten dieser Effekte zeigen deshalb auch keinerlei Differenzierung zwischen den beiden Klassen 1 und 2 (siehe Abbildung 6-16a und 6-16b).

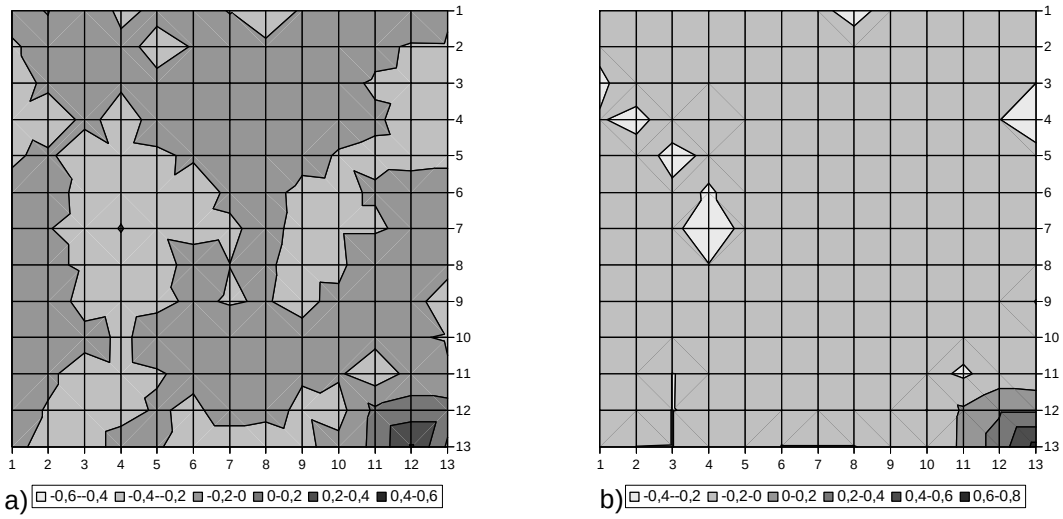


Abb. 6-16: a) Darstellung der Schicht 27 des Kohonen-Netzes: Einfluß der π-Elektronegativitätsdifferenz zwischen den Atomen N4 und N5;
b) Darstellung der Schicht 28 des Kohonen-Netzes: Einfluß der Gesamtladungsdifferenz zwischen den Atomen N4 und N5.

Keine der untersuchten Schichten zeigt alleine eine Zweiteilung, wie sie in Abbildung 6-12 dargestellt ist. Erst das Zusammenwirken aller physikochemischen Effekte – mit Ausnahme der Deskriptoren im Hydrazinderivat – führt zu einer Einteilung der Reaktionen, die zu experimentell bestätigten und generierten regioisomeren Produkten führen. Aus diesem Grund kann man sich auch nicht auf einige wenige physikochemische Effekte beschränken, die als Ausgangspunkt für ein mathematisches Modell zur Abschätzung des Regioisomerenverhältnisses dienen könnten.

In mehreren diskutierten Schichten wird ein kleiner Bereich im rechten unteren Eck des Netzes vom übrigen Teil abgetrennt. In diesem Bereich findet man Reaktionen, bei denen ein Säurehydrazid anstelle einer Hydrazinverbindung umgesetzt wird. Die physikochemischen Effekte der Hydrazingruppe werden durch die benachbarte Carbonylgruppe mit ihren besetzten und unbesetzten p-Orbitalen und dem daraus resultierenden mesomeren Effekt stark verändert. Dieser Effekt wirkt sich vor allem auf die Differenz der π-Elektronegativität zwischen den beiden Stickstoffatomen am stärksten aus (siehe Abbildung 6-16a). Andererseits wirkt sich auch der starke elektronenziehende Effekt der Carbonylgruppe, der über σ-Bindungs-

elektronen vermittelt wird, auf die Differenz der σ-Elektronegativität (siehe Abbildung 6-16b) aus. Sowohl der mesomere als auch der induktive Effekt führen zu einer deutlich erhöhten Differenz der Gesamtladung zwischen den beiden Stickstoffatomen N4 und N5.

Diese physikochemischen Effekte der Säurehydrazide führen zu einer deutlichen Abtrennung der Neuronen (13,12) und (13,13) vom übrigen Teil der Netzes. In diesem Bereich findet man nur Konfliktsituationen, da in diesen Neuronen sowohl die Reaktionen eingetragen werden, die zu den experimentell erhaltenen Produkten, als auch zu den generierten Regioisomeren führen. In Neuron (13,13) ist beispielsweise die experimentell bestätigte Reaktion #262 zusammen mit der generierten Reaktion #575 eingetragen. Im Gegensatz zu den Reaktionen, bei denen Hydrazine reagieren, kann das neuronale Netz mit den ausgewählten Deskriptoren keine zuverlässige Vorhersage zur Regioselektivität der Reaktionen treffen, bei denen Säurehydrazide mit 1,3-Dicarbonylverbindungen umgesetzt werden. Daher sollte man in den folgenden Testdatensätzen generell Reaktionen ausschließen, bei denen Säurehydrazide als Edukte eingesetzt werden.

Die Auftragung der Euklidischen Distanz für das mit 626 Reaktionen trainierte Netz zeigt nochmals deutlich diese Abtrennung der beiden Neuronen im unteren rechten Teil des Netzes (siehe Abbildung 6-17).

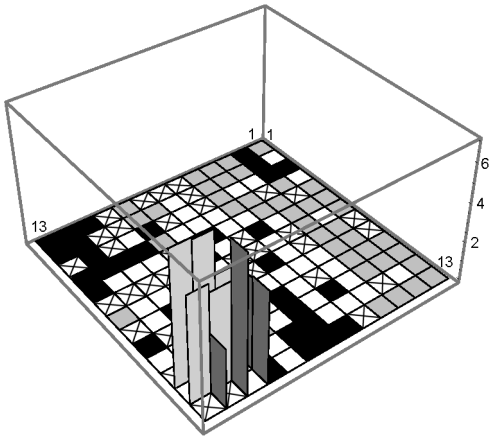


Abb. 6-17: Auftragung der Euklidischen Distanzen des mit 626 Reaktionen trainierten Kohonen-Netzes in die dritte Dimension. Es sind nur die Distanzen eingezeichnet, die einen Wert größer 1,5 aufweisen. Je höher die Barriere zwischen zwei Neuronen ist, desto unähnlicher sind sie zueinander.

Die hohen Euklidischen Distanzen trennen aber nicht nur diese beiden Neuronen (13,12) und (13,13) vom restlichen Teil des Netzes ab, sondern auch die Neuronen in der ersten Nachbarschaftssphäre, in denen im Training keine Reaktion eingetragen wurde. Diese sechs Neuronen (12,11) bis (13,13) werden daher der Umsetzung von Säurehydraziden mit 1,3-Dicarbonylverbindungen zugeschrieben. Da für solche Reaktionen nicht zwischen den entstehenden Regioisomeren unterschieden werden kann, werden diese sechs Neuronen im folgenden als Konfliktsituationen gekennzeichnet.

Um diese Karte zur Vorhersage des bevorzugt gebildeten Regioisomers einsetzen zu können, sollte man möglichst auch allen leeren Neuronen eine Klassenzugehörigkeit zuordnen. Da mit Ausnahme des rechten unteren Teils des Netzes die Euklidischen Distanzen niedrig sind, kann man für diesen Teil das in Kapitel 2.5.3.2 beschriebene Verfahren zur Einfärbung leerer Neuronen heranziehen, das die Euklidische Distanzen unberücksichtigt läßt. Dabei wird einem leeren Neuron die Klasse zugeteilt, die am häufigsten in den acht angrenzenden Neuronen auftritt. Während leere Neuronen in der Nachbarschaft ignoriert werden, zählen Konfliktsituationen genauso wie jede andere Gruppe. Um eine gegenseitige Beeinflussung benachbarter, leerer Neuronen auszuschließen, werden die Neuronen, für die man eine Klasse bestimmen konnte, erst am Ende eines Durchgangs eingefärbt. Dieser iterative Prozeß wird so lange durchlaufen, bis allen leeren Neuronen eine Klasse zugewiesen werden konnte oder bis keinem leeren Neuron mehr eine Klasse zugeteilt werden kann, da verschiedene Klassen zu gleichen Anteilen die Nachbarschaftssphäre ausfüllen.

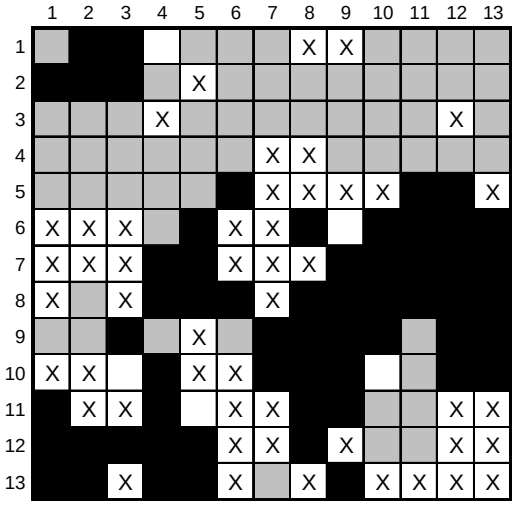


Abb. 6-18: Kohonen-Karte der Abbildung 6-12 nach Einfärbung aller möglicher leerer Neuronen. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen werden mit einem X symbolisiert.

Der Anteil der Konfliktneuronen ist durch das Einfärben der leeren Neuronen von 20,7% auf 27,8% gestiegen. Zur Reduzierung der Konfliktneuronenzahl wird daher folgendes Verfahren angewandt: Ausgehend von der Klasseneinteilung des trainierten Kohonen-Netzes der Abbildung 6-12 wird jedem Konfliktneuron die am häufigsten eingetragene Klasse 1 oder 2 zugeordnet, sofern nicht gleich viele Reaktionen der Klasse 1 und 2 vorliegen (siehe Abbildung 6-19a). Danach werden die leeren Neuronen einer Klasse zugeteilt unter Berücksichtigung der häufigsten Klasse in den benachbarten Neuronen (siehe Abbildung 6-19b).

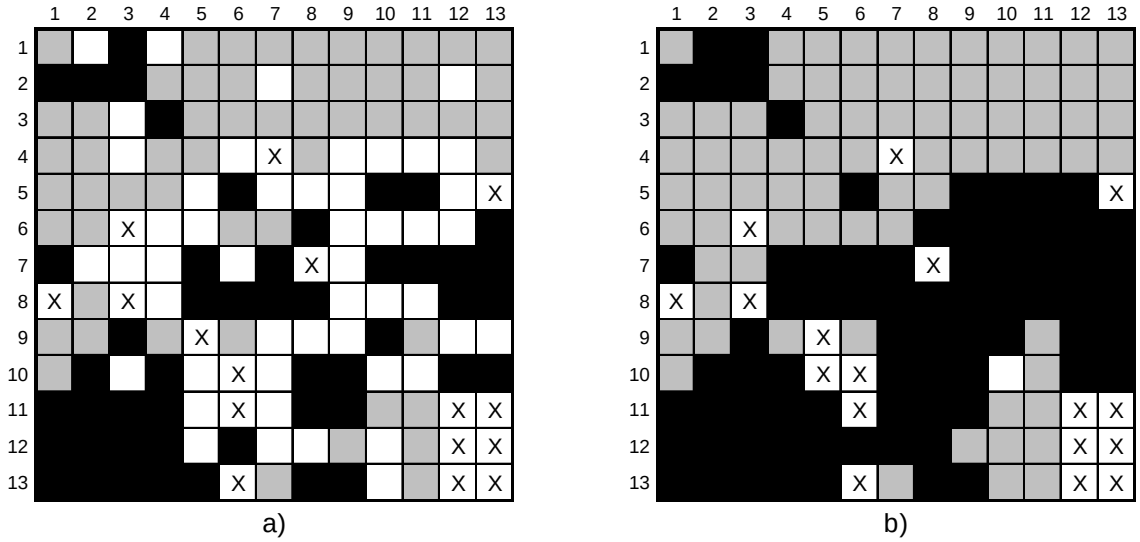
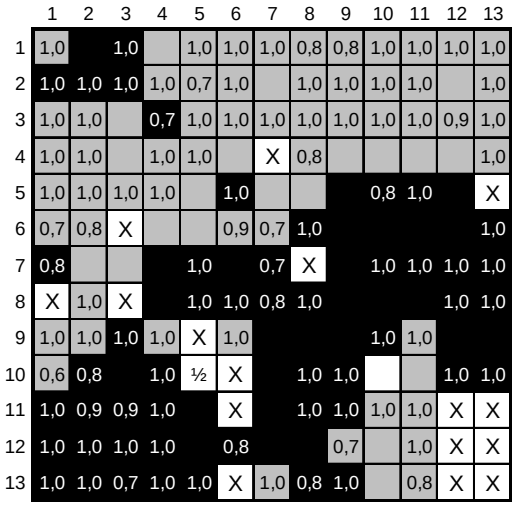


Abb. 6-19: a) Kohonen-Karte der Abbildung 6-12 nach Zuteilung der Konfliktneuronen zu den beiden Klassen. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktsituationen, die wegen gleich vieler Beispiele der beiden Klassen nicht zugeteilt werden können, werden weiterhin mit einem X symbolisiert;
b) Kohonen-Karte der Abbildung 6-19a nach Einfärbung aller möglicher leerer Neuronen.

Einen detaillierten Einblick in die Anteile einer Klasse pro Neuron ist in Abbildung 6-20 gezeigt. Neuronen mit dem Zahlenwert 1,0 enthalten ausschließlich Reaktionen einer einzigen Klasse, Neuronen mit einem Zahlenwert kleiner 1,0 enthalten in dem ursprünglichen Netz der Abbildung 6-12 noch Konfliktsituationen. Eingefärbte Neuronen ohne einen Zahlenwert waren in dem ursprünglichen Netz leer. Die mit einem X gekennzeichneten Neuronen waren in dem ursprünglichen Netz Konfliktneuronen, die gleich viele Reaktionen der Klasse 1 und Klasse 2 enthalten. Schließlich war das mit ½ gekennzeichnete Neuron in dem ursprünglichen Netz leer und ist nun ein Konfliktneuron, da in der Nachbarschaft am häufigsten Konfliktsituationen vorliegen.

In dieser Karte liegt der Anteil der Konfliktneuronen nur noch bei 8,9%, inklusive den sechs Konfliktsituationen im unteren rechten Teil des Netzes. Diese mit dem erweiterten Labeling-Verfahren eingefärbte Karte wird im folgenden bei der Auswertung der Testdatensätze herangezogen.



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1,0		1,0		1,0	1,0	1,0	0,8	0,8	1,0	1,0	1,0	1,0
2	1,0	1,0	1,0	1,0	0,7	1,0		1,0	1,0	1,0	1,0		1,0
3	1,0	1,0		0,7	1,0	1,0	1,0	1,0	1,0	1,0	1,0	0,9	1,0
4	1,0	1,0		1,0	1,0		X	0,8					1,0
5	1,0	1,0	1,0	1,0		1,0				0,8	1,0		X
6	0,7	0,8	X			0,9	0,7	1,0					1,0
7	0,8				1,0		0,7	X		1,0	1,0	1,0	1,0
8	X	1,0	X		1,0	1,0	0,8	1,0				1,0	1,0
9	1,0	1,0	1,0	1,0	X	1,0				1,0	1,0		
10	0,6	0,8		1,0	½	X		1,0	1,0			1,0	1,0
11	1,0	0,9	0,9	1,0		X		1,0	1,0	1,0	1,0	X	X
12	1,0	1,0	1,0	1,0		0,8			0,7		1,0	X	X
13	1,0	1,0	0,7	1,0	1,0	X	1,0	0,8	1,0		0,8	X	X

Abb. 6-20: Kohonen-Karte der Abbildung 6-12 nach Einfärbung aller möglicher leerer Neuronen. Konfliktneuronen wurden zuvor der häufigsten Klasse zugeteilt. Der Zahlenwert gibt den Anteil der häufigsten Klasse wieder. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen werden mit einem X oder ½ symbolisiert.

6.4.4 Validierung der Vorhersageleistung

Um die Vorhersagequalität des Netzes zu überprüfen wird eine Validierung durch Aufteilung des Datensatzes in einen Trainings- und Testdatensatz durchgeführt. Mit Hilfe dieser Methode kann man Rückschlüsse auf die Genauigkeit der Verallgemeinerung eines trainierten Netzes ziehen, d.h. je höher der Prozentsatz an richtig vorhergesagten Reaktionen ist, desto größere Fähigkeiten zur Verallgemeinerung hat das Netz während des Trainings erworben. Der Datensatz aus 626 Reaktionen wird hierfür in zwei gleich große Datensätze D1 und D2 aufgeteilt, wobei die Verteilung der Reaktionen auf die beiden Datensätze zufällig erfolgte. In beiden Datensätzen sollten in möglichst gleichen Anteilen Reaktionen enthalten sein, die zu den experimentell bestätigten Regioisomeren oder den generierten Regioisomeren führen. Diese Anforderung wurde in beiden Datensätzen gut verwirklicht, denn im Datensatz D1 sind 159 Reaktionen der Klasse 1 und 154 Reaktionen der Klasse 2 enthalten, im Datensatz D2 sind es folglich 154 Reaktionen der Klasse 1 und 159 Reaktionen der Klasse 2.

Anschließend wird der Datensatz D1 als Trainingsdatensatz einem neuronalen Netz zur Klassifizierung übergeben. Obwohl der Trainingsdatensatz nur aus 159 Reaktionen besteht, wurde dennoch die Dimension des Kohonen-Netzes nicht reduziert, da nach Projektion des Testdatensatzes insgesamt wieder 626 Reaktionen eingetragen sind.

Das Ergebnis der Klassifizierung des Datensatzes D1 ist in der Abbildung 6-21a dargestellt. Bevor man diesem Kohonen-Netz den Datensatz D2 zur Vorhersage übergibt, sollte man noch die leeren Neuronen einer Klasse zuordnen. Außerdem werden auch die Konfliktneuronen einer Klasse 1 oder 2 zugeteilt, um ihren hohen Anteil an der Kohonen-Karte herab-

zusetzen. Auch hier wird das in Kapitel 2.5.3.2 beschriebene Verfahren eingesetzt, nach dem die Häufigkeit der Klassen in der Nachbarschaft des leeren Neurons zur Klassenzuordnung herangezogen wird. Das Ergebnis der resultierenden Kohonen-Karte ist in Abbildung 6-21b wiedergegeben.

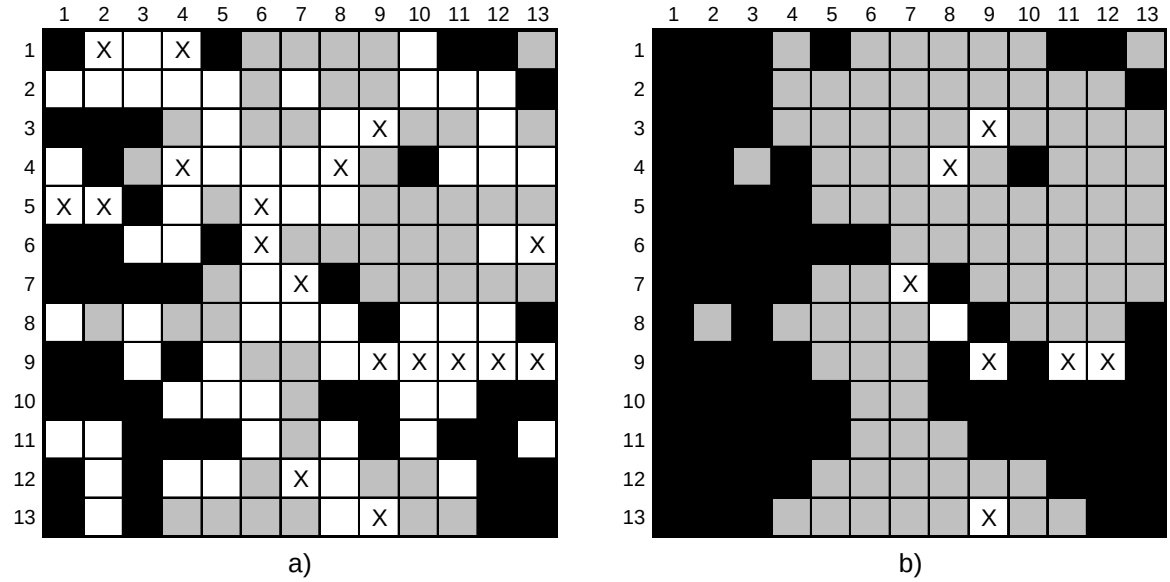


Abb. 6-21: a) Kohonen-Karte nach Trainieren mit 313 Reaktionen des Datensatzes D1. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen werden mit einem X symbolisiert;
b) Kohonen-Karte der Abbildung 6-21a nach Einfärbung aller möglicher leerer Neuronen. Konfliktneuronen wurden zuvor der häufigsten Klasse zugeteilt.

In diese Kohonen-Karte, die in der Abbildung 6-21b dargestellt ist, wird nun der Testdatensatz D2 projiziert. Das Ergebnis ist in Abbildung 6-22 dargestellt.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	3		~~11~~	1	1	7	2	4		1		
2						1		1	1				3
3	5	3		~~3~~		1	2		1	6	2		9
4		8		2					1	1			
5	~~5~~	10	1			~~9~~		2	2	2			~~3~~
6	5	2	4	1		~~5~~	1		4	3	6		~~3~~
7		1		1		~~2~~	~~3~~		~~3~~	4	6	6	6
8		2			~~4~~	1				1			1
9		3	2	2		1			2	~~5~~		1	~~4~~
10							4	2				3	1
11			3	1	1		2	2	4	3	3	2	
12	4		14				8		~~2~~	3		10	1
13	4		1	2	2		2		~~6~~	~~3~~	1	7	2

Abb. 6-22: Projektion der Reaktionen des Testdatensatzes D2 in die mit Datensatz D1 trainierte Kohonen-Karte der Abbildung 6-21b. Die Zahlen in den Neuronen stellen die Anzahl der eingetragenen Reaktionen dar; eine durchgestrichene Nummer zeigt eine Konfliktsituation an.

Für jede Reaktion des Testdatensatzes wird die Klasse mit der vorhergesagten Klasse verglichen. In 253 Fällen ist die Klasse 1 oder 2 der angefragten Reaktion identisch mit der vorhergesagten Klasse, d.h. die Bildungstendenz des Regioisomeren wird in 80,8% korrekt vorhergesagt (siehe Tabelle 6-4). In 47 Fällen (15,0%) sagt das neuronale Netz eine falsche Bildungstendenz voraus. In diesen Fällen wird entweder eine Reaktion der Klasse 1 in ein Gewinnerneuron eingetragen, in dem Reaktionen der Klasse 2 vorherrschend sind, oder eine Reaktion der Klasse 2 wird in ein Gewinnerneuron eingetragen, in dem hauptsächlich Reaktionen der Klasse 1 vorliegen. Schließlich werden 13 Reaktionen (4,2%) in Gewinnerneuronen eingetragen, die eine Konfliktsituation enthalten, d.h. für die Regioisomere dieser Reaktionen kann keine Bildungstendenz abgeschätzt werden.

Reaktionsanzahl	tatsächl. Klasse	vorhergesagte Klasse	Beurteilung
127 (40,6%)	Klasse 1	Klasse 1	richtig
126 (40,2%)	Klasse 2	Klasse 2	richtig Σ 253 (80,8%)
20 (6,4%)	Klasse 1	Klasse 2	falsch
27 (8,6%)	Klasse 2	Klasse 1	falsch
7 (2,2%)	Klasse 1	Konfliktneuron	---
6 (1,9%)	Klasse 2	Konfliktneuron	--- Σ 60 (19,2%)

Tab. 6-4: Erstes Teilergebnis zur Validierung des Datensatzes aus 626 Reaktionen.

Das gleiche Verfahren wird nun für D2 als Trainings- und D1 als Testdatensatz angewandt. Das Ergebnis der Klassifizierung des Datensatzes D2 ist in der Abbildung 6-23a dargestellt. Auch in diesem Fall werden die leeren Neuronen und die Konfliktneuronen – sofern möglich – der Klasse 1 oder 2 zugeteilt. Das Ergebnis der resultierenden Kohonen-Karte ist in der Abbildung 6-23b wiedergegeben.

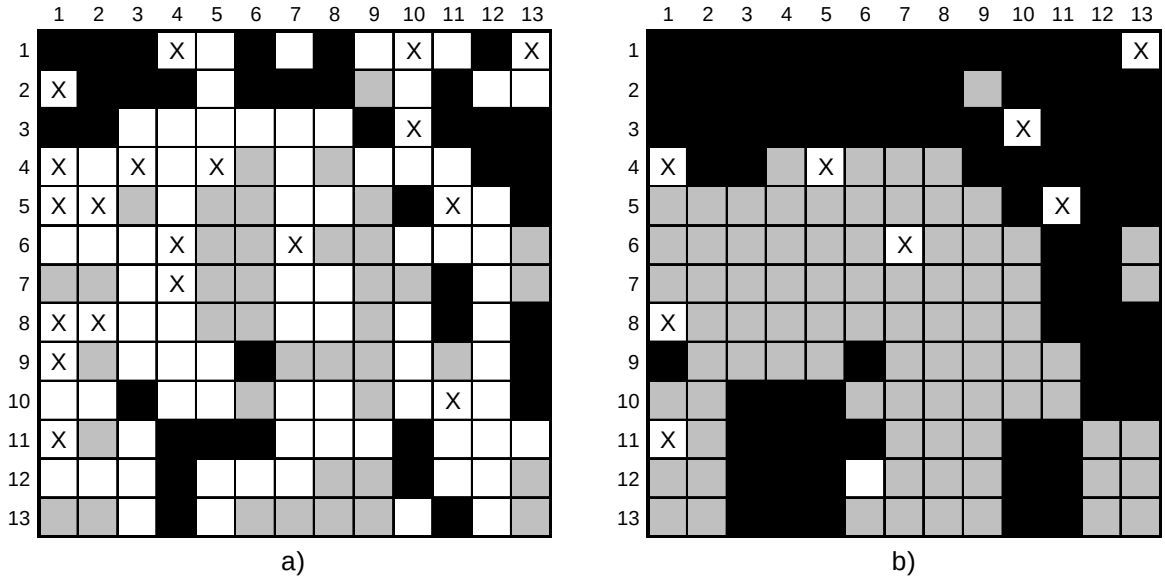


Abb. 6-23: a) Kohonen-Karte nach Trainieren mit 313 Reaktionen des Datensatzes D2;
b) Kohonen-Karte der Abbildung 6-23a nach Einfärbung aller möglicher leerer Neuronen.

In diese Kohonen-Karte, die auf der rechten Seite der Abbildung 6-23 wiedergegeben ist, wird nun der Testdatensatz D1 projiziert (siehe Abbildung 6-24).

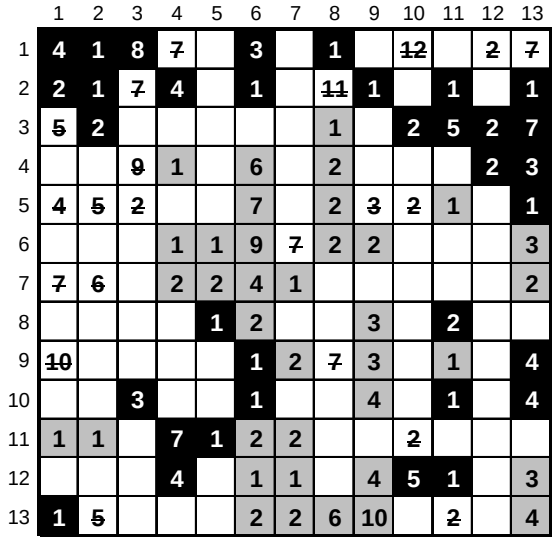


Abb. 6-24: Projektion der Reaktionen des Testdatensatzes D1 in die mit Datensatz D2 trainierte Kohonen-Karte der Abbildung 6-23b. Reaktionen des Testdatensatzes, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Die Zahlen in den Neuronen stellen die Anzahl der eingetragenen Reaktionen dar; eine durchgestrichene Nummer zeigt eine Konfliktsituation an.

In 246 Fällen (78,6%) wird die Bildungstendenz des Regioisomeren korrekt vorhergesagt. In 48 Fällen (15,3%) sagt das neuronale Netz eine falsche Bildungstendenz voraus. 18 Reaktionen (5,8%) werden in Konfliktneuronen eingetragen und eine Reaktion (0,3%) wird in ein leeres Neuron eingetragen (siehe Tabelle 6-5).

Reaktionsanzahl	tatsächl. Klasse	vorhergesagte Klasse	Beurteilung	
120 (38,3%)	Klasse 1	Klasse 1	richtig	
126 (40,3%)	Klasse 2	Klasse 2	richtig	Σ 246 (78,6%)
27 (8,6%)	Klasse 1	Klasse 2	falsch	
21 (6,7%)	Klasse 2	Klasse 1	falsch	
1 (0,3%)	Klasse 1	leeres Neuron	---	
11 (3,5%)	Klasse 1	Konfliktneuron	---	
7 (2,2%)	Klasse 2	Konfliktneuron	---	Σ 67 (21,4%)

Tab. 6-5: Zweites Teilergebnis zur Validierung des Datensatzes aus 626 Reaktionen.

Die zwei Untersuchungen zur Vorhersageleistung des Kohonen-Netzes zeigen, daß es die Bildungstendenz eines Regioisomeren zu rund 79,7% richtig vorhersagen kann. In rund 15,2% der untersuchten Fälle findet eine falsche Vorhersage statt und in rund 5,1% kann keine Entscheidung getroffen werden. Eine tabellarische Aufstellung der Ergebnisse zu dieser Validierung findet man im Anhang A.4.

Eine durchgeführte Kreuzvalidierung nach der „leave-one-out“-Methode liefert ein ähnliches Ergebnis wie die Aufteilung in Trainings- und Testdatensatz. Nach dieser Methode wird für 494 Reaktionen (78,9%) die korrekte Bildungstendenz vorhergesagt, während das neuronale Netz für 111 Reaktionen (17,7%) die falsche Bildungstendenz ermittelt und für 21 Reaktionen (3,4%) keine Tendenz vorhersagen kann.

Nach dieser Verifizierung der Vorhersageleistung des neuronalen Netzes erfolgt nun in Kapitel 6.4.5 eine detaillierte Betrachtung der Reaktionsbedingungen der klassifizierten Reaktionen dieses Trainingsdatensatzes sowie die Vorhersage des bevorzugt gebildeten Produktes für zwei ausgewählte Beispiele (siehe Kapitel 6.4.6). In den Kapiteln 7.3.2 und 7.3.4 wird nochmals auf das trainierte neuronale Netz der Abbildung 6-20 zurückgegriffen, um bevorzugt gebildete Regioisomere vorherzusagen. In diesem Kapitel werden dann tausende von Reaktionen untersucht, die eine kombinatorische Bibliothek aufbauen.

6.4.5 Darstellung der Reaktionsbedingungen und der Ausbeuten des Trainingsdatensatzes

Die Reaktionsklassifizierung kann auch eingesetzt werden, um geeignete Reaktionsbedingungen oder zu erwartende Ausbeuten für Anfragereaktionen vorherzusagen. Die mit jeder Reaktion gespeicherten Angaben über eingesetzte Lösungsmittel, Katalysatoren, Reaktionstemperaturen und Ausbeuten werden dazu herangezogen. Für jedes Neuron wird im Falle des Lösungsmittels und des Katalysators der häufigste Eintrag aller darin enthaltenen Reaktionen ermittelt, im Falle der Temperatur, der Reaktionszeit und der Ausbeute wird der Durchschnittswert berechnet. Da zu den generierten Reaktionen keine Bedingungen vorliegen, werden für die Neuronen, in denen hauptsächlich generierte Reaktionen eingetragen werden, keine Angaben gemacht. Die vier Reaktionsbedingungen (Lösungsmittel, Katalysator, Reaktionszeit und Reaktionstemperatur) sind für jedes Neuron der Klasse 1 des Trainingsdatensatzes aus 626 Reaktionen in Abbildung 6-25 dargestellt.

Abbildung 6-25 verdeutlicht die für den gesamten Datensatz sehr ähnlichen Reaktionsbedingungen. Meist werden Pyrazolsynthesen in Ethanol durchgeführt, wobei häufig Essigsäure als Katalysator eingesetzt wird. Daneben finden sich auch einige andere Lösungsmittel, wie Diethylether in Neuron (1,6) oder Dioxan in den Neuronen (4,2) und (5,3), die jedoch auf einzelne Neuronen beschränkt bleiben. Häufig werden die Ausgangsverbindungen unter erhöhter Reaktionstemperatur umgesetzt. Interessanterweise werden alle Reaktionen des Trainingsdatensatzes, die unter Basenkatalyse umgesetzt werden, in drei Neuronen im rechten oberen Eck der Kohonen-Karte eingetragen. Bei all diesen Reaktionen wird Natriumacetat in Essigsäure eingesetzt, um 4-(1*H*-Pyrazol-4-yl)-azo-benzolsulfonamid-Verbindungen zu bilden.

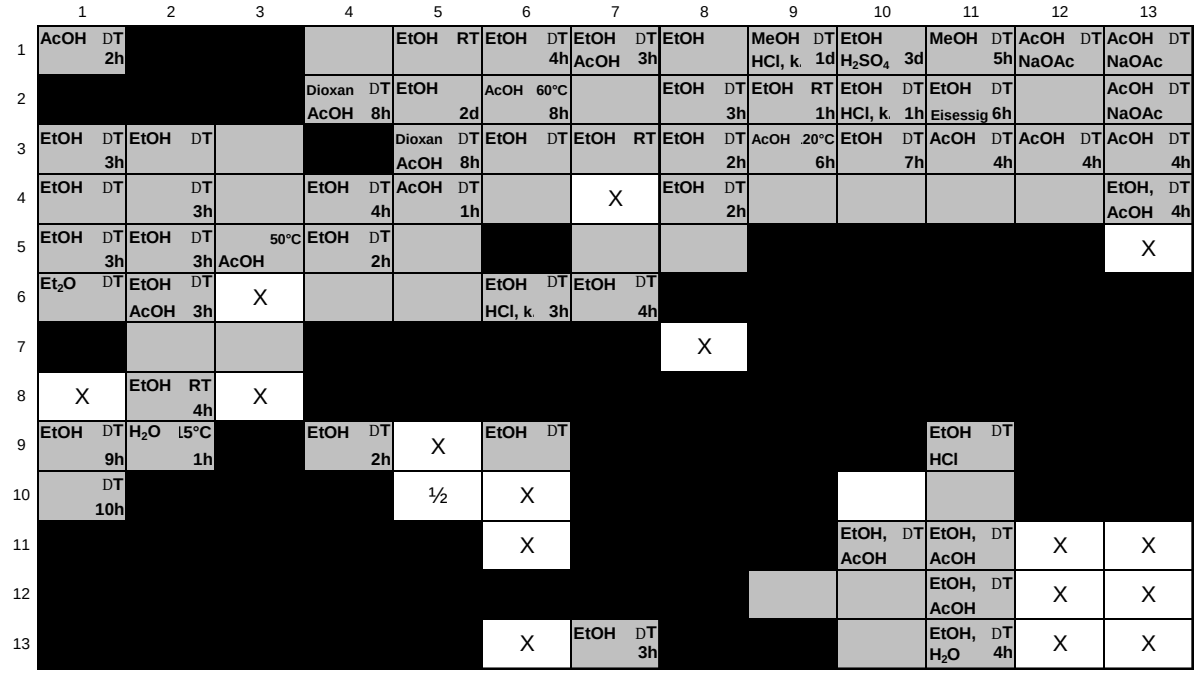
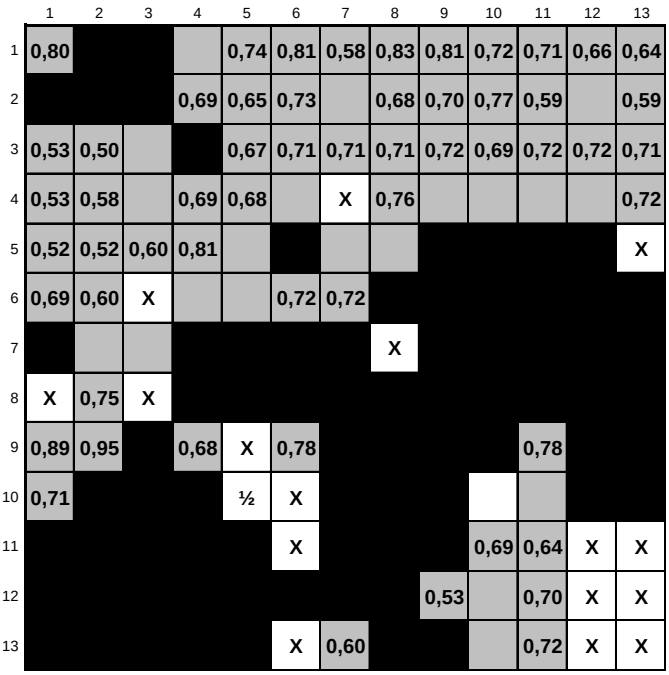


Abb. 6-25: Kohonen-Karte der Abbildung 6-20 mit den häufigsten Reaktionsbedingungen pro Neuron: Lösungsmittel (links oben), Katalysator (links unten), Reaktionstemperatur (rechts oben) und Reaktionszeit (rechts unten). ΔT steht für eine erhöhte Reaktionstemperatur. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen werden mit einem X oder ½ symbolisiert.

In Abbildung 6-26 ist die durchschnittliche Ausbeute aller in einem Neuron der Klasse 1 eingetragenen Reaktionen des Trainingsdatensatzes aus 626 Reaktionen dargestellt.

Die durchschnittlichen Ausbeuten der Reaktionen im oberen mittleren Bereich der Karte liegen zwischen 58% und 83%, wobei der Durchschnittswert bei rund 70% liegt. Auffallend niedrig sind im rechten oberen Eck der Karte die drei durchschnittlichen Ausbeuten, in denen die basenkatalysierten Pyrazolsynthesen eingetragen werden: Für die Neuronen (12,1), (13,1) und (13,2) werden Ausbeuten zwischen 59% und 66% ermittelt. Daraus läßt sich folgern, daß bei basenkatalysierten Pyrazolsynthesen kleinere Ausbeuten zu erwarten sind, als bei säurekatalysierter Umsetzung.

In den Neuronen (1,3) bis (2,5) findet man ebenfalls nur geringe Ausbeuten, hier liegen die Werte zwischen 50% und 58%. Bei allen in diesen Neuronen eingetragenen Reaktionen wurde auf einen Katalysator verzichtet. Dies könnte als eine der Ursache für die geringe Ausbeute angesehen werden.



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0,80				0,74	0,81	0,58	0,83	0,81	0,72	0,71	0,66	0,64
2				0,69	0,65	0,73		0,68	0,70	0,77	0,59		0,59
3	0,53	0,50			0,67	0,71	0,71	0,71	0,72	0,69	0,72	0,72	0,71
4	0,53	0,58		0,69	0,68		X	0,76					0,72
5	0,52	0,52	0,60	0,81									X
6	0,69	0,60	X			0,72	0,72						
7								X					
8	X	0,75	X										
9	0,89	0,95		0,68	X	0,78					0,78		
10	0,71				½	X							
11						X				0,69	0,64	X	X
12									0,53		0,70	X	X
13						X	0,60				0,72	X	X

Abb. 6-26: Kohonen-Karte der Abbildung 6-20 mit den durchschnittlichen Ausbeuten pro Neuron: Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen werden mit einem X oder ½ symbolisiert.

Obwohl die Reaktionen nicht nach den Reaktionsbedingungen oder der Ausbeute klassifiziert wurden, erkennt man in der Kohonen-Karte Bereiche, in denen Reaktionen ähnliche Ausbeuten oder Reaktionsbedingungen aufweisen. Auf beide Karten wird zurückgegriffen, wenn man für eine Pyrazolsyntheseanfrage geeignete Reaktionsbedingungen vorhersagen will, oder die zu erwartende Ausbeute abschätzen will, wie es im zweiten Reaktionsbeispiel des nächsten Kapitels erläutert wird.

6.4.6 Vorhersage des Hauptreaktionsproduktes zweier Pyrazolsynthesen

Die Vorhersageleistung des in Kapitel 6.4.3 trainierten neuronalen Netzes soll anhand zweier Beispiele zur Reaktionsvorhersage erläutert werden. Dabei sollen unsymmetrisch substituierte 1,3-Dicarbonylverbindungen mit einfach substituierten Hydrazinderivaten unter den für Pyrazolsynthesen üblichen Reaktionsbedingungen (EtOH, AcOH, 4h Erhitzen bei ca. 40 bis 60°C) umgesetzt werden. Das neuronale Netz sagt jeweils das regioisomere Hauptreaktionsprodukt vorher.

6.4.6.1 Hauptreaktionsprodukt von Beispiel I

Im ersten Beispiel reagieren 1-Methoxy-3-(4'-nitro-benzyl)-heptan-2,4-dion und Methylhydrazin miteinander, wobei zwei regioisomere Produkte möglich sind (siehe Abbildung 6-27).

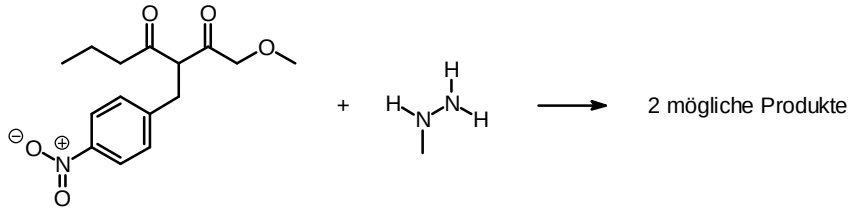


Abb. 6-27: Die Reaktion von 1-Methoxy-3-(4'-nitro-benzyl)-heptan-2,4-dion und Methylhydrazin kann zu 2 verschiedenen regioisomeren Produkten führen.

Ein Identitätsvergleich basierend auf den Hashcodes der 2 möglichen Reaktionsprodukte mit dem Datensatz der 626 Reaktionen aus der Beilstein Reaktionsdatenbank ergab keine Übereinstimmung. Somit muß auf einen Ähnlichkeitsvergleich, nämlich auf die Vorhersageleistung des trainierten neuronalen Netzes, zurückgegriffen werden. Dazu werden für alle zwei Reaktionsprodukte die entsprechenden Reaktionen erzeugt, anschließend nach dem in Kapitel 6.4.2 beschriebenen Verfahren codiert und als Testdatensatz dem trainierten neuronalen Netz übergeben (siehe Abbildung 6-20).

Das neuronale Netz ermittelt für die beiden Anfragereaktionen 2 Gewinnerneuronen. Die erste Reaktion wird in Neuron (2,9) eingetragen, in das während des Trainings eine einzige experimentelle Reaktion #198 eingetragen wurde, die mit einer Ausbeute von 95% zu dem Pyrazolderivat 5-Diethoxymethyl-1,3-dimethyl-1*H*-pyrazol führt (siehe Abbildung 6-28).

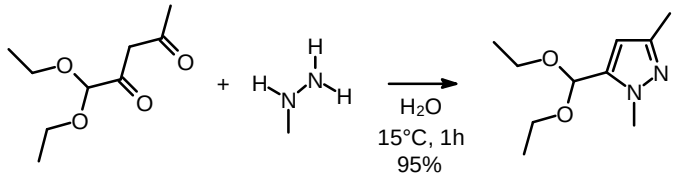


Abb. 6-28: Reaktion #198 in Neuron (2,9) aus der Beilstein Reaktionsdatenbank (#3517409).

Die zweite Reaktion wird in Neuron (5,8) projiziert, das der Klasse 2 angehört, da in dieses Neuron die Reaktion #511 eingetragen wurde. In dieser erzeugten Reaktion wird das zur Reaktion #198 regioisomere Produkt gebildet (siehe Abbildung 6-29).

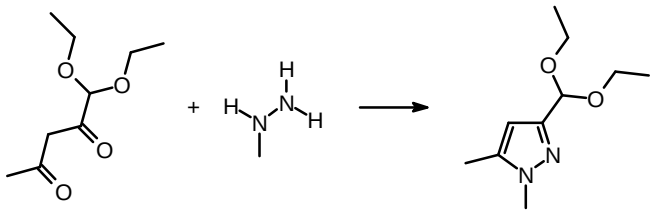


Abb. 6-29: Reaktion #511 in Neuron (5,8), die das zur Reaktion #198 regioisomere Produkt erzeugt.

Das neuronale Netz sagt somit eine eindeutige Bildungstendenz für das Regioisomere der ersten Reaktion vorher, während für die zweite Reaktion nur noch eine geringe Bildungstendenz zu erwarten ist (siehe Abbildung 6-30). Wie in Kapitel 6.4.3 beschrieben wurde, verläuft der Übergangsbereich der zwei Klassen 1 und 2 von der linken unteren Ecke der Kohonen-Karte bis in die Mitte auf der rechten Seite. Da beide Gewinnerneuronen relativ nahe an diesem Übergangsbereich liegen, sollte das Regioisomerenverhältnis etwas ausgeglichener ausfallen, als man aufgrund der 95%-igen Ausbeute der ähnlichsten Reaktion erwarten könnte.

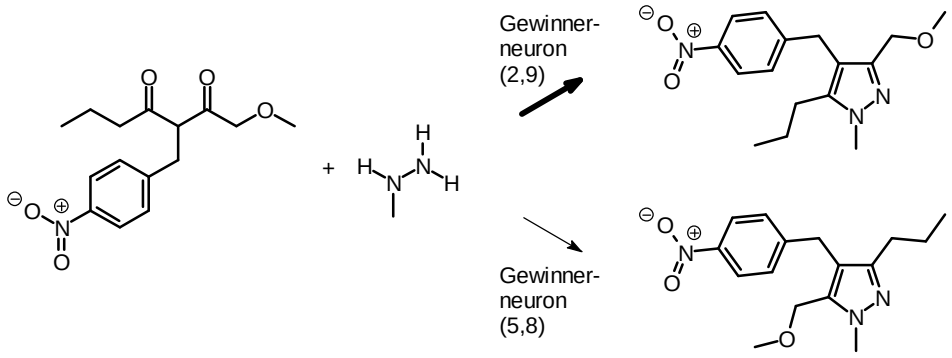


Abb. 6-30: Das trainierte neuronale Netz sagt für das erste Pyrazolderivat, 3-Methoxymethyl-1-methyl-4-(4'-nitro-benzyl)-5-propyl-1*H*-pyrazol, eine hohe Bildungstendenz voraus, für das zweite regioisomere Produkt, 5-Methoxymethyl-1-methyl-4-(4'-nitro-benzyl)-3-propyl-1*H*-pyrazol, nur eine geringe Tendenz.

Das vom neuronalen Netz vorhergesagte Regioisomere kann nun mit dem experimentell erhaltenen verglichen werden, da eine entsprechende Reaktion in der Beilstein Datenbank enthalten ist. Diese Reaktion zählt zu den drei Reaktionen, die nicht mit in den Trainingsdatensatz aufgenommen wurden, da jeweils zwei regioisomere Produkte in einer Reaktionsgleichung angegeben sind (siehe Kapitel 6.4.1). Für diese Reaktion sind – im Gegensatz zu den beiden anderen Reaktionen – die Ausbeuten aller möglicher Regioisomere bestimmt worden. Im Experiment, das von Nicolai et al. durchgeführt wurde, wurde dasselbe Regioisomere mit 67% Ausbeute isoliert, das auch dem neuronalen Netz zufolge entstehen sollte[96]. Das andere regioisomere Produkt bildet sich mit 32% Ausbeute (siehe Abbildung 6-31).

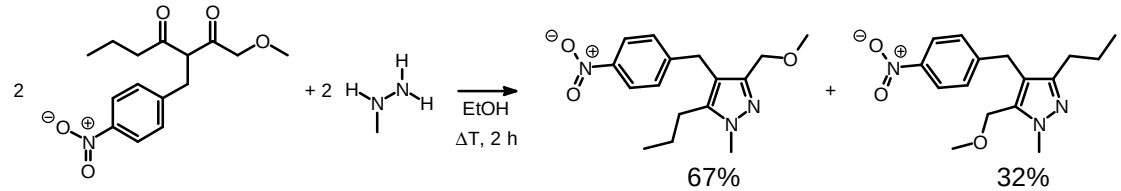


Abb. 6-31: Reaktion #3993875 aus der Beilstein Reaktionsdatenbank.

Das neuronale Netz sagt also in diesem ersten Beispiel das richtige Regioisomere, nämlich 3-Methoxymethyl-1-methyl-4-(4'-nitro-benzyl)-5-propyl-1*H*-pyrazol, voraus. Für das andere Regioisomere kann man nur aufgrund der Lage des Gewinnerneurons eine sehr geringe Bildungstendenz erwarten.

6.4.6.2 Hauptreaktionsprodukt von Beispiel II

Im zweiten Beispiel werden die in Abbildung 6-32 gezeigten Moleküle, 3-Benzoyl-5-methyl-hexan-2,4-dion und Methylhydrazin, unter den für Pyrazolsynthesen üblichen Reaktionsbedingungen umgesetzt. Bei dieser Umsetzung sind insgesamt sechs Reaktionsprodukte möglich.

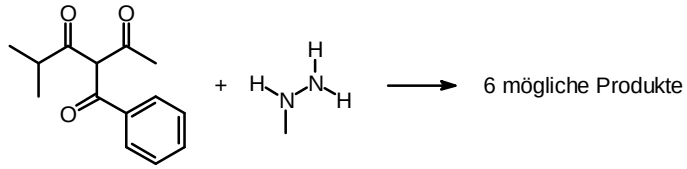


Abb. 6-32: Die Reaktion von 3-Benzoyl-5-methyl-hexan-2,4-dion und Methylhydrazin kann zu 6 verschiedenen Produkten führen.

Eine mit dem 3D-Generator CORINA erzeugte Konformation der Tricarbonylverbindung zeigt, daß alle 3 Carbonylgruppen nach oben gerichtet mit nahezu identischen Trajektorien für einen Angriff von Methylhydrazin zugänglich sind. Somit sollte die Bildung des Hauptprodukts in erster Linie von den physikochemischen Effekten beeinflußt werden. Mit Hilfe des neuronalen Kohonen-Netzes in Abbildung 6-20, das mit 626 Reaktionen trainiert wurde, soll auch für dieses Beispiel das Hauptprodukt ermittelt werden.

Auch in diesem Fall muß auf die Vorhersageleistung des trainierten neuronalen Netzes zurückgegriffen werden, da ein Identitätsvergleich der 6 möglichen Reaktionsprodukte mit dem Datensatz der 626 Reaktionen keine Übereinstimmung ergab. Es werden deshalb für alle sechs möglichen Reaktionsprodukte die entsprechenden Reaktionen erzeugt, anschließend nach dem in Kapitel 6.4.2 beschriebenen Verfahren codiert und als Testdatensatz dem trainierten neuronalen Netz übergeben (siehe Abbildung 6-20).

Das neuronale Netz bestimmt für die sechs Anfragereaktionen 2 Gewinnerneuronen. In Neuron (3,8) werden fünf Anfragereaktionen eingetragen. Bei diesem Neuron handelt es sich allerdings um ein Konfliktneuron, so daß für diese Reaktionen keine zuverlässige Vorhersage möglich ist. In das andere Gewinnerneuron (9,2) wird dagegen eine einzige Reaktion der sechs Anfragereaktionen eingetragen. In dieses Neuron wurden während des Trainings ausschließlich experimentell bestätigte Reaktionen aus der Beilstein Datenbank eingetragen. Somit wird für diese eine Reaktion, die in Abbildung 6-33 gezeigt ist, die höchste Bildungstendenz vorhergesagt.

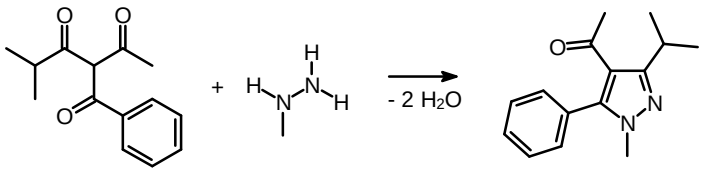


Abb. 6-33: Das vom trainierten neuronalen Netz ermittelte, bevorzugt entstehende Reaktionsprodukt des zweiten Beispiels.

Methylhydrazin sollte demnach bevorzugt mit den beiden Carbonylgruppen reagieren, die zur Phenyl- und *i*-Propylgruppe benachbart sind. Dabei wird bevorzugt das 4-Acetyl-3-isopropyl-1-methyl-5-phenyl-1*H*-pyrazol-Regioisomere gebildet.

Für diese Anfragereaktion kann man schließlich noch die Reaktionsbedingungen vorhersagen lassen. In Neuron (9,2) der Kohonen-Karte in Abbildung 6-25 wird Ethanol als häufigstes Lösungsmittel ermittelt, die meisten Reaktionen wurden bei Raumtemperatur durchgeführt, und eine Stunde zur Reaktion gebracht. Die durchschnittliche Ausbeute der Reaktionen in diesem Neuron liegt bei 70 (±11)%.

Eine Reaktionssubstruktursuche unter Beachtung der aromatischen und aliphatischen Reste der Anfragereaktion ergab in der Beilstein Reaktionsdatenbank keinen Treffer, d.h. diese Reaktion wurde bisher experimentell nicht untersucht. Daher ist für dieses Beispiel – im Gegensatz zum ersten Beispiel – keine Überprüfung anhand eines experimentellen Ergebnisses möglich.

6.5 Diskussion des Einsatzes der Reaktionsklassifizierung in der Reaktionsvorhersage

Wie in dem Kapitel zur Validierung der Vorhersageleistung 6.4.4 gezeigt wurde, sagt das neuronale Kohonen-Netz für eine Pyrazolsynthese nach dem Knorr-Verfahren das korrekte Regioisomere in rund 79,7% richtig voraus. Die Ursache für diese akzeptable, aber nicht besonders überzeugende Vorhersageleistung ist hauptsächlich in dem Trainingsdatensatz zu suchen, in zweiter Linie in der Auswahl der physikochemischen Effekte.

Weitere Untersuchungen zur Variation des Trainingsdatensatzes haben nämlich gezeigt, daß die Auswahl der Reaktionen für die Vorhersagequalität des Netzes von größter Wichtigkeit ist. Setzt man alle Reaktionen mit derselben Suchanfrage (siehe Abbildung 6-5) aber beliebiger Ausbeute als Trainingsdatensatz ein, so werden in der Kreuzvalidierung nur noch rund 70% korrekt vorhergesagt. In diesem Falle besteht der Datensatz aus insgesamt 1.490 experimentellen und erzeugten Reaktionen, d.h. 432 Reaktionen aus der Beilstein Reaktionsdatenbank weisen in diesem Datensatz eine Ausbeute von weniger als 50% auf. Bei diesen Reaktionen könnte theoretisch mit mindestens 51% das andere Regioisomere entstehen, das aber nicht unbedingt in der Datenbank abgespeichert sein muß. In diesem Fall müßte man das abgespeicherte Regioisomere dem generierten Teil des Datensatzes hinzuzählen, während das Regioisomere mit mindestens 51% Ausbeute dem experimentellen Datensatz angehörte. Diese Untersuchung zeigt, daß man bei dem Aufbau eines Trainingsdatensatzes sowohl zwischen einer quantitativen und qualitativen Auswahl abwägen muß. Viele Reaktionsbeispiele sind notwendig, damit das neuronale Netz möglichst gut generalisieren kann, andererseits können Reaktionen mit einer niedrigen Ausbeute die Vorhersageleistung wesentlich verringern.

Manchmal findet man in der Originalliteratur zu den Pyrazolsynthesen auch den Hinweis, daß das Regioisomere nicht experimentell bestimmt wurde, sondern theoretisch mit Abschätzungen der Nucleophilie und Elektrophilie abgeleitet wurde[97]. Daher können auch solche Reaktionen, bei denen das Regioisomere nicht experimentell verifiziert wurde, die Vorhersageleistung beeinflussen. Eine auf die Originalliteratur zurückgehende Überprüfung der Reaktionsbeispiele würde allerdings sehr zeitintensiv ausfallen, wenn man wie in dieser Arbeit große Datensätze untersuchen will.

Des weiteren könnte die Klassifizierung sicherlich noch im Hinblick auf eine bessere Vorhersageleistung optimiert werden, indem man die physikochemischen Effekte des Standardverfahrens verändert oder weitere hinzu nimmt. Die Stabilisierung einer negativen Ladung am Atom 2 nach Bruch der Bindung zwischen den Atomen 1 und 2 (D^{-}_{21}) und die Stabilisierung einer negativen Ladung am Atom 2 nach Bruch der Bindung zwischen den Atomen 2 und 3 (D^{-}_{23}) spielen angesichts des Reaktionsverlaufs über polare Zwischenstufen eine wichtige Rolle. Außerdem blieben bisher auch physikochemische Effekte zwischen den Wasserstoffatomen und Stickstoffatomen in den Hydrazinderivaten unberücksichtigt, die sicherlich auch einen nicht zu unterschätzenden Einfluß auf den Reaktionsverlauf nehmen können.

6.6 Anschluß an das Reaktionsvorhersagesystem EROS

Für das Reaktionsvorhersagesystem EROS7 wurde bereits eine Regel zur Pyrazolsynthese, ausgehend von 1,3-Dicarbonylverbindungen und Hydrazinen, geschrieben. Allerdings generiert EROS7 bisher alle theoretisch möglichen Produkte, einschließlich aller Regioisomere. Im Experiment werden aber viele Regioisomere gar nicht gebildet, da häufig nur eins der beiden Regioisomere bevorzugt entsteht. Um dieser unterschiedlichen Bildungstendenz der Regioisomere Rechnung zu tragen, und eine kleinere Bibliothek zu bilden, die dem Experiment näher kommt, verknüpfte man das Ergebnis der Reaktionsklassifizierung mit dem Reaktionsvorhersagesystem. Die Reaktionsgeschwindigkeiten für alle Reaktionen des Netzwerkes müssen dazu weder bekannt sein noch berechnet werden.

Das Reaktionsvorhersagesystem EROS ruft für jede erzeugte Reaktion ein trainiertes Kohonen-Netz auf, das mit entsprechenden, experimentell durchgeführten Reaktionsbeispielen trainiert wurde. Falls die zur Anfragereaktion ähnlichste Reaktion aus einem experimentellen Versuch stammt, so wird auch für die Anfragereaktion eine große Bildungstendenz vorhergesagt. Wenn die ähnlichste Reaktion zu dem generierten Regioisomeren führt, so ist die Reaktionswahrscheinlichkeit für diese Reaktion nur gering. Auf diese Weise gelangen nur die Regioisomere in die Bibliothek, denen eine hohe Bildungstendenz vorhergesagt wird.

Diese verbesserte Reaktionssimulation kommt der praktischen Versuchsdurchführung einen Schritt näher.

7 Praktische Anwendung: Planung kombinatorischer Bibliotheken

Die rasante Entwicklung der kombinatorischen Chemie in den letzten Jahren wurde vor allem von der pharmazeutischen Industrie vorangetrieben[98]. Im Gegensatz zur konventionellen organischen Synthese, bei der beispielsweise zwei Edukte A und B zu einem Produkt AB umgesetzt werden, setzt man bei der kombinatorischen Synthese strukturchemisch unterschiedliche Bausteine vom Typ A mit Bausteinen vom Typ B um[99]. Nach kombinatorischen Prinzipien entstehen gleichzeitig beispielsweise aus 10 Verbindungen vom Typ A ($A_1 - A_{10}$) und 10 Verbindungen vom Typ B ($B_1 - B_{10}$) 100 Verbindungen. Diese Synthesetechnik liefert eine Vielzahl von Testsubstanzen, die anschließend einem Massenscreening (High-Throughput-Screening) zum raschen und effizienten Auffinden neuer Leitstrukturen unterzogen werden oder aus denen eine optimale Leitstruktur gewonnen werden kann.

Die kombinatorische Synthese wurde erstmals zum Aufbau von Peptidbibliotheken basierend auf dem Verfahren der Festphasenpeptidsynthese nach Merrifield eingesetzt[100]. In der Zwischenzeit wurden zahlreiche Methoden und Verfahren zum Aufbau von Bibliotheken niedermolekularer organischer Verbindungen entwickelt[101]. Da biologisch aktiven Verbindungen meist eine heterocyclische Molekülstruktur zu Grunde liegt, stehen oft Heterocyclensynthesen im Mittelpunkt der kombinatorischen Synthese. Als ein Heterocyclenvertreter mit rigidem, hochfunktionalisierbarem Molekülgerüst wurde für die folgenden Untersuchungen das Pyrazolsystem ausgewählt.

Nach einleitenden Abschnitten zu den chemischen Eigenschaften der Pyrazole und deren Einsatzgebiete wird der Aufbau zweier kombinatorischer Bibliotheken und deren Bewertung basierend auf der Reaktionsklassifizierung erörtert.

7.1 Eigenschaften und Bedeutung der Pyrazole

7.1.1 Chemische Eigenschaften

Pyrazol (genauer 1*H*-Pyrazol) gehört zur Heterocyclenfamilie der 1,2-Azole und zeigt aufgrund der Delokalisierung des freien Elektronenpaars am 2N-Atom aromatischen Charakter mit hoher π-Elektronendichte. Das freie Elektronenpaar des anderen Stickstoffatoms ist orthogonal hierzu angeordnet und steht für Protonierungsreaktionen zur Verfügung, weshalb Pyrazole auch basisches Verhalten zeigen. Pyrazol beispielsweise hat einen pK_B-Wert von 11,5. Bei unsymmetrisch C-substituierten Pyrazolen tritt ein Tautomeriegleichgewicht auf, so daß beide N-Atome chemisch ununterscheidbar werden. 5-Methyl-1*H*-pyrazol und 3-Methyl-

1*H*-pyrazol existiert beispielsweise in Lösung als rasch äquilibrierendes Gemisch (siehe Abbildung 7-1).

Abb. 7-1: Tautomeriegleichgewicht zwischen 5-Methyl-1*H*-pyrazol und 3-Methyl-1*H*-pyrazol.

7.1.2 Bedeutung der Pyrazole

Pyrazole sind in der Natur nur selten anzutreffen, zum Beispiel im Samen der Wassermelone (Citrullus vulgaris) als Bestandteil der Aminosäure β-(Pyrazol-1-yl)-L-alanin[102]. Pyrazolderivate werden großtechnisch neben Stilben-Derivaten als optische Aufheller (Blankophor) für Textilien, Papier, Waschmittel und Kunststoffe eingesetzt. Einige Pyrazolderivate finden als Herbizide, Akarizide und Insektizide Anwendung. Die Herbizide Azimsulfuron und Halosulfuron setzt man gegen Unkräuter in Reis- bzw. Maiskulturen ein, Fenpyroximat und Tebufenpyrad zeigen Kontakt- und Fraßgiftwirkung gegen Milben im Obst-, Gemüse-, Wein-, und Teeanbau. Das Pestizid Fipronil schließlich wirkt gegen ein breites Insektenspektrum (siehe Abbildung 7-2).

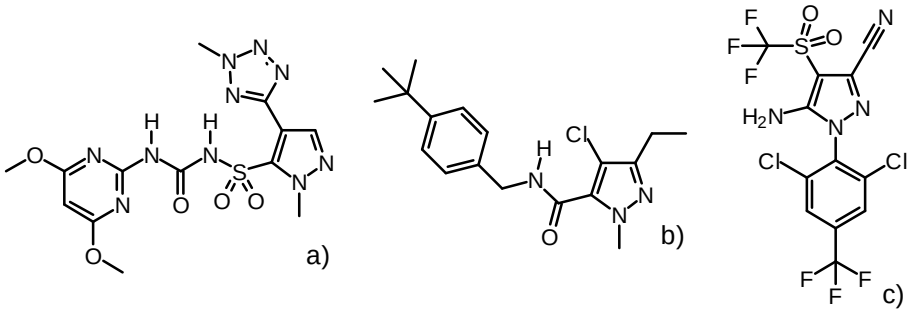


Abb. 7-2: Als Pestizide eingesetzte Pyrazolderivate: a) Azimsulfuron, b) Tebufenpyrad und c) Fipronil.

Pyrazolderivate findet man auch in einigen Pharmazeutika. Lonazolac, das als Grundstruktur eine Pyrazolessigsäure enthält, zeigt eine analgetische Wirkung (siehe Abbildung 7-3a). Von größerer pharmazeutischer Bedeutung sind jedoch die Oxoderivate des Pyrazolins, nämlich Pyrazolin-5-one und Pyrazolidin-3,5-dione (siehe Abbildung 7-3b und 7-3c), die hier allerdings nicht näher untersucht werden.

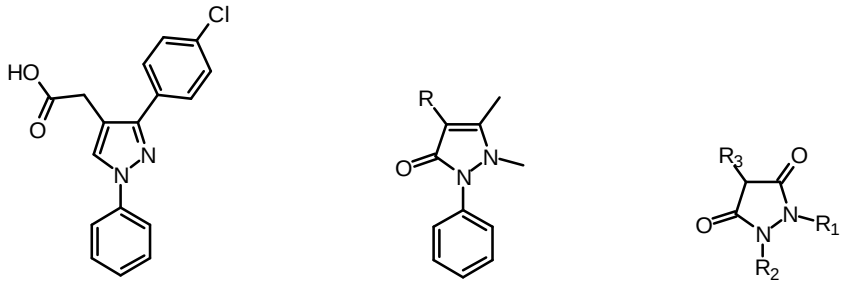


Abb. 7-3: a) Das Analgetikum Lonazolac mit einer Pyrazolessigsäure-Grundstruktur;
b) Grundgerüst für 1-Phenyl-2,3-dimethyl-3-pyrazolin-5-on Derivate, wie Phenazon (R = H), Propylphenazon (R = $CH(CH_3)_2$) und Metamizol-Natrium (R = $N(CH_3)$-CH_2-SO_3Na);
c) Grundgerüst für 1,2-Diphenyl-pyrazolidin-3,5-dione, zu denen beispielsweise Phenylbutazon (R_1 = Phenyl, R_2 = Phenyl, R_3 = C_4H_9) gehört.

Zur Zeit werden vierfach substituierte Pyrazole als Estrogenrezeptorliganden diskutiert[103]. Kombinatorische Bibliotheken von dreifach substituierten Pyrazolen sind in den vergangenen Jahren intensiv von Mitarbeitern pharmazeutischer Unternehmen untersucht worden[104][105].

7.2 Problemstellungen beim Aufbau von Bibliotheken

Der Aufbau einer kombinatorischen Bibliothek kann in vier Stufen eingeteilt werden, wobei jede Stufe einen unterschiedlichen Zeit- und Investitionsaufwand erfordert: Der Literatursuche nach einer geeigneten Reaktion zum Aufbau der Bibliothek, die nur ca. 5% an Zeit- und Investitionsaufwand erfordert, folgt das Verifizieren dieser Reaktion in eigenen Labors (10%). Den größten Anteil mit 70% nimmt die Reaktionsoptimierung und das Erforschen des Bereiches ein, innerhalb dessen die Reaktion zu den gewünschten Produkten führt. Schließlich läuft die experimentelle Durchführung der kombinatorischen Synthese mit nur 15% der Gesamtkosten ab[106]. Bei der Reaktionsoptimierung treten für einen Chemiker vor allem folgende Fragestellungen auf:

- Auswahl geeigneter Reaktionsbedingungen
  Welche Reaktionsbedingungen – wie Lösungsmittel, Katalysator, Reaktionstemperatur etc. – sollen zum optimalen Aufbau der Bibliothek eingesetzt werden?
- Selektivität der Reaktion
  Reagieren alle ausgewählten Edukte zu den erwarteten Produkten oder treten Nebenreaktionen auf?
- Diversität des Reaktionsdatensatzes
  Sind die möglichst diversen Ausgangsverbindungen trotzdem ähnlich genug, um alle nach demselben Reaktionsmechanismus zu reagieren?

Angesichts dieses Löwenanteils an Zeit- und Investitionskosten, der in die Optimierung der Reaktionen fließt, ist man bestrebt, Methoden zu entwickeln, die zu einer Senkung des Zeit- und Investitionsaufwandes führen. Wie in den folgenden Kapiteln erläutert wird, kann bei all diesen Fragestellungen die Reaktionsklassifizierung wertvolle Hilfestellungen geben.

Beim Aufbau von Bibliotheken müssen meist auch Fragen zur Diversität der Ausgangsmoleküle oder Produkte beantwortet werden. Diese Fragen müssen mit anderen Methoden als der Reaktionsklassifizierung angegangen werden, da man prinzipiell mit einer Klassifizierung von Reaktionen keine Ähnlichkeitsanalyse von Molekülen vornehmen kann. Aus diesem Grund darf die Diversität von Reaktionen nicht mit der Diversität von Molekülen, die auf der Ähnlichkeit physikochemischer und biologischer Eigenschaften beruht, verwechselt werden.

7.2.1 Auswahl des Reaktionsmediums

Bevor man mit der Optimierung einer Reaktion beginnen kann, muß man sich für ein geeignetes Reaktionsmedium entscheiden. Die kombinatorischen Bibliotheken der Vergangenheit wurden überwiegend an fester Phase erzeugt, während man heutzutage auch häufig Bibliotheken in flüssiger Phase erzeugt. Will man auf bereits bekannte Synthesen zurückgreifen, so kann das codierte Reaktionszentrum einer kombinatorischen Synthese zum einen in ein Netz projiziert werden, das mit Reaktionen im flüssigen Medium trainiert wurde, wie beispielsweise die Theilheimer Reaktionsdatenbank. Zum anderen kann man das Reaktionszentrum auch als Testdatensatz für ein Netz verwenden, das zuvor mit Reaktionen an fester Phase trainiert wurde, die beispielsweise der SPORE Reaktionsdatenbank entnommen wurden. Gibt das neuronale Netz nur für ein Reaktionsmedium ähnliche Reaktionen aus, so sollte man dieses Medium für das kombinatorische Experiment auswählen. Falls für beide Reaktionsmedien ähnliche Reaktionen gefunden werden, so kann eventuell anhand der folgenden Fragen eine Entscheidung getroffen werden, welches Reaktionsmedium bevorzugt eingesetzt werden sollte.

7.2.2 Selektivität der Reaktion

Neben der Wahl eines Reaktionsmediums muß auch ein Reaktionsmechanismus geplant werden, der die unterschiedlichen Edukte in hoher Ausbeute umsetzt und dabei zu möglichst wenigen Nebenreaktionen führt. Falls manche Edukte in ihrem chemischen Reaktionsverhalten zu unterschiedlich sind, kann es vorkommen, daß manche Reaktionsprodukte gar nicht gebildet werden oder daß Nebenreaktionen in Konkurrenz zur gewünschten Reaktion treten und somit unerwartete Nebenprodukte entstehen. Darüber hinaus wird beim Testen von Mischungen gefordert, daß nicht nur alle möglichen Produkte entstehen, sondern daß die Produkte auch noch in annähernd äquimolaren Mengen vorliegen. Die Reaktionsklassifizierung kann vor allem bei der Klärung der Größe der Bibliothek wertvolle Information liefern. Fin-

det das neuronale Netz zu jeder der möglichen Reaktionen eine ähnliche Reaktion mit hoher Ausbeute, so sollten alle theoretisch möglichen Produkte auch gebildet werden. Werden dagegen für einige Reaktionen nur Gewinnerreaktionen gefunden, die nur in geringer Ausbeute ablaufen oder gar künstlich erzeugt wurden, so verringert sich die Größe der Bibliothek um diese Anzahl an Reaktionen.

Anhand der unterschiedlichen Reaktionstypen in einem Gewinnerneuron kann auf mögliche Nebenreaktionen geschlossen werden. Finden sich in dem Gewinnerneuron ausschließlich Reaktionen des gewünschten Reaktionstyps, so ist die Wahrscheinlichkeit von auftretenden Nebenreaktionen sehr gering. Andererseits ist bei Reaktionen, die in Neuronen projiziert werden, in denen viele verschiedene Reaktionstypen eingetragen wurden, mit Nebenreaktionen zu rechnen. Da chemische Reaktionen sehr von Reaktionsbedingungen abhängig sind, sollten Lösungsmittel, Katalysatoren etc. der Anfragereaktion mit denen der ähnlichsten Reaktionen verglichen werden, um die Wahrscheinlichkeit von auftretenden Nebenreaktionen besser abschätzen zu können.

7.2.3 Diversität des Reaktionsdatensatzes

Während zur Diversität von kombinatorischen Bibliotheken schon viele Methoden entwikkelt[107] und viele Studien veröffentlicht wurden[108], ist die Diversität von Reaktionen noch relativ unerforscht. Das bereits in Kapitel 4.1 vorgestellte Verfahren zum Vergleich von Reaktionsdatenbanken wird dazu auf die kombinatorisch erzeugten Datensätze angewendet. Als erstes wird ein neuronales Netz mit Reaktionen trainiert, die möglichst umfassend den gesamten Reaktionsraum dieses Reaktionstyps abdecken. Anschließend berechnet man für alle Reaktionen, die die Bibliothek aufbauen, die entsprechenden Codierungsvektoren. Da die verschiedenen Substituenten der Ausgangsverbindungen unterschiedliche physikochemische Effekte auf die Reaktionszentren ausüben, werden bei der Projektion dieses Testdatensatzes in das neuronale Netz mehrere Gewinnerneuronen ermittelt, die in ihrer Summe einen Reaktionsteilraum bilden (siehe Abbildung 4-1). Aus der Größe und Position des Reaktionsteilraums kann man Rückschlüsse auf die Diversität der kombinatorischen Synthesereaktionen ziehen.

7.3 Planung kombinatorischer Bibliotheken mittels Reaktionsklassifizierung

In den folgenden Kapiteln wird der Aufbau und die Planung zweier kombinatorischer Bibliotheken aus Pyrazolderivaten erörtert. Die Einsatzmöglichkeiten der Reaktionsklassifizierung bei der Auswahl des Reaktionsmediums wurde für die Pyrazolsynthese bereits in Kapitel 5.4.2 diskutiert. Daher wird in dem folgenden Kapitel die Reaktionsklassifizierung

nur noch zur Abschätzung der Selektivität und Diversität bei der Pyrazolsynthese eingesetzt. Als Wissensbasis zur Vorhersage des Regioisomeren wird das in Kapitel 6.4.3 beschriebene, mit 626 Reaktionen trainierte Netz verwendet (siehe Abbildung 6-20). In der Validierung zeigte dieses Netz eine korrekte Vorhersageleistung von 79,7% (siehe Kapitel 6.4.4).

7.3.1 Aufbau der kombinatorischen Bibliothek I

Zum Aufbau einer kombinatorischen Bibliothek aus Pyrazolderivaten geht man von kommerziell erhältlichen Ausgangsverbindungen aus. Es werden zwei Substruktursuchen im Fluka Chemikalienkatalog (Version 1995/96) mit den in Abbildung 7-4 gezeichneten Anfragestrukturen durchgeführt.

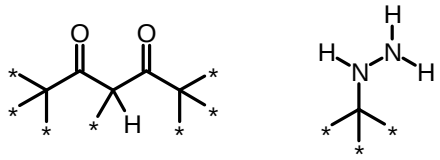


Abb. 7-4: Anfragemoleküle für die zwei Substruktursuchen im Fluka Chemikalienkatalog. Die Sterne symbolisieren die freien Valenzen.

Die erste Suche ergibt für die 1,3-Dicarbonylverbindungen 29 Treffer. Davon werden zwei doppelte Einträge und 13 cyclische 1,3-Dione ausgeschlossen, da diese nicht zu bicyclischen Systemen reagieren. Somit verbleiben 14 1,3-Dicarbonylverbindungen. Diese 14 Moleküle kann man in sechs symmetrische und acht unsymmetrische 1,3-Dicarbonylverbindungen unterteilen. Da nur unsymmetrische 1,3-Dicarbonylverbindungen zu regioisomeren Produkten führen, beschränkt man sich auf die 8 unsymmetrischen Verbindungen. Die zweite Substruktursuche im selben Chemikalienkatalog führte zu 67 Hydrazinen und Hydraziden. Nach Streichung von neun doppelten Einträgen erhält man 58 Verbindungen. Davon sind 31 Hydrazinderivate, 24 Carbonsäurehydrazide und 3 Thiocarbonsäurehydrazide. Bei den Hydrazinderivaten wird auf die 27 Säurehydrazide verzichtet, da im Trainingsdatensatz des Kohonen-Netzes nur wenige Hydrazide enthalten waren und diese ausnahmslos in Konfliktneuronen eingetragen wurden. Die ausgewählten Edukte sind als Strukturformeln in Anhang A.5 abgebildet oder im World-Wide-Web abrufbar (siehe Anhang A.1). Somit wird aus den 31 Hydrazinderivaten und den 8 unsymmetrischen 1,3-Dicarbonylverbindungen eine kombinatorische Bibliothek aus 496 Reaktionen erzeugt. Dabei werden aus je einem Hydrazin-Molekül und einer 1,3-Dicarbonylverbindung alle möglichen Pyrazole gebildet, einschließlich der Regioisomeren. In dieser kombinatorischen Bibliothek werden die Reaktionen, die zu korrespondierenden Regioisomeren führen, paarweise abgespeichert. Reaktion #1 und #2, #3 und #4 etc. führen zu Produkten, die jeweils zueinander regioisomer sind. Eine solche Bibliothek aller möglicher Pyrazolderivate kann relativ einfach mit dem EROS Programmsystem aufgebaut werden[109]. Nach dem Generieren aller möglicher Reaktionen werden diese wie

auf der in Kapitel 6.4.2 beschriebenen Weise codiert. Insgesamt können alle 496 Reaktionen codiert werden.

7.3.2 Vorhersage der Regioisomeren der kombinatorischen Bibliothek I

Bevor der codierte Testdatensatz dem trainierten Netz aus Kapitel 6.4.3 zur Klassifizierung und damit zur Vorhersage der bevorzugt gebildeten Regioisomere übergeben wird, wird noch nach identischen Reaktionen in beiden Datensätzen gesucht. Insgesamt haben der Trainingsdatensatz aus 626 Reaktionen und der Testdatensatz aus 496 Reaktionen 4 Reaktionen gemeinsam. Reaktion #13 aus dem Trainingsdatensatz und #280 aus dem Testdatensatz haben identische Produkte, ebenso die korrespondierenden regioisomeren Produkte, die in den Reaktionen #326 im Trainingsdatensatz und #279 im Testdatensatz zu finden sind. Die dritte und vierte gemeinsame Reaktion ist Reaktion #269 im Trainingsdatensatz und #392 im Testdatensatz und die Reaktion #582 im Trainingsdatensatz und #391 im Testdatensatz. Das Ergebnis des projizierten Testdatensatzes ist in der Kohonen-Karte der Abbildung 7-5 dargestellt. Eine detaillierte Aufstellung über alle Gewinnerneuronen ist in Anhang A.6 wiedergegeben. Dort sind auch die 496 regioisomeren Produkte aufgelistet. Aus Platzgründen werden diese aber nicht – wie die Edukte – als Strukturformeln, sondern als Zahlencode wiedergegeben. Im World-Wide-Web ist der projizierte Datensatz außerdem in einer interaktiven Version abgelegt, die alle Edukte und Produkte als Strukturformeln anzeigt (siehe Anhang A.1).

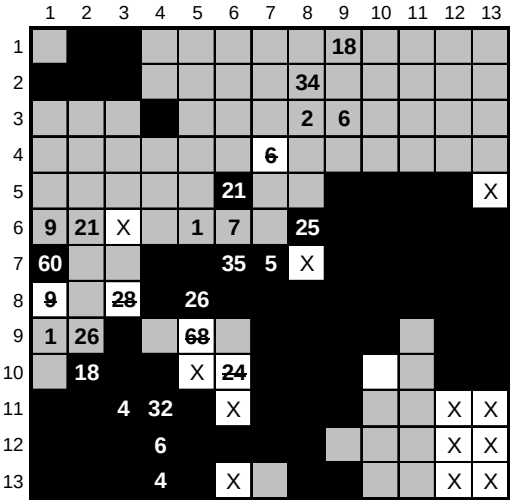


Abb. 7-5: Kohonen-Karte nach Projektion der aus 496 Reaktionen bestehenden kombinatorischen Bibliothek in das mit 626 Reaktionen trainierte Netz der Abbildung 6-20. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen sind mit einem X symbolisiert. Die Zahlen in den Neuronen stellen die Anzahl der eingetragenen Reaktionen dar; eine durchgestrichene Nummer zeigt ein Konfliktneuron im Trainingsdurchgang an.

Von den insgesamt 496 Reaktionen werden 69 Reaktionen (13,9%) in Neuronen eingetragen, die während des Trainingsprozesses ausschließlich mit Reaktionen der Klasse 1 gefüllt

wurden. Diese 69 Reaktionen reagieren also gemäß der Vorhersageleistung des Kohonen-Netzes auf alle Fälle zu den angegebenen Regioisomeren. Für 55 (11,1%) von 496 Reaktionen wird ebenfalls eine Bildungstendenz vorhergesagt, die aber kleiner ist als für die 69 Reaktionen, da diese 55 Reaktionen in Gewinnerneuronen eingetragen werden, in denen im Trainingsdurchlauf sowohl Reaktionen der Klasse 1 und 2 eingetragen wurden. Die Anzahl der Reaktionen der Klasse 1 war in diesen Gewinnerneuronen aber stets größer als die Anzahl der Reaktionen der Klasse 2. Eine einzige Reaktion wird in Neuron (5,6) eingetragen, in das im Training keine Reaktion eingetragen wurde. Aufgrund der benachbarten Neuronen, die am häufigsten der Klasse 1 angehören, zählt dieses Neuron ebenfalls zur Klasse 1 und für die eine Reaktion wird eine gewisse Bildungstendenz vorhergesagt. 122 (24,6%) von 496 Reaktionen werden in Neuronen eingetragen, in die hauptsächlich Reaktionen der Klasse 2 projiziert wurden. Für diese 122 Regioisomere wird nur noch eine sehr kleine Bildungstendenz vorhergesagt. Schließlich sollten dem neuronalen Netz nach 114 Regioisomere (23,0%) gar nicht entstehen.

135 Reaktionen (27,2%) werden in Konfliktneuronen eingetragen. In die Konfliktneuronen wurden im Trainingsdurchgang gleich viele Reaktionen der Klasse 1 und 2 eingetragen, so daß sie weder der Klasse 1 noch der Klasse 2 zugeteilt werden können. Im Testdurchlauf werden insgesamt 132 von den 135 Reaktionen paarweise in Konfliktneuronen projiziert. Diese 132 Reaktionen werden in die Neuronen (7,4), (1,8), (3,8), (5,9) und (6,10) eingetragen. Die Reaktionen, die in diese Konfliktneuronen eingetragen werden, enthalten als Edukte hauptsächlich 2-Acetylcyclohexanon oder 2-Acetylcyclopentanon. Die physikochemischen Effekte der beiden Carbonylgruppen dieser zwei Acetylcycloalkanone sind sehr ähnlich, so daß das Netz zwischen den Regioisomeren nicht differenzieren kann. Eine Zusammenfassung der Bildungstendenzen der regioisomeren Pyrazole des Datensatzes I findet man in Tabelle 7-1.

Reaktionsanzahl	vorhergesagte Klasse	Beurteilung
69 (13,9%)	Klasse 1 (1,0)	hohe Bildungstendenz
55 (11,1%)	Klasse 1 (<1,0)	mittlere Bildungstendenz
1 (0,2%)	Klasse 1 (leeres N.)	mittlere Bildungstendenz
		Σ 125 (25,2%)
135 (27,2%)	Konfliktneuron	--- Σ 135 (27,2%)
122 (24,6%)	Klasse 2 (<1,0)	niedrige Bildungstendenz
114 (23,0%)	Klasse 2 (1,0)	keine Bildungstendenz
		Σ 236 (47,6%)

Tab. 7-1: Ergebnisse zur Vorhersage der Bildungstendenz regioisomerer Pyrazole des kombinatorischen Datensatzes I.

Zählt man alle regioisomeren Produkte aus Reaktionen, die in Neuronen der Klasse 1 eingetragen wurden, zusammen, so besteht die kombinatorische Bibliothek nur noch aus 125

(25,2%) von 496 möglichen Produkten. Schließt man die 135 Reaktionen mit ein, die in Konfliktneuronen eingetragen werden, so kann man für insgesamt 260 Reaktionen (52,4%) eine Bildungstendenz erwarten. Für 236 Regioisomere (47,6%) sagt das neuronale Netz aufgrund der elektronischen Eigenschaften der entsprechenden Edukte nur eine sehr geringe Bildungstendenz der entsprechenden Produkte voraus.

7.3.3 Aufbau der kombinatorischen Bibliothek II

Der zweite Datensatz wurde der Veröffentlichung von Marzinzik und Felder entnommen[110]. Ausgehend von 4 Acetylcarbonsäuren, 35 Carbonsäureester und 41 Hydrazinen synthetisierten sie an fester Phase mit der sogenannten *split and mix* Methode eine Bibliothek von insgesamt 11.480 Pyrazolderivaten. Dabei werden die Acetylcarbonsäuren zuerst über eine Amidbindung zum Linker an das Trägermaterial gebunden, anschließend erfolgt eine Claisen-Kondensationsreaktion zwischen dem Acetylrest und dem Carbonsäureester. Die entstandene 1,3-Dicarbonylverbindung wird anschließend mit der Hydrazinverbindung zu einem dreifach substituierten Pyrazolsystem umgesetzt (siehe Abbildung 7-6).

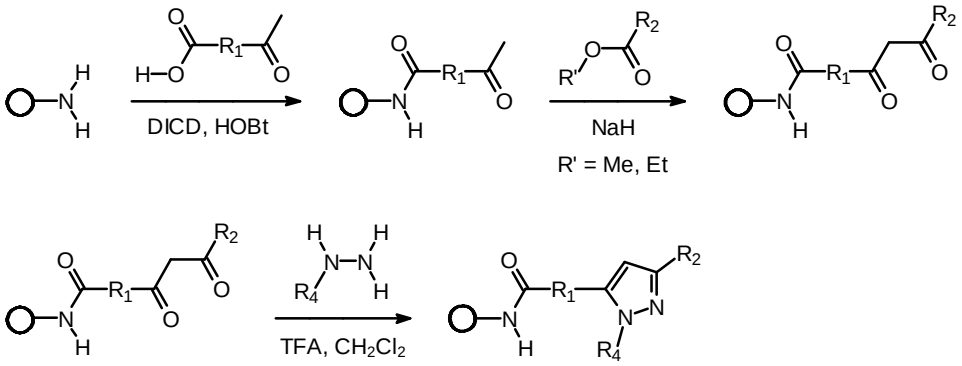


Abb. 7-6: Reaktionsabfolge zum Aufbau der kombinatorischen Bibliothek II an fester Phase.

Für die zweite zu untersuchende Bibliothek werden alle 4 Acetylcarbonsäuren und 33 von 35 Carbonsäureester ausgewählt. Ein Carbonsäureester ist im Anhang der Veröffentlichung falsch wiedergegeben, ein anderer Carbonsäureester enthält einen Pyrazolring, der bei Hinzunahme zu den Ausgangsverbindungen ein sehr zeitintensives manuelles Atom-Atom-Mapping erfordern würde. Daher werden diese zwei Carbonsäureester weggelassen. Bei den 41 Hydrazinverbindungen wird auf die Stammverbindung Hydrazin verzichtet, da diese keine zwei Regioisomere bilden kann. Im Anhang A.7 sind alle Ausgangsverbindungen als Strukturformeln dargestellt und im World-Wide-Web abrufbar (siehe Anhang A.1). Aus den 4 Acetylcarbonsäuren, 33 Carbonsäureester und 40 Hydrazinverbindungen erhält man eine Bibliothek von 10.560 Pyrazolen.

Obwohl die kombinatorische Bibliothek ursprünglich von Marzinzik et al. an fester Phase aufgebaut wurde, werden die Reaktionen des kombinatorischen Datensatzes I nach einer

modifizierten Reaktionsabfolge erzeugt, die einer Synthese in einem Lösungsmittel entspricht (siehe Abbildung 7-7). Die Acetylcarbonsäuren werden nämlich ohne Kupplungsreaktion mit einer Amingruppe direkt mit den Carbonsäureestern zur Reaktion gebracht, bevor die generierten 1,3-Dicarbonylverbindungen mit den Hydrazinverbindungen kombiniert werden. Da einerseits die physikochemischen Unterschiede des am Linker gebundenen Säureamids zur freien Carbonsäuregruppe auf die weit entfernte 1,3-Dicarbonylgruppe nur sehr gering ausfallen, andererseits bei der Codierung der Reaktionen keine Reaktionsbedingungen einfließen, sind die Ergebnisse der Reaktionsklassifizierung mit den Ergebnissen des experimentellen Aufbaus der Bibliothek an fester Phase vergleichbar.

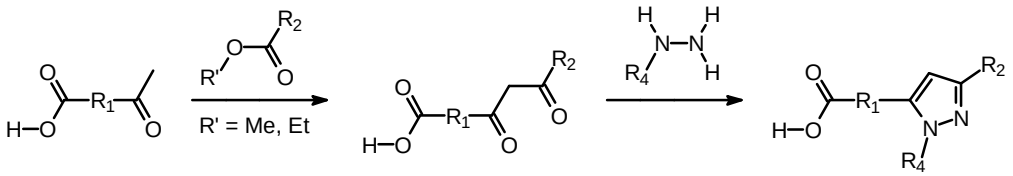


Abb. 7-7: Reaktionsabfolge zum Aufbau der kombinatorischen Bibliothek II in einem Lösungsmittel.

Wie schon im Falle des kombinatorischen Datensatzes I werden die Reaktionen, die zu korrespondierenden Regioisomeren führen, wieder paarweise abgespeichert, d.h. Reaktion #1 und #2, #3 und #4 etc. führen zu regioisomeren Produkten. Nach dem Generieren aller möglicher Reaktionen werden diese wie auf der in Kapitel 6.4.2 beschriebenen Weise codiert. Insgesamt können alle 10.560 Reaktionen codiert werden.

7.3.4 Vorhersage der Regioisomeren der kombinatorischen Bibliothek II

Die Suche nach identischen Reaktionen in dem Trainingsdatensatz aus 626 Reaktionen und der codierten kombinatorischen Bibliothek aus 10.560 Reaktionen ergab keine Übereinstimmung. Somit muß das trainierte Netz für alle 10.560 Regioisomeren die Bildungstendenzen vorhersagen. Das Ergebnis der projizierten kombinatorischen Bibliothek II in die Kohonen-Karte der Abbildung 6-20 ist in Abbildung 7-8 dargestellt.

Von den insgesamt 10.560 Reaktionen werden 1.484 Reaktionen (14,1%) in Neuronen eingetragen, die während des Trainingsprozesses ausschließlich mit Reaktionen der Klasse 1 gefüllt wurden. Die Regioisomere dieser 1.484 Reaktionen sollten also auf alle Fälle entstehen. Für 4.502 (42,6%) von 10.560 Reaktionen wird ebenfalls eine Bildungstendenz vorhergesagt, die aber kleiner ist als für die 1.484 Reaktionen, da diese Reaktionen in Gewinnerneuronen eingetragen werden, in denen im Trainingsdurchlauf mehr Reaktionen der Klasse 1 als der Klasse 2 eingetragen wurden. Der umgekehrte Fall, daß nämlich die Gewinnerneuronen hauptsächlich Reaktionen der Klasse 2 enthalten, betrifft 3.004 Reaktionen (28,4%). Für diese 3.004 Regioisomere wird nur noch eine sehr kleine Bildungstendenz vor-

hergesagt. Schließlich sollten dem neuronalen Netz nach 1.436 Regioisomere (13,6%) gar nicht entstehen.

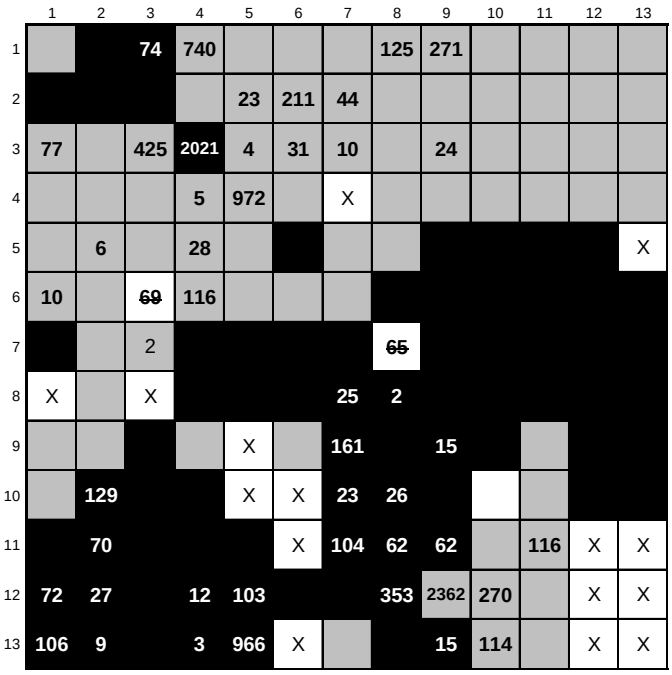


Abb. 7-8: Kohonen-Karte nach Projektion der aus 10.560 Reaktionen bestehenden kombinatorischen Bibliothek II in das mit 626 Reaktionen trainierte Netz der Abbildung 6-20. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen sind mit einem X symbolisiert. Die Zahlen in den Neuronen stellen die Anzahl der eingetragenen Reaktionen dar; eine durchgestrichene Nummer zeigt ein Konfliktneuron im Trainingsdurchgang an.

134 Reaktionen (1,3%) werden in Konfliktneuronen eingetragen. Häufig findet man ein Pyrazolsystem, das entweder mit einer Furanylgruppe und einem Pyrrolderivat oder mit einer Furanylgruppe und einer p-Benzoesäuregruppe substituiert ist, in dem einen Konfliktneuron, während die entsprechenden regioisomeren Produkte in dem anderen Konfliktneuron anzutreffen sind. Bei diesen physikochemisch ähnlichen aromatischen Systemen kann das trainierte Netz also keine zuverlässige Vorhersage in Bezug auf die Bildungstendenzen treffen.

Zählt man alle regioisomeren Produkte aus Reaktionen, die in Neuronen der Klasse 1 eingetragen wurden, zusammen, so besteht die kombinatorische Bibliothek nur noch aus 5.986 (56,7%) von 10.560 möglichen Produkten. Für 4.440 Regioisomere (42,0%) sagt das neuronale Netz aufgrund der elektronischen Eigenschaften der entsprechenden Edukte nur eine sehr geringe Bildungstendenz der entsprechenden Produkte voraus. Eine Zusammenfassung der Bildungstendenzen der regioisomeren Pyrazole des Datensatzes II findet man in Tabelle 7-2.

Reaktionsanzahl	vorhergesagte Klasse	Beurteilung
1.484 (14,1%)	Klasse 1 (1,0)	hohe Bildungstendenz
4.502 (42,6%)	Klasse 1 (<1,0)	Bildungstendenz
		Σ 5.986 (56,7%)
134 (1,3%)	Konfliktneuron	--- Σ 134 (1,3%)
3.004 (28,4%)	Klasse 2 (<1,0)	niedrige Bildungstendenz
1.436 (13,6%)	Klasse 2 (1,0)	keine Bildungstendenz
		Σ 4.440 (42,0%)

Tab. 7-2: Ergebnisse zur Vorhersage der Bildungstendenz regioisomerer Pyrazole des kombinatorischen Datensatzes II.

Unterteilt man die Bibliothek aus 10.560 Reaktionen nach den 40 Hydrazinderivaten in genauso viele kleinere Bibliotheken, und trägt für jede Teilbibliothek die Anzahl der vorhergesagten Reaktionen der Klasse 1 und 2 tabellarisch auf (siehe Tabelle 7-3), so kann man Rückschlüsse auf den Einfluß der Hydrazinkomponente auf die Bildungstendenzen der Regioisomere ableiten. Beispielsweise werden Reaktionen mit den Hydrazinkomponenten 31 oder 35 überdurchschnittlich oft in Konfliktneuronen eingetragen. Bei Einsatz von 3-Nitrophenylhydrazin oder 4-Nitrophenylhydrazin kann für viele Reaktionen, an denen 1,3-Dicarbonylverbindungen umgesetzt werden, die mit aromatischen Resten substituiert sind, die Bildungstendenz nicht mehr eindeutig vorhergesagt werden.

Setzt man dagegen noch elektronenärmere aromatische Hydrazinderivate, wie beispielsweise die mit den Nummern 11, 29, 33 oder 38 ein, so wird überdurchschnittlich vielen Reaktionen eine hohe Bildungstendenz vorhergesagt. Diese vier Hydrazinderivate, nämlich (4-Trifluormethyl-pyrimidin-2-yl)-hydrazin, (7-Chloro-chinolin-4-yl)-hydrazin, (1,3,4-Trimethyl-1*H*-pyrazolo[3,4-*b*]pyridin-6-yl)-hydrazin und (6-Methyl-pyridazin-3-yl)-hydrazin, reagieren ausschließlich mit 1,3-Dicarbonylverbindungen, die mit zwei aromatischen Systemen substituiert sind, zu beiden Regioisomeren.

Aus Tabelle 7-3 geht außerdem hervor, daß keine Teilbibliothek bei der Synthese stark benachteiligt ist, da in allen Subbibliotheken mehr als die Hälfte der erwarteten Regioisomere entstehen sollten, es entsteht also immer wenigstens ein Regioisomer. Die Autoren bestätigen in der Veröffentlichung dieses Ergebnis, indem sie auf die durchwegs hohen Ausbeuten der kombinatorischen Synthese hinweisen. Für einige exemplarisch angegebene Subbibliotheken liegt die durchschnittliche Ausbeute häufig bei 95%, nur selten bei 75% oder 80%.

Hydra-zinderiv.	Konflikt-neuron	leeres Neuron	Klasse 1	Klasse 2
1	0	0	146	118
2	0	0	146	118
3	0	0	147	117
4	3	0	148	113
5	0	0	145	119
6	0	0	146	118
7	3	0	149	112
8	3	0	156	105
9	3	0	150	111
10	4	0	147	113
11	4	0	**158**	**102**
12	3	0	155	106
13	3	0	150	111
14	3	0	149	112
15	2	0	157	105
16	3	0	149	112
17	0	0	146	118
18	3	0	149	112
19	9	0	136	119
20	3	0	148	113

Hydra-zinderiv.	Konflikt-neuron	leeres Neuron	Klasse 1	Klasse 2
21	0	0	146	118
22	3	0	149	112
23	3	0	149	112
24	3	0	148	113
25	3	0	150	111
26	3	0	150	111
27	4	0	147	113
28	3	0	157	104
29	2	0	**158**	104
30	3	0	155	106
31	**17**	0	137	110
32	3	0	148	113
33	3	0	**159**	**102**
34	0	0	146	118
35	**18**	0	134	112
36	3	0	157	104
37	3	0	155	106
38	4	0	**158**	**102**
39	3	0	155	106
40	4	0	151	109

Tab. 7-3: Aufteilung der Bibliothek aus 10.560 Reaktionen in 40 Subbibliotheken. Auffällige Zahlenwerte sind markiert hervorgehoben.

7.3.5 Vergleich der beiden kombinatorischen Bibliotheken

Um die beiden kombinatorischen Bibliotheken mit 496 und 10.560 Reaktionen (siehe Kapitel 7.3.1–7.3.4) vergleichen zu können, werden die Projektionen der Datensätze in die trainierte Kohonen-Karte aus Abbildung 6-20 in Abbildung 7-9 gegenübergestellt. Dabei werden nur die Neuronen mit der jeweiligen Klassenfarbe des trainierten Netzes eingefärbt, in denen mindestens eine Reaktion eingetragen wird.

Man erkennt, daß die beiden Datensätze in der trainierten Karte unterschiedliche Bereiche einnehmen. Die kombinatorische Bibliothek aus 496 Reaktionen, die aus den Ausgangsverbindungen des Fluka-Katalogs aufgebaut wurde, besetzt überwiegend ein Gebiet, das sich diagonal von rechts oben nach links unten erstreckt (siehe Abbildung 7-9a). Wie in Kapitel 6.4.3 diskutiert, treffen in diesem Gebiet häufig die beiden Bereiche der Klasse 1 – hauptsächlich im oberen linken Teil der Karte – und Klasse 2 – im unteren rechten Teil – aufeinander. Daher ist es auch nicht verwunderlich, daß insgesamt 135 Reaktionen (27,2%) in 5 Konfliktneuronen eingetragen werden. Für viele Reaktionen, nämlich 236 (47,6%) von 496 Reaktionen, sagt das neuronale Netz nur eine sehr geringe Bildungstendenz voraus.

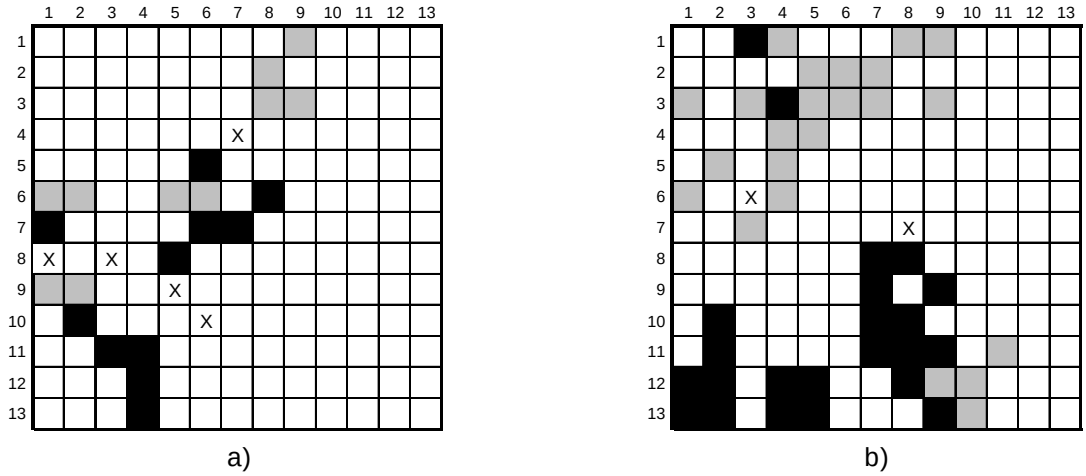


Abb. 7-9: Gegenüberstellung der Projektionen zweier kombinatorischer Bibliotheken mit 496 (a) und 10.560 (b) Reaktionen. Reaktionen, die zu experimentell bestätigten Regioisomeren führen, gehören der Klasse 1 (grau) an, Reaktionen, die zu generierten Regioisomeren führen, der Klasse 2 (schwarz). Konfliktneuronen sind mit einem X symbolisiert.

Die Diversität der Datensätze kann man aus der Zahl der besetzten Neuronen berechnen. Betrachtet man die 13x13 Neuronen des trainierten Kohonen-Netzes als maximal mögliche Diversität von 100%, so beträgt die Diversität der 626 Reaktionen des Trainingsdatensatzes, die insgesamt 164 Neuronen besetzten, 97,0%. Entsprechend ergibt sich für die 26 besetzten Neuronen des kombinatorischen Datensatzes I eine Diversität von 15,4%. Obwohl die Anzahl der Reaktionen des ersten kombinatorischen Datensatzes zum Trainingsdatensatz 79,2% beträgt, fällt die Reaktionsdiversität sehr gering aus. Der Grund ist in den wenigen 1,3-Dicarbonylverbindungen zu suchen. Möchte man die Diversität der Reaktionen in dieser kombinatorischen Bibliothek erhöhen, so sollte man etwa gleich viele Ausgangsverbindungen unter dem Gesichtspunkt einer möglichst hohen Diversität der Moleküle auswählen.

Der zweite kombinatorische Datensatz ist dagegen in zwei Teilbereichen lokalisiert (siehe Abbildung 7-9b). Im linken oberen Gebiet der Kohonen-Karte werden überwiegend Reaktionen eingetragen, bei denen das entsprechende Regioisomere gebildet wird, da hier Neuronen der Klasse 1 vorherrschen. Davon abgetrennt werden im unteren linken und mittleren Gebiet der Kohonen-Karte hauptsächlich die Reaktionen eingetragen, die nicht ablaufen sollten. In die diagonal verlaufenden Grenze der beiden Gebiete werden dagegen nur wenige Reaktionen eingetragen. In die zwei Konfliktneuronen werden insgesamt 134 Reaktionen eingetragen, das entspricht nur 1,3% von insgesamt 10.560 Reaktionen. Im Vergleich zum ersten kombinatorischen Datensatz mit 27,2% können für diesen Datensatz die Bildungstendenzen nahezu aller Reaktionsprodukte eindeutig vorhergesagt werden. Das Netz sagt zudem für die meisten Produkte, nämlich 5.986 (56,7%) von 10.560, eine hohe Bildungstendenz voraus. Marzinzik et al. ist es somit gelungen, die Ausgangsverbindungen für diese kombinatorische Synthese so auszuwählen, daß stets mindestens eines der beiden möglichen Produkte entsteht.

Obwohl der zweite Datensatz fast 21 mal mehr Reaktionen enthält als der erste kombinatorische Datensatz sind nur insgesamt 48 Neuronen besetzt. Das entspricht einer Reaktionsdiversität von 28,4%. Ursache für diese relativ kleine Diversität des sehr großen Datensatzes liegt hier in der Auswahl von nur 4 Acetylcarbonsäuren, bei denen zudem die Acetylgruppe an einem aromatischen System gebunden ist. Zur Vergrößerung der Diversität der Reaktionen sollte man daher als erste Maßnahme mehrere Acetylcarbonsäuren unter dem Gesichtspunkt einer möglichst hohen Diversität der Moleküle auswählen.

7.4 Diskussion des Einsatzes der Reaktionsklassifizierung beim Aufbau kombinatorischer Bibliotheken

Wie in den vorangegangenen Kapiteln gezeigt wurde, liefert die Reaktionsklassifizierung wertvolle Information beim effizienten Aufbau kombinatorischer Bibliotheken, beispielsweise bei der Bestimmung der Diversität von Reaktionsdatensätzen.

Die Auswahl der Ausgangsverbindungen durch Substruktursuche im Fluka-Katalog führte zu einer aus 496 Reaktionen bestehenden Bibliothek, die nur eine kleine Reaktionsdiversität von rund 15% aufweist. Angesichts der sehr hohen Reaktionsanzahl von 10.560 Reaktionen weist auch der zweite Datensatz nur eine kleine Reaktionsdiversität von rund 28% auf.

Die Vorhersagen zur Auswahl des Reaktionsmediums, der Selektivität und Diversität von Reaktionen können beliebig oft wiederholt werden, falls sich beispielsweise die Ausgangsverbindungen ändern sollten. Ist eine Ausgangsverbindung beispielsweise nicht mehr kommerziell erhältlich oder will man eine weitere Verbindung hinzunehmen, weil man mit ihr eine größere Diversität der Produkte erwartet, so bietet die Reaktionsklassifizierung eine Möglichkeit, die resultierenden Veränderungen auf die Selektivität, Diversität der Reaktionen etc. unmittelbar abzurufen und darzustellen.

Obwohl momentan zur Codierung der Reaktionen nur physikochemische Effekte verwendet werden, liegt die Vorhersagequalität bei rund 80%. Durch Hinzunahme weiterer Effekte, vor allem sterischer Einflüsse, ließe sich die Vorhersageleistung bestimmt steigern.

8 Realisierung der Reaktionsklassifizierung mittels CORA

8.1 Die Programmiersprache Tcl/Tk

Tcl/Tk ist eine leistungsfähige Skriptsprache mit graphischer Benutzerschnittstelle. Sie ist eine interpretierte Sprache, die kostenfrei erhältlich[111] ist und auch in kommerziellen Paketen – ohne Entrichten von Lizenzgebühren – verwendet werden darf.

Die Entwicklung von Tcl (***T**ool **C**ommand **L**anguage*) begann 1988 an der University of California in Berkeley unter der Leitung von John Ousterhout. Oberstes Ziel war die Entwicklung einer allgemein verwendbaren Kommandosprache für bestehende Applikationen; diese wurde als eine C-Bibliothek implementiert, was zur Folge hatte, daß diese Sprache leicht in beliebige Anwendungsprogramme integriert werden kann. Weiterhin achtete man darauf, daß die Sprache durch Erweiterungen an Anwendungsprogramme angepaßt werden kann. Die Anforderung einer graphischen Benutzerschnittstelle (***G**raphical **U**ser **I**nterface*, GUI), die flexibel und wieder verwendbar implementiert werden sollte, führte zur Entwicklung von Tk, dem graphischen ***T**ool**k**it* auf der Basis der unter UNIX standardisierten X-Window-Bibliothek. Mit den Arbeiten an Tk wurde im Jahre 1989 begonnen. In den 90er Jahren fand Tcl/Tk eine schnelle Verbreitung in der UNIX-Welt und behauptete sich nach der Portierung auf IBM-kompatible PCs und Macintosh-Rechner im Mai 1996 auch auf diesen Systemen. Gegenwärtig ist Tcl/Tk für folgende Plattformen erhältlich: Windows95/98/NT, Macintosh, UNIX.

Zwischenzeitlich wurde das Paket von Mitarbeitern von Sun Microsystems Laboratories, Inc. unter der Leitung von John Ousterhout gewartet und weiterentwickelt. Im Januar 1998 wurde von John Ousterhout eine eigene Firma, Scriptics, zur Weiterentwicklung dieser Interpretersprache gegründet, die im Mai 2000 in „Ajuba Solutions“ umbenannt wurde.

Als Dokumentation kann man entweder auf die Online-Manualseiten einer Tcl-Distribution zurückgreifen, oder auf viele kommerziell erhältliche Bücher[112],[113].

Aus all diesen Gründen ist Tcl/Tk gerade auch für den universitären Einsatz eine hervorragende Programmiersprache.

8.2 Das CACTVS-Informationssystem

Das von Ihlenfeldt entwickelte Informationssystem für chemische Anwendungen, heißt CACTVS[114]. Es ist ein verteiltes Client/Server-System zur Berechnung, Verwaltung, Analyse und Visualisierung chemischer Information[115]. Die meisten Funktionalitäten des in C geschriebenen, modularen Programmsystems können über eine Programmiersprache aufgerufen werden, die in ihrer Syntax der Tcl/Tk-Sprache gleicht. Alle Tcl/Tk-Befehle sind auf dem

Stand der Tcl/Tk-Version 8.2 ebenfalls in dem Programmsystem integriert. Die Mächtigkeit dieses sich laufend weiterentwickelten Systems wird nicht nur anhand der ca. 370.000 Quellcodezeilen deutlich, sondern auch anhand zahlreicher Publikationen[116], [117].

Um die Funktionalität anderer im Arbeitskreis von Gasteiger entwickelter Programmsysteme in dieses Informationssystem einzubringen, ging man in den letzten Jahren dazu über, CORINA, ein Programm zur Erzeugung 3-dimensionaler Strukturen, und KMAP als zusätzliche Module zu CACTVS zu implementieren. Sämtliche Funktionalitäten zum Erzeugen von Kohonen-Netzen sind bereits in der neu implementierten Programmversion 4.0 als Modul verfügbar[118]. Angesichts der Mächtigkeit des CACTVS-Systems und der Homogenität eines Programms zur Reaktionsklassifizierung, das sowohl physikochemische Effekte berechnen kann, als auch einen Kohonen-Generator enthalten muß, wurde bei der Implementierung eines Programmsystems zur Reaktionsklassifizierung auf dieses chemische Informationssystem zurückgegriffen.

8.3 CORA

Die Entwicklung eines Programmsystems zur Klassifizierung organischer Reaktionen begann im Jahre 1996. Vor dieser Zeit wurde mit einer Vielzahl kurzer Programme, die darüber hinaus in unterschiedlichen Programmiersprachen abgefaßt waren, die Codierung von Reaktionszentren durchgeführt. Ein automatisch durchgehender Ablauf von der Angabe eines Reaktionsdatensatzes bis zum trainierten neuronalen Netz war nicht möglich. Außerdem konnte man auch nur bestimmte Reaktionszentren behandeln, weil Programmteile auf den jeweiligen Typ angepaßt werden mußten. Aus diesem Grund wurde im Rahmen dieser Arbeit ein Programmsystem namens CORA konzipiert und verwirklicht, das in diesem Kapitel vorgestellt wird. Der Name CORA ist ein Akronym für „***C****lassification of* ***O****rganic* ***R****eactions for* ***A****pplications*“.

Bei der Planung dieses Programmsystems wurde hoher Wert auf eine leichte Integration bzw. Anknüpfung an etablierte Programmsysteme des Arbeitskreises gelegt. Ein Zugriff auf die Wissensbasis sollte sowohl von dem Reaktionsvorhersagesystem EROS als auch von dem Syntheseplanungsprogramm WODCA aus möglich sein. Da sowohl die graphische Benutzeroberfläche von EROS und WODCA in Tk geschrieben sind, fiel die Wahl für die Benutzeroberfläche von CORA ebenfalls auf diese Programmiersprache. Das gut gewartete CACTVS Programmsystem enthält immer alle Standardbefehle der aktuellen Tcl/Tk-Version. Darüber hinaus bietet es einen für die chemische Informationsverarbeitung erweiterten Befehlssatz von Tcl/Tk an (siehe Kapitel 8.2). CACTVS stellt somit die ideale Programmiersprache für das Klassifizierungsprogramm dar.

Die zuerst geschriebene Programmversion 1.0 von CORA konnte nur spezielle Reaktionstypen klassifizieren, da für jeden zu untersuchenden Reaktionstyp eine kurze Regeldatei

geschrieben werden mußte. Diese vom Programmsystem getrennt gehaltenen Regeldateien sind notwendig, um die Reihenfolge der zu codierenden Atome oder Bindungen des Reaktionszentrums eindeutig festzulegen. Die Entwicklung einer Methode, welche die Reihenfolge von Atomen und Bindungen mittels des modifizierten USMILES-Codes eindeutig festlegt, führte zu einer überarbeiteten Version 2.0, die nun jedes Reaktionszentrum in einer eindeutigen Weise codieren kann.

Der typische Ablauf einer Reaktionsklassifizierung ist in Abbildung 8-1 dargestellt. Aus dieser Abbildung geht auch das Zusammenspiel der verschiedenen Programmsysteme, wie CORA, PETRA, KMAP, WODCA und EROS hervor.

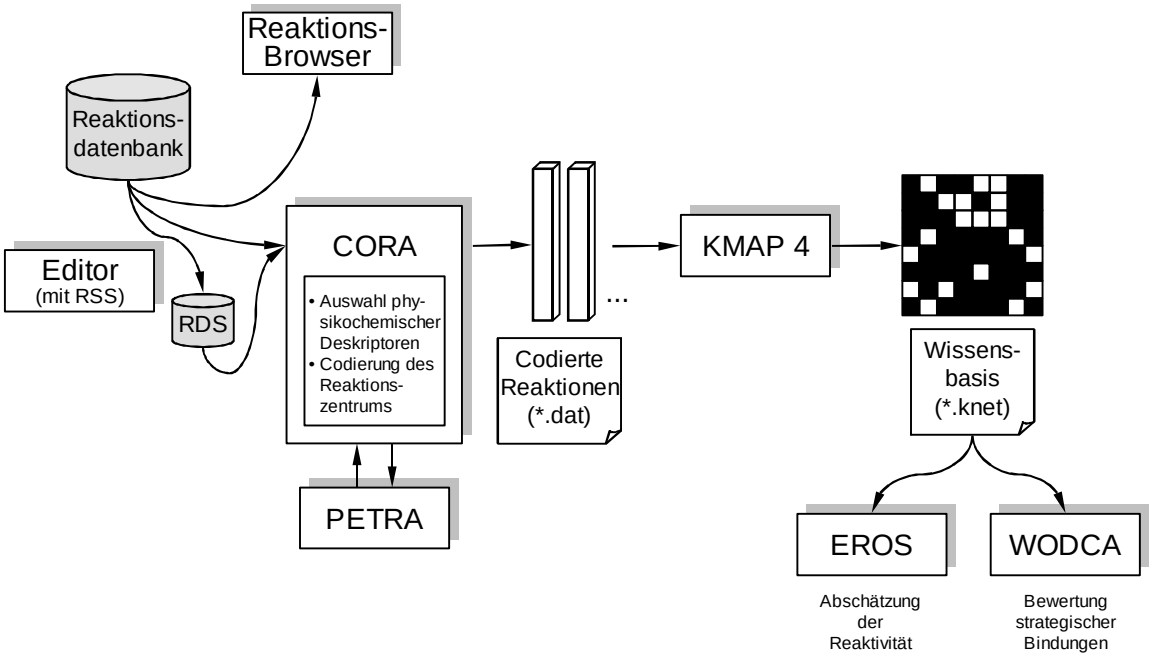


Abb. 8-1: Darstellung des Ablaufs einer Reaktionsklassifizierung: Die Reaktionen einer Reaktionsdatenbank oder eines Reaktionsdatensatzes (RDS) werden mittels CORA codiert und dem Kohonen Map Generator (KMAP) als Datei übergeben. Die klassifizierte Kohonen-Karte stellt die Wissensbasis für andere Programmsysteme wie EROS und WODCA dar.

CORA kann sowohl komplette Reaktionsdatenbanken, als auch Reaktionsdatensätze codieren. Einen Reaktionsdatensatz erhält man beispielsweise, wenn man alle Reaktionen einer Datenbank einer Reaktionssubstruktursuche unterzieht. Hierzu kann der um eine Reaktionssubstruktursuche erweiterte Moleküleditor des CACTVS-Systems herangezogen werden. In diesem Moleküleditor wird das Reaktionszentrum als Suchanfrage für die Substruktursuche formuliert (siehe Abbildung 8-2).

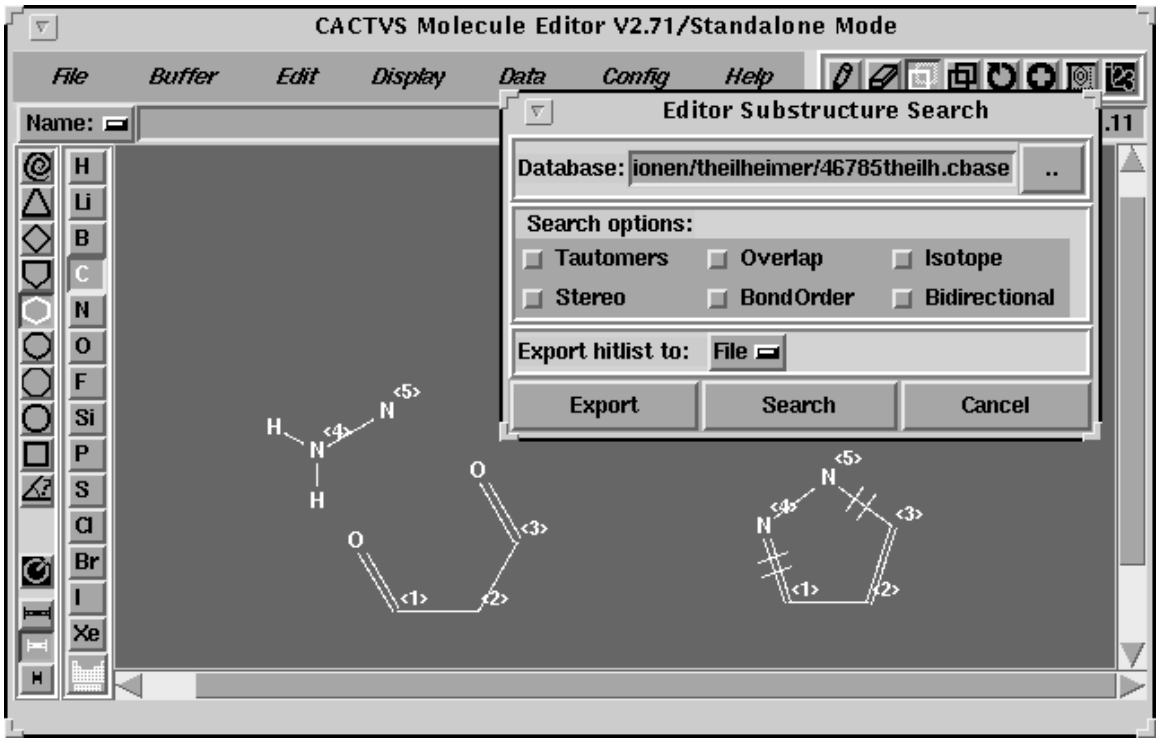


Abb. 8-2: Formulierung einer Suchanfrage für eine Reaktionssubstruktursuche im Moleküleditor. In den beiden Edukten und dem Produkt sind an den Atomsymbolen die Atom-Atom-Mapping-Nummern angegeben, im Produkt sind neu geknüpfte Bindungen mit zwei Strichen markiert. In dem kleineren Fenster zur Substruktursuche können weitere Suchoptionen angegeben werden.

Die Trefferliste dieser Suche wird anschließend als Datei in einem Dateiformat für Reaktionen exportiert, wobei das mit der Suchanfrage übereinstimmende Reaktionszentrum für jede Reaktion abgespeichert wird. Somit ist gewährleistet, daß die Reihenfolge aller Atome oder Bindungen des Reaktionszentrums eindeutig festgelegt ist. Im Falle der Pyrazole ist dadurch die Unterscheidung der Reaktionszentren für beide Regioisomere möglich (siehe Kapitel 6.4.2). Nach dem Export dieses Reaktionsdatensatzes kann dieser in CORA eingelesen werden.

In Abbildung 8-3 ist das Hauptfenster von CORA wiedergegeben, von dem aus alle Funktionen und verschiedene Konfigurationsfenster aufrufbar sind. Die Konfiguration umfaßt zunächst die Festlegung, ob Atome oder Bindungen auf der Edukt- oder Produktseite untersucht werden sollen. Zweitens wird festgelegt, welche Methode zur Reaktionszentren-Erkennung angewandt werden soll, sofern keine Information über das Reaktionszentrum in der Datei abgespeichert vorliegt. Anschließend werden die Deskriptoren ausgewählt und die Skalierungsmethode bestimmt. Bei dem anschließenden Codierungsprozeß wird für jede Reaktion das PETRA Programmsystem aufgerufen, das sämtliche physikochemische Atom- und Bindungseigenschaften berechnet. CORA schreibt für jede Reaktion die ausgewählten Eigenschaften nach deren Skalierung in einer einheitlichen Weise in eine sogenannte DAT-Datei

heraus. In den anderen erzeugten Dateien werden beispielsweise die einzelnen Reaktionszentren in einem Molekülformat und dem USMILES-Code heraus geschrieben, sowie die Ausbeuten der Reaktionen oder Informationen zur Konfiguration von CORA.

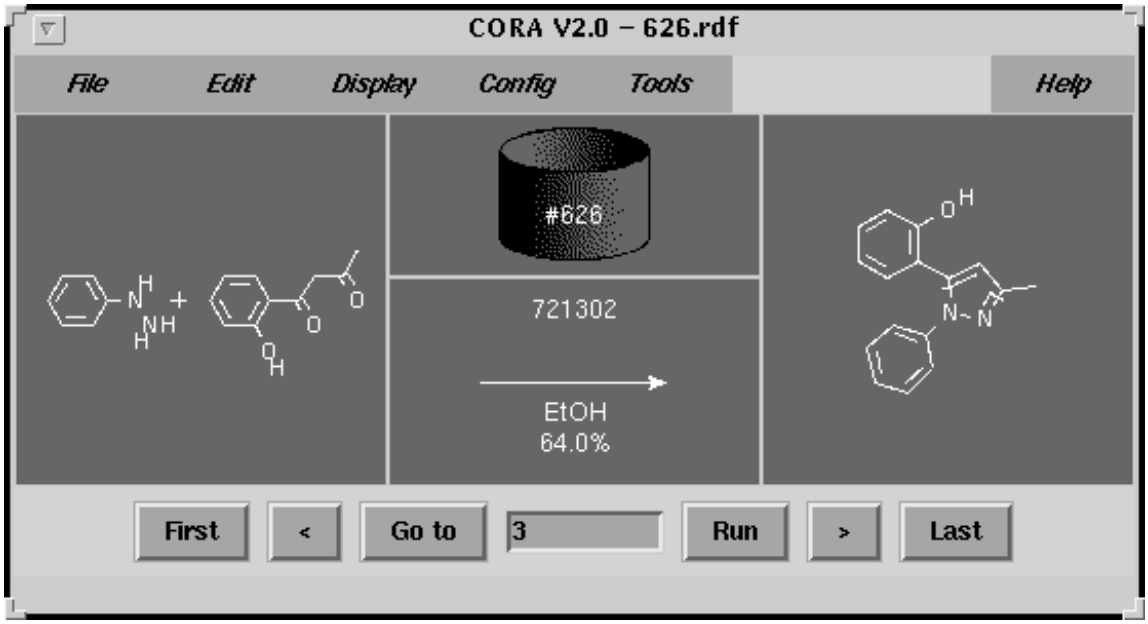


Abb. 8-3: Das Hauptfenster von CORA. Die Edukte jeder Reaktion werden im linken Fenster, die Produkte im rechten Fenster dargestellt. In der Mitte sind die Größe des Datensatzes, die Reaktionsnummer und die Reaktionsbedingungen (Katalysator, Lösungsmittel, Ausbeute) wiedergegeben.

Die DAT-Datei, in der die codierten Reaktionen abgespeichert sind, wird danach dem Kohonen-Map-Generator zum Trainieren eines Kohonen-Netzes übergeben. KMAP führt die Klassifizierung des Datensatzes anhand einer Regeldatei aus und schreibt schließlich mehrere Dateien heraus, die als Wissensbasis für andere Programmsysteme herangezogen werden können. So können beispielsweise das Reaktionsvorhersagesystem EROS zur Abschätzung der Reaktivität und das Syntheseplanungsprogramm WODCA zur Bewertung strategischer Bindungen auf die klassifizierten Reaktionen zugreifen.

Alle Reaktionsdatensätze aus den Anwendungskapiteln dieser Arbeit wurden mit CORA in der Version 2.0 codiert. Im Anwendungskapitel 4 wurde die Theilheimer Reaktionsdatenbank mit insgesamt 46.785 Reaktionen, die viele verschiedene Reaktionszentren aufweisen, mit der in Kapitel 3.3 beschriebenen Standardmethode codiert. Diese codierte Datenbank wurde in den Kapiteln 5.4 und 5.5 bei der Planung organischer Synthesen eingesetzt, um strategische Bindungen zu finden und zu bewerten. Die momentan noch in CORA implementierten Prozeduren zur Suche und Bewertung strategischer Bindungen sollen zukünftig in WODCA eingebunden werden. Die zweite Codierungsmöglichkeit, nämlich die Beschreibung eines Reaktionsdatensatzes, der aus Reaktionen mit gleichen Reaktionszentren besteht, wurde ebenfalls mit CORA verwirklicht und ist in Kapitel 6.4.2 erläutert. Auf diesen klassifizierten Datensatz kann EROS bzw. CORA zugreifen, um für eine Anfragereaktion die ähn-

lichsten Reaktionen zu bestimmen und somit das Hauptreaktionsprodukt vorherzusagen (siehe Kapitel 6.4.6).

Neben den Algorithmen zur Codierung von Reaktionen basierend auf ausgewählten physikochemischen Effekten der Atome oder Bindungen des Reaktionszentrums enthält CORA weitere nützliche Hilfsmittel. Dazu zählt beispielsweise ein Programm, das die resultierenden Kohonen-Karten in das World-Wide-Web spezifische Dateiformat HTML und die Reaktionsbeispiele in das GIF-Graphikformat konvertieren kann. Somit ist es möglich die klassifizierten Datensätze weltweit abrufbar anzubieten.

9 Diskussion

In den vorangegangenen Kapiteln wurde gezeigt, daß sich die vorgestellte Methode zur Codierung und Klassifizierung organischer Reaktionen in vielen Anwendungsgebieten als sehr leistungsstark erwiesen hat. Die Vor- und Nachteile der entwickelten Methode werden in diesem Kapitel nochmals zusammengefaßt dargestellt. Zu den Stärken und Vorteilen der ausgearbeiteten Methode zählen:

- Erweiterbarkeit und Substituierbarkeit der Deskriptoren
  Die gewählten physikochemischen Effekte können leicht durch andere Deskriptoren ersetzt werden. Falls PETRA in Zukunft neue Atom- oder Bindungseigenschaften berechnen kann, können diese entweder auf Edukt- oder Produktseite hinzugenommen werden oder vorhandene Eigenschaften ersetzen. Des weiteren können auch Effekte hinzugenommen werden, die Reaktionsbedingungen beschreiben können, wie beispielsweise den Einfluß des Katalysators über die Säuredissoziationskonstante[119] oder den Einfluß des Lösungsmittel über die Lösungsmittelpolarität nach Reichardt[120].
- Schnelle Berechnung der Codierungsvektoren
  Durch den Einsatz von physikochemischen Deskriptoren, die durch empirische Methoden berechnet werden, ist eine schnelle Berechnung der Codierungsvektoren möglich. Dies ist eine wichtige Voraussetzung um auch große Datensätze in der Größenordnung von 10.000−100.000 Reaktionen behandeln zu können.
- Interpretierbarkeit des Vektors
  Der Codierungsvektor enthält die physikochemischen Deskriptoren in einer zuordbaren, einfach transformierten Form. Jede Position des Codierungsvektors kann eindeutig einem Atom oder einer Bindung und einem physikochemischen Effekt zugeordnet werden. Die Skalierung der physikochemischen Zahlenwerte basiert auf einer linearen Transformationsgleichung. Daher sind Rückschlüsse auf die physikochemischen Effekte am Reaktionszentrum ausgehend von einem Codierungsvektor leicht möglich.
- Codierung von Konstitutionsisomeren
  Der Algorithmus zur Codierung von Reaktionen erlaubt es, für konstitutionsisomere Reaktionszentren einen unterschiedlichen Codierungsvektor zu berechnen (siehe Kapitel 6.4.2).
- Breites Anwendungsgebiet
  Das Klassifizieren von Reaktionen nach der hier vorgestellten Methode kann in einer Reihe von Anwendungsgebieten erfolgreich eingesetzt werden. Zum einen können im Bereich der Syntheseplanung strategische Bindungen erkannt und bewertet werden

(siehe Kapitel 5), zum anderen können Reaktionsprodukte bei der Reaktionsvorhersage vorausgesagt werden (siehe Kapitel 6).

- Flexible, erweiterbare und laufend aktualisierbare Wissensbasis
  Infolge der einfachen und schnellen Codierung und Klassifizierung eines Reaktionsdatensatzes ist die abgeleitete Wissensbasis sehr flexibel, erweiterbar und leicht auf neuestem Stand zu halten. Die Aktualität der zunehmenden Informationsmenge an organischen Reaktionen kann somit leicht auf die Wissensbasis übertragen werden. Auf die Vorteile einer austauschbaren, klassifizierten Kohonen-Karte wurde vor allem in den Kapiteln 5.4 und 5.5 bei der Planung organischer Synthesen eingegangen.

Diesen Vorteilen der neu entwickelte Methode stehen aber auch folgende Einschränkungen und Nachteile gegenüber:

- Verfügbarkeit einer qualitativ hochwertigen, elektronischen Reaktionsdatenbank
  Eine Wissensbasis kann nur von solchen Reaktionen aufgebaut werden, die bereits in Reaktionsdatenbanken gespeichert sind. Stehen keine eigenen Datenbanken zur Verfügung, so muß man unter Umständen teure, kommerzielle Reaktionsdatenbanken einsetzen. Zur Zeit spiegelt sich auch die unvollständige oder fehlerhafte Codierung von Reaktionen in elektronischen Datenbanken in der Wissensbasis wider. Je qualitativ höherwertig eine Reaktionsdatenbank in Bezug auf Reproduzierbarkeit der Ergebnisse und sorgfältiger Erfassung der Reaktionen ist, desto genauere Vorhersagen lassen sich mit dem trainierten Netz treffen.

- Beschränkung auf eine maximale Größe des Reaktionszentrums
  Die entwickelte Codierungsmethode setzt eine maximale Größe des Reaktionszentrums voraus, um eine konstante Vektorlänge zu garantieren. Theoretisch kann die Anzahl der maximalen Atome und Bindungen so groß gewählt werden, wie es dem größten Reaktionszentrum entspricht. Dann würde keine Reaktion wegen eines zu großen Reaktionszentrums ausscheiden. Allerdings steigen mit zunehmender Vektorlänge auch die Rechenzeiten zum Trainieren eines neuronalen Netzes linear an, so daß man aus diesem Grund vor allem bei großen Datensätzen die Vektorlänge möglichst klein wählen sollte. Daher wurde die maximale Größe des Reaktionszentrums deutlich unterhalb des größten Reaktionszentrums (siehe Kapitel 3.3.3) gewählt. Würden beispielsweise Folgereaktionen in den Reaktionsdatenbanken in separaten Gleichungen abgespeichert werden, so wären immer nur eine geringe Anzahl von gebrochenen und geknüpften Bindungen zu codieren.

- Existenz eines Reaktionszentrums
  Die vorgestellte Methode zur Codierung von Reaktionen setzt voraus, daß das Reaktionszentrum bereits in der Reaktion markiert vorliegt. In allen kommerziell erhältlichen

Datenbanken ist für jede Reaktion ein Reaktionszentrum abgespeichert, so daß dieser Nachteil nicht sehr häufig in Erscheinung tritt. Da ein automatisches Erkennen des Zentrums zur Zeit nicht implementiert ist, können Reaktionen ohne vorgegebenes Zentrum auch nicht codiert werden. Außerdem liefert die Methode unvollständig codiert Reaktionszentren, falls das Reaktionszentrum nicht komplett in der Datenbank angegeben wird. Dies ist vor allem ein Problem, wenn im Reaktionszentrum ein Wasserstoffatom enthalten ist. In der Datenbank sind solche Wasserstoffatome nur sehr selten abgespeichert. Ein Reaktionsdatensatz ohne vorgegebenes Reaktionszentrum kann jedoch – wie in Kapitel 6.4.2 gezeigt wurde – mittels einer Reaktionssubstruktursuche codiert werden.

– Fehlende Codierung der Stereochemie
  Der Codierungsvektor eines Reaktionszentrums enthält momentan keine stereochemische Information. Da der Anteil der Reaktionen aus Datenbanken mit Stereoisomerie-Information momentan noch gering ist, wird zur Zeit noch auf die Codierung der Stereochemie verzichtet. Beispielsweise sind in der Theilheimer Reaktionsdatenbank mit insgesamt 46.785 Reaktionen nur für 6.485 Reaktionen (13,9%) Angaben zur Stereochemie abgespeichert worden. Die Berechnung eines Deskriptors, der die Stereochemie des Reaktionszentrums ausdrückt, könnte auf den Arbeiten von Sousa et al. aufbauen[121].

Sowohl bei der Codierung, der Klassifizierung als auch bei den Anwendungsmöglichkeiten ergeben sich eine Reihe erfolgversprechender Ansätze, die in weiterführenden Forschungsarbeiten untersucht werden sollten.

10 Zusammenfassung

Durch die rasche Entwicklung der Technologien im Bereich der elektronischen Datenverarbeitung ist es möglich geworden, riesige Datenbanken im Bereich der Chemie aufzubauen, die einen effizienten Zugang zu chemischen Verbindungen, Reaktionen und den zugehörigen Fakten – wie chemische, physikalische und biologische Eigenschaften, NMR-, IR-, MS-Spektren, Reaktionsbedingungen etc. – ermöglichen.

Angesichts dieser Informationsflut im chemischen Sektor, die in den letzten Jahren durch das Aufkommen der kombinatorischen Chemie noch verschärft wurde, ist ein Bedarf an neuen Methoden zur Auswertung der immensen Datenmengen entstanden. Der Schwerpunkt sollte bei einer solchen Auswertung auf einer Wissensextraktion liegen, d.h. dem Ableiten von Regeln und Gesetzmäßigkeiten aus den wiedergewonnenen Daten. Die so erhaltenen Ergebnisse können wieder zur effizienten Lösung von Problemen der heutigen Chemie einfließen.

In der vorliegenden Arbeit wurde eine neu entwickelte Methode vorgestellt, die einerseits eine Wissensextraktion aus Reaktionsdatenbanken vornimmt, andererseits aber auch bei vielen Fragestellungen in diversen Bereichen der Chemie nutzbringend eingesetzt werden kann. Diese Methode beschreibt eine organische Reaktion anhand physikochemischer Effekte von Atomen und Bindungen des Reaktionszentrums und klassifiziert den so codierten Reaktionsdatensatz mit neuronalen Netzen nach Kohonen. Aufgrund ihrer wichtigsten Eigenschaft, der Lernfähigkeit, können Kohonen-Netze induktiv aus einer Reihe von Einzelreaktionen Wissen in generalisierter Form im neuronalen Netz speichern. Die Methode wendet somit dasselbe Lernverfahren an, mit dem auch jeder Chemiker sein Wissen erlangt hat, nämlich an einer Reihe von Einzelreaktionen. Durch die Erkenntnis von Gemeinsamkeiten und Unterschiede werden die verschiedenen Reaktionstypen wie nucleophile oder elektrophile Substitutionsreaktionen, Friedel-Crafts-Alkylierung oder -Acylierung etc. gelernt, und ein Verständnis für die Reaktivität funktioneller Gruppen entwickelt.

Wie in dieser Arbeit gezeigt wurde, ergeben sich vielfältige Anwendungsmöglichkeiten für die in Reaktionsdatenbanken gespeicherte Information. Das Ausarbeiten eines Standardverfahrens, das in allen Anwendungsbereichen gute Ergebnisse liefert, war eines der Ziele dieser Arbeit.

Die Möglichkeiten der Wissensextraktion wurden anhand der Theilheimer Reaktionsdatenbank vorgestellt, in der alle wichtigen Reaktionstypen der organischen Chemie enthalten sind. Die Klassifizierung dieser Datenbank führt zu einer chemisch gut interpretierbaren Karte, in der die verschiedenen Reaktionstypen voneinander getrennt dargestellt werden. Das erste Anwendungsgebiet, der auf einem Ähnlichkeitsmaß basierende Datenbankenvergleich, führte die Unterschiede von Reaktionsdatenbanken, in denen Reaktionen an fester Phase

gespeichert sind, im Vergleich zur Theilheimer Reaktionsdatenbank, in die ausschließlich Reaktionen im flüssigen Medium aufgenommen wurden, vor Augen. Der Datenbankenvergleich kann auch zum Aufzeigen der zeitlichen Entwicklung von Reaktionsdatenbanken herangezogen werden, wie am Beispiel der ChemInform RX-Reaktionsdatenbanken erläutert wurde.

Im Bereich der computergestützten Syntheseplanung wurden ebenfalls Einsatzmöglichkeiten für klassifizierte Reaktionsdatenbanken diskutiert. Am Beispiel zweier Zielverbindungen wurde die Suche und Bewertung strategischer Bindungen erläutert, wobei eine Retrosynthese sowohl über Reaktionen in flüssiger Phase, als auch über Festphasenreaktionen ins Auge gefaßt wurde.

Als drittes Einsatzgebiet der Reaktionsklassifizierung wurde die Reaktionsvorhersage angeführt. Hier wurde eine Wissensbasis mit regioisomeren Pyrazolderivaten aufgestellt, um die Bildungstendenzen regioisomerer Pyrazole abschätzen zu können. Das trainierte neuronale Netz kann in rund 80% der Fälle das korrekte Pyrazolderivat vorhersagen.

Bei der Planung kombinatorischer Bibliotheken kann das aus Reaktionsdatenbanken erworbene Wissen ebenfalls erfolgversprechend eingesetzt werden. Wie am Beispiel der Pyrazole gezeigt wurde, kann das neuronale Netz bei der Auswahl des Reaktionsmediums, der Selektivität der Reaktionen sowie der Diversität des Reaktionsdatensatzes wertvolle Hilfestellungen geben. Dies wurde anhand zweier kombinatorischer Datensätze aus 496 und 10.560 Einzelreaktionen erläutert.

In einem weiteren Kapitel wurde noch das neu entwickelte Programmsystem namens CORA vorgestellt, mit dem die Reaktionscodierung durchgeführt wird. Bei der Konzeption wurde darauf Wert gelegt, daß die Information leicht mit anderen Programmsystemen, die im Arbeitskreis von Gasteiger entwickelt und gepflegt werden, ausgetauscht werden kann. Somit können die Ergebnisse der Reaktionsklassifizierung unter anderem in dem Syntheseplanungsprogramm WODCA und dem Reaktionsvorhersagesystem EROS eingesetzt werden.

Abschließend wurden in Kapitel 9 die Vor- und Nachteile der entwickelten Codierung- und Klassifizierungsmethode diskutiert, sowie Verbesserungsmöglichkeiten aufgezeigt. Sind die dort beschriebenen Nachteile und Einschränkungen eines Tages behoben, so kann man sich vorstellen, daß diese leistungsfähige Methode oder andere Verfahren zur Wissensextraktion aus Reaktionsdatenbanken unentbehrlich bei dem Erschließen von chemischer Information werden. Dann nämlich wenn Datenbanken zukünftig die Stelle einnehmen werden, die heute noch die Fachzeitschriften einnehmen: Die wichtigste Informationsquelle für Chemiker[122].

11 Literaturverzeichnis

[1] Barth, A.
„Online Databases in Chemistry“
in „Encyclopedia of Computational Chemistry“, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer III, H. F.; Schreiner, P. R. (Eds.), John Wiley & Sons, Chichester, UK, 1998, Vol. 3, S. 1968 – 1981.

[2] Gelernter, H.; Rose, J. R.; Chen, C.
„Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning“
J. Chem. Inf. Comput. Sci. **30** (1990) 492 – 504.

[3] Chen, L.; Gasteiger, J.; Rose, J. R.
„Extraction of Chemical Knowledge from Organic Reaction Data by Automatic Hierarchical Classification and Generalization“
in „Software Development in Chemistry 9“, R. Moll (Ed.), GDCh, Frankfurt/Main, 1994, S. 169 – 182.

[4] Rose, J. R.; Gasteiger, J.
„HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions“
J. Chem. Inf. Comput. Sci. **34** (1994) 74 – 90.

[5] Chen, L.; Gasteiger, J.
„Organische Reaktionen mit Hilfe neuronaler Netze klassifiziert: Michael-Additionen, Friedel-Crafts-Alkylierungen durch Alkene und verwandte Reaktionen“
Angew. Chem. **108** (1996) 844 – 846;
„Organic Reactions Classified by Neural Networks: Michael Additions, Friedel-Crafts Alkylations by Alkenes, and Related Reactions“
Angew. Chem. Int. Ed. Engl. **35** (1996) 763 – 765.

[6] Chen, L.; Gasteiger, J.
„Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network“
J. Am. Chem. Soc. **119** (1997) 4033 – 4042.

[7] Laue, T.; Plagens, A.
„Namen- und Schlagwort-Reaktionen der Organischen Chemie“
B. G. Teubner, Stuttgart, 1994.

[8] Garagnani, E.; Bart, J. C. J.
„Organic Reaction Schemes and General Reaction-Matrix Types, III - A Quantitative Analysis“
Z. Naturforsch. **32b** (1977) 465 – 468.

[9] Die Theilheimer Reaktionsdatenbank wird von MDL Information Systems, Inc., San Leandro, CA, USA, vertrieben.

[10] A Guide to Reaction databases, MDL Information Systems, Inc., 1990, Abschnitt 3.

[11] Parlow, A.; Weiske, C.; Gasteiger, J.
„ChemInform - An Integrated Information System on Chemical Reactions“
J. Chem. Inf. Comput. Sci. **30** (1990) 400 – 402.

[12] Die ChemInform RX-Reaktionsdatenbank wird vom Fachinformationszentrum Chemie, Berlin produziert und von MDL Information Systems, Inc., San Leandro, CA, USA, vertrieben (siehe http://www.mdli.com und http://www.fiz-chemie.de).

[13] Die Einträge für die SPORE Reaktionsdatenbank werden ebenfalls vom Fachinformationszentrum Chemie, Berlin gesammelt und abstrahiert. Der Vertrieb unterliegt MDL Information Systems, Inc., San Leandro, CA, USA.

[14] Merrifield, R. B.
„Solid phase peptide synthesis. I. The synthesis of a tetrapeptide“
J. Am. Chem. Soc. **85** (1963) 2149 – 2154.

[15] Hicks, M. G.
„Surfing the Organic Chemistry Hyperdocument with CrossFire plus Reactions“
J. Chem. Inf. Comput. Sci. **37** (1997) 146 – 147.

[16] Trümbach, D.; Gasteiger, J.
„Interaktive biochemische Stoffwechselwege“
in: Hofestädt, R.; Pleißner, K.-P.; Stephanik, A. (Hrsg.), „Informationssysteme in der Biotechnologie“, Workshop der GI-FG 4.0.2 Informatik in den Biowissenschaften, 10.-11. Februar 2000, Magdeburg.

[17] Behnke, C.; Bargon, J.
„Computer-Assisted Topological Analysis and Completion of Chemical Reactions“
J. Chem. Inf. Comput. Sci. **30** (1990) 228 – 237.

[18] Borkent, J. H.; Oukes, F.; Noordik, J. H.
„Chemical Reaction Searching Compared in REACCS, SYNLIB, and ORAC“
J. Chem. Inf. Comput. Sci. **28** (1988) 148 – 150.

[19] Boiten, J.-W.; Ott, M. A.; Noordik, J. H.
„Automated Overlap Analysis of Reaction Databases“
J. Chem. Inf. Comput. Sci. **35** (1995) 115 – 120.

[20] Hendrickson, J. B.; Zhang, L.
„Duplications among Reaction Databases“
J. Chem. Inf. Comput. Sci. **40** (2000) 380 – 383.

[21] Spitzer, M.
„Geist im Netz: Modelle für Lernen, Denken und Handeln“
Spektrum Akademischer Verlag, Heidelberg, 2000, S. 3 – 7.

[22] tom Dieck, S.; Gundelfinger, E. D.
„Chemische Synapsen des Zentralnervensystems“
Chem. i. u. Zeit **34** (2000) 140 – 148.

[23] Koolman, J.; Röhm, K. H.
„Taschenatlas der Biochemie“
2. Auflage, Georg Thieme, Stuttgart, 1998, S. 332.

[24] Brause, R.
„Neuronale Netze“
2. Auflage, B. G. Teubner, Stuttgart, 1995, S. 33 – 37.

[25] Ritter, H.; Martinetz, T.; Schulten, K.
„Neuronale Netze: Eine Einführung in die Neuroinformatik selbstorganisierender Netzwerke“
Addision-Wesley, Bonn, 1990.

[26] McCulloch, W. S.; Pitts, W.
„A Logical Calculus of the Ideas Immanent in Nervous Activity“
Bulletin of Mathematical Biophysics **5** (1943) 115 – 133.

[27] Rosenblatt, F.
„The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain“
Psychological Review **65** (1958) 386 – 408.

[28] Zupan, J.; Gasteiger, J.
„Neural Networks in Chemistry and Drug Design“
2nd Edition, Wiley-VCH, Weinheim, 1999.

[29] Zupan, J.; Novic, M.; Ruisanchez, I.
„Kohonen and counterpropagation artificial neural networks in analytical chemistry“
Chemom. Intell. Lab. Syst. **38** (1997) 1 – 23.

[30] Selzer, P.
„Vorhersage von Infrarotspektren mittels neuronaler Netze zur Identifikation organischer Verbindungen“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1998.

[31] Novic, M.; Zupan, J.
„Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Network“
J. Chem. Inf. Comput. Sci. **35** (1995) 454 – 466.

[32] Majcen, N.; Xavier R., F.; Zupan, J.
„Linear and non-linear multivariate analysis in the quality control of industrial titanium dioxide white pigment“
Anal. Chim. Acta **348** (1997) 87 – 100.

[33] Schuur, J. H.
„Die Codierung der 3D-Struktur von Molekülen und ihre Anwendung zur Simulation von IR-Spektren und für QSAR-Untersuchungen“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1998.

[34] Simon, V.; Gasteiger, J.; Zupan, J.
„A combined application of two different neural network types for the prediction of chemical reactivity“
J. Am. Chem. Soc. **115** (1993) 9148 – 9159.

[35] Gasteiger, J.; Li, X.; Rudolph, C.; Sadowski, J.; Zupan, J.
„Representation of Molecular Electrostatic Potentials by Topological Feature Maps“
J. Am. Chem. Soc. **116** (1994) 4608 – 4620.

[36] Wessel, M. D.; Jurs, P. C.
„Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure“
J. Chem. Inf. Comput. Sci. **35** (1995) 841 – 50.

[37] Hall, L. H.; Story, C. T.
„Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks“
J. Chem. Inf. Comput. Sci. **36** (1996) 1004 – 1014.

[38] Cherqaoui, D.; Villemin, D.
„Use of a Neural Network to determine the Boiling Point of Alkanes“
J. Chem. Soc. Faraday Trans. **90** (1994) 97 – 102.

[39] Schimke, G.; Clark, T.
„Estimation of the Partition Coefficients of Organic Compounds - A Neural Network Study in Combination with Semiempirical AM1 MO-Calculations“
J. Mol. Graph. **11** (1993) 65 – 66.

[40] Grunenberg, J.; Herges, R.
„Prediction of Chromatographic Retention Values (R_M) and Partition Coefficients (log P_{oct}) Using a Combination of Semiempirical Self-Consistent Reaction Field Calculations and Neural Networks“
J. Chem. Inf. Comput. Sci. **35** (1995) 905 – 911.

[41] Vracko, M.; Novic, M.; Zupan, J.
„Study of structure-toxicity relationship by a counterpropagation neural network“ .
Anal. Chim. Acta **384** (1999) 319 – 332.

[42] Brinn, M. W.; Payne, M. P.; Walsh, P. T.
„Neural Network Prediction of Mutagenicity Using Structure-Property Relationships“
Chem. Eng. Res. Des. **71** (1993) 337 – 339.

[43] Villemin, D.; Cherqaoui, D.; Cense, J. M.
„Neural Networks Studies: Quantitative Structure-Activity Relationship of Mutagenic Aromatic Nitro Compounds“
J. Chim. Phys. **90** (1993) 1505 – 1519.

[44] Gasteiger, J.; Zupan, J.;
„Neuronale Netze in der Chemie“
Angew. Chem. **105** (1993) 510 – 536;
„Neural Networks in Chemistry“,
Angew. Chem. Int. Ed. Engl. **32** (1993) 503 – 527.

[45] Kohonen, T.
„Self-organized Formation of Topologically Correct Feature Maps“
Biol. Cybern. **43** (1982) 59 – 69.

[46] Kohonen, T.
„Self-Organization and Associative Memory“
3rd Edition; Springer Verlag, Berlin, 1989.

[47] Kohonen, T.
„Self-Organizing Maps“
Huang, T. S.; Kohonen, T.; Schröder, M. R. (Eds.), Springer Verlag, Berlin, 1995.

[48] Otto, M.
„Chemometrie - Statistik und Computereinsatz in der Analytik“
VCH, Weinheim, 1997, S. 135 – 196.

[49] Gasteiger, J.; Marsili, M.
„A New Model for Calculating Atomic Charges in Molecules“
Tetrahedron Lett. **34** (1978) 3181 – 3184.

[50] Gasteiger, J.; Marsili, M.
„Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges“
Tetrahedron **36** (1980) 3219 – 3228.

[51] Guillen, M. D.; Gasteiger, J.
„Extension of the Method of Iterative Partial Equalization of Orbital Electronegativity to Small Ring Systems“
Tetrahedron **39** (1983) 1331 – 1335.

[52] Mortier, W. J.; Van Genechten, K.; Gasteiger, J.
„Electronegativity Equalization: Application and Parametrization“
J. Amer. Chem. Soc. **107** (1985) 829 – 835.

[53] Hammarström, L. G.; Liljefors, T.; Gasteiger, J.
„Electrostatic Interactions in Molecular Mechanics (MM2) Calculations via PEOE Partial Charges. I. Haloalkanes“
J. Comput. Chem. **9** (1988) 424 – 440.

[54] Marsili, M.; Gasteiger, J.
„Pi-Charge Distributions from Molecular Topology and Pi-Orbital Electronegativity“
Croat. Chem. Acta **53** (1980) 601 – 614.

[55] Gasteiger, J.; Saller, H.
„Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomeriekonzeptes“
Angew. Chem. **97** (1985) 699 – 701;
„Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept“
Angew. Chem. Int. Ed. Engl. **24** (1985) 687 – 689.

[56] Kang, Y. K.; Jhon, M. S.
„Additivity of Atomic Static Polarizabilities and Dispersion Coefficients“
Theor. Chim. Acta **61** (1982) 41 – 48.

[57] Gasteiger, J.; Hutchings, M. G.
„Quantification of Effective Polarisability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation“
J. Chem. Soc. Perkin **2** (1984) 559 – 564.

[58] Ihlenfeldt, W.-D.; Gasteiger, J.
„Hash Codes for the Identification and Classification of Molecular Structure Elements“
J. Comput. Chem. **15** (1994) 793 – 813.

[59] Information zu der Linearnotation SMILES ist unter der URL
http://www.daylight.com/dayhtml/smiles/index.html abrufbar.

[60] Weininger, D.
„SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules“
J. Chem. Inf. Comput. Sci. **28** (1988) 31 – 36.

[61] Weininger, D.; Weininger, A.; Weininger, J. L.
„SMILES. 2. Algorithm for Generation of Unique SMILES Notation“
J. Chem. Inf. Comput. Sci. **29** (1989) 97 – 101.

[62] Aylward, G. H.; Findlay, T. J. V.
„Datensammlung Chemie in SI -Einheiten“
2. Auflage, VCH, Weinheim, 1986, S. 127.

[63] Moreau, G.; Turpin, C.
„Use of Similarity Analysis to Reduce Large Molecular Libraries to Smaller Sets of Representative Molecules“
Analusis **24** (1996) M17 – M22.

[64] Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J.
„Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists“
J. Chem. Inf. Comput. Sci. **36** (1996) 1205 – 1213.

[65] Bogdanov, B.; Zdravkovski, Z.; Hristovski, K.
„Reaction Index: Name Reactions in Organic Chemistry“
http://www.pmf.ukim.edu.mk/PMF/Chemistry/reactions/rindex.htm

[66] Corey, E. J.
„General Methods for the Construction of Complex Molecules“
Pure Appl. Chem. **14** (1967) 19 – 37.

[67] Corey, E. J.; Wipke, W. T.
„Computer-Assisted Design of Complex Organic Syntheses“
Science **166** (1969) 178 – 192.

[68] Corey, E. J.; Wipke, W. T.; Cramer III, R. D.; Howe, W. J.
„Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates“
J. Am. Chem. Soc. **94** (1972) 440 – 459.

[69] Wipke, W. T.; Dyott, T. M.
„Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry“
J. Am. Chem. Soc. **96** (1974) 4825 – 4834.

[70] Blair, J.; Gasteiger, J.; Gillespie, C.; Gillespie, P. D.; Ugi, I.
„CICLOPS - A Computer Program for the Design of Synthesis on the Basis of a Mathematical Model“
in: Wipke, W. T.; Heller, S.; Feldmann, R.; Hyde, E. (Hrsg.), „Computer Representation and Manipulation of Chemical Information“, J. Wiley, New York, 1974, S. 129 – 145.

[71] Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K.
„A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design“
Topics Curr. Chem. **137** (1987) 19 – 73.

[72] Gelernter, H.; Sridharan, N. S.; Hart, A. J.; Yen, S. C.; Fowler, F. W.; Shue, H. J.
„The Discovery of Organic Synthetic Routes by Computer“
Top. Curr. Chem. **41** (1973) 113 – 150.

[73] Hendrickson, J. B.; Grier, D. L.; Toczko, A. G.
„A Logic-Based Program for Synthesis Design“
J. Am. Chem. Soc. **107** (1985) 5228 – 5238.

[74] Satoh, K.; Azuma, S.; Satoh, H.; Funatsu, K.
„Development of a Program for Construction of a Starting Material Library for AIPHOS“
Journal of Chemical Software **4** (1997) 101 – 107.
http://cssjweb.chem.eng.himeji-tech.ac.jp/jcs/v4n3/a3/text.html

[75] Ihlenfeldt, W.-D.; Gasteiger, J.
„Computergestützte Planung organisch-chemischer Synthesen: Die zweite Programmgeneration“
Angew. Chem. **107** (1995) 2807 – 2829;
„Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs“
Angew. Chem. Int. Ed. Engl. **34** (1995) 2613 – 2633.

[76] Gasteiger, J.; Ihlenfeldt, W.-D.; Röse, P.
„A collection of computer methods for the synthesis design and reaction prediction“
Recl. Trav. Chim. Pays-Bas **111** (1992) 270 – 290.

[77] M. Pförtner,
„Entwicklung von Methoden zur Bestimmung strategischer Bindungen in der computergestützten Syntheseplanung“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1999, S. 109 – 137.

[78] Braun, H.,
„Syntheseplanung. Strategische Werkzeuge für den Chemiker“
Chem. Ind. **40** (1988) 43 – 53.

[79] Nakayama, T.
„Computer-Assisted Knowledge Acquisition System for Synthesis Planning“
J. Chem. Inf. Comput. Sci. **31** (1991) 495 – 503.

[80] Krohn, K.; Wolf, U.
„Kurze Einführung in die Chemie der Heterocyclen“
B. G. Teubner, Stuttgart, 1994, S. 103 – 104.

[81] M. Pförtner,
„Entwicklung von Methoden zur Bestimmung strategischer Bindungen in der computergestützten Syntheseplanung“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1999, S. 201 – 206.

[82] Hall, S. S.; McEnroe, F. J.
„Alkylation-Reduction of Carbonyl Systems. IV. The Convenient and Selective Synthesis of Simple and Complex Aromatic Hydrocarbons by Phenylation-Reduction of Aldehyds and Ketons“
J. Org. Chem. **40** (1975) 271 – 275.

[83] Satoh, H.; Funatsu, K.
„Further Development of a Reaction Generator in the SOPHIA System for Organic Reaction Prediction. Knowledge-Guided Addition of Suitable Atoms and/or Atomic Groups to Product Skeleton“
J. Chem. Inf. Comput. Sci. **36** (1996) 173 – 184.

[84] Gasteiger, J.; Jochum, C.
„EROS - A Computer Program for Generating Sequences of Reactions“
Topics Curr. Chem. **74** (1978) 93 – 126.

[85] Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schulz, K.-P.; Herwig, A.
„Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis“
J. Chem. Inf. Comput. Sci. **40** (2000) 482 – 494.

[86] Bauerschmidt, S.
„Repräsentation von Molekülstrukturen zur computergestützten Behandlung chemischer Reaktionen“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1997.

[87] Höllering, R.
„Simulation von Massenspektren und Entwicklung eines Systems zur Reaktionsvorhersage“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1999.

[88] Gasteiger, J.; Hondelmann, U.; Röse, P.; Witzenbichler, W.
„Computer Assisted Prediction of the Degradation of Chemicals: Hydrolysis of Amides and Benzoylphenylureas“
J. Chem. Soc. Perkin Trans. 2 (1995) 193 – 204.

[89] Bhattacharya, B. K.; Eirich, F. R.
„Synthesis and Mass Spectral Studies of 1-[5-Chloro-1-substituted-2(1H)-pyrazin-2-on-3-yl]-5-aryl-3-methylpyrazoles“
J. Heterocycl. Chem. **22** (1985) 229 – 234.

[90] Axenrod, T.; Watnick, C. M.; Wieder, M. J.
„Structural assignments in isomeric pyrazoles based on $^{3}J(^{15}NH)$ coupling contants“
Org. Magn. Reson. **12** (1979) 476 – 480.

[91] Katritzky, A. R.; Rees, C. W.
„Comprehensive Heterocyclic Chemistry: The Structure, Reactions, Synthesis and Uses of Heterocyclic Compounds“, Volume 5, Part 4A
Pergamon Press, Oxford, 1984, S. 278.

[92] Selivanov, S. I.; Golodova, K. G.; Abbasov, Ya. A.; Ershov, B. A.;
„Investigation of the Mechanisms of Formation of Heterocycles by NMR Spectroscopy. III. The Effect of Electronic Factors on the Kinetics of the Dehydration of Dihydroxypyrazolidines and Hydroxypyrazolines - The Intermediates in the Reaction of Hydrazine with 1,3-Diketones“
Zh. Org. Chem. **20** (1984) 1494 – 1497.

[93] Katritzky, A. R.; Ostercamp, D. L., Yousaf, T. I.
„The Mechanisms of Heterocyclic Ring Closures“
Tetrahedron **43** (1987) 5171 – 5186.

[94] Davies, D. T.
„Aromatische Heterocyclen“
VCH, Weinheim, 1995, S. 89

[95] Mayr, H.; Patz, M.
„Nucleophilie- und Elektrophilie-Skalen als Ordnungsprinzipien polarer organischer und metallorganischer Reaktionen“
Angew. Chem. **106** (1994) 990 – 1010;
„Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions“
Angew. Chem. Int. Ed. Engl. **33** (1994) 938 – 957.

[96] Nicolai, E.; Cure, G.; Goyard, J.; Kirchner, M.; Teulon, J.-M.; et al.
„Synthesis and Angiotensin II Receptor Antagonist Activity of C-Linked Pyrazole Derivatives“
Chem. Pharm. Bull. **42** (1994) 1617 – 1630.

[97] Joshi, K. C.; Jain, R.; Dandia, A.; Sharma, K.
„Investigation of the Reactions of 2-Hydrazinobenzimidazoles with β-Diketones: Synthesis of 2-(3,5-Disubstituted-1*H*-pyrazol-1-yl-)benzimidazoles“
J. Heterocycl. Chem. **25** (1988) 1641 – 1643.

[98] Warr, W. A.
„Combinatorial Chemistry“
in „Encyclopedia of Computational Chemistry“, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer III, H. F.; Schreiner, P. R. (Eds.), John Wiley & Sons, Chichester, UK, 1998, Vol. 1, S. 407 – 417.

[99] Balkenhohl, F.; von dem Bussche-Hünnefeld, C.; Lansky, A.; Zechel, C.
„Kombinatorische Synthese niedermolekularer organischer Verbindungen“
Angew. Chem. **108** (1996) 2436 – 2488;
„Combinatorial Synthesis of Small Organic Molecules“
Angew. Chem. Int. Ed. Engl. **35** (1996) 2288 – 2337.

[100] Jung, G.; Beck-Sickinger, A. G.
„Methoden der multiplen Peptidsynthese und ihre Anwendungen“
Angew. Chem. **104** (1992) 375 – 391;
„Multiple Peptide Synthesis Methods and Their Applications“
Angew. Chem. Int. Ed. Engl. **31** (1992) 367 – 383.

[101] Booth, S. E.; Dreef-Tromp, C. M.; Hermkens, P. H. H.; de Man, J. A. P. A.; Ottenheijm, H. C. J.
„Survey of Solid-Phase Organic Reactions“
in „Combinatorial Chemistry - Synthesis, Analysis, Screening“
Jung, G. (Ed.), Wiley-VCH, Weinheim, 1999, S. 35 – 76.

[102] Noe, F. F.; Fowden L.
„β-Pyrazol-1-ylalanine, an Amino Acid from Water-Melon Seeds (Citrullus vulgaris)“
Biochem. J. **77** (1960) 543 – 546.

[103] Stauffer, S. R.; Katzenellenbogen, J. A.
„Solid-Phase Synthesis of Tetrasubstituted Pyrazoles, Novel Ligands for the Estrogen Receptor“
J. Comb. Chem. **2** (2000) 318 – 329.

[104] Marzinzik, A. L.; Felder, E. R.
„Key Intermediates in Combinatorial Chemistry: Access to Various Heterocycles from α,β-Unsaturated Ketones on the Solid Phase“
J. Org. Chem. **63** (1998) 723 – 727.

[105] Marzinzik, A. L.; Felder, E. R.
„Solid Support Synthesis of Highly Functionalized Pyrazoles and Isoxazoles; Scaffolds for Molecular Diversity“
Tetrahedron Lett. **37** (1996) 1003 – 1006.

[106] Wagener, M.
„Selecting Templates for Combinatorial Libraries Based on Reaction Classification“
Organon, Oss, Niederlande, Vortrag am 07.09.2000 im Computer-Chemie-Centrum.

[107] Matter, H.; Rarey, M.
„Design and Diversity Analysis of Compound Libraries for Lead Discovery“
in „Combinatorial Chemistry - Synthesis, Analysis, Screening“
Jung, G. (Ed.), Wiley-VCH, Weinheim, 1999, S. 409 – 439.

[108] Sadowski, J.; Wagener, M.; Gasteiger, J.
„Bewertung der Ähnlichkeit und Vielfalt von Verbindungsbibliotheken mit räumlichen Autokorrelationsvektoren und neuronalen Netzen“
Angew. Chem. **107** (1995) 2892 – 2895;
„Assessing Similarity and Diversity of Combinatorial Libraries by Spatial Autocorrelation Functions and Neural Networks“
Angew. Chem. Int. Ed. Engl. **34** (1995) 2674 – 2677.

[109] R. Höllering
„Simulation von Massenspektren und Entwicklung eines Systems zur Reaktionsvorhersage“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1998, S. 180 – 182.

[110] Marzinzik, A. L.; Felder, E. R.
„Combinatorial Libraries on Rigid Scaffolds: Solid Phase Synthesis of Variably Substituted Pyrazoles and Isoxazoles“
Molecules **2** (1997) 17 – 30.
http://mdpi.org/molecules/papers/jan97p5.pdf

[111] Eine aktuelle Version der unter dem Namen Tcl/Tk bekanntgewordenen Programmiersprache Scriptics kann unter http://dev.scriptics.com/software/tcltk/download82.html heruntergeladen werden.
Weitere Information findet man unter http://www.ajubasolutions.com

[112] Ousterhout, J. K.
„Tcl and the Tk Toolkit“
Addison-Wesley, Reading, 1994.

[113] Welch, B. B.
„Practical Programming in Tcl and Tk“
Prentice Hall, London, 1995.

[114] Informationen zu CACTVS findet man unter:
http://www2.ccc.uni-erlangen.de/software/cactvs/

[115] Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S.
„Computation and Management of Chemical Properties in CACTVS: An extensible Networked Approach toward Modularity and Flexibility“
J. Chem. Inf. Comp. Sci. **34** (1994) 109 – 116.

[116] Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.
„Computational Algorithm Management in a Global Networked Context“
in „AIP-Conference Proceedings 330: E.C.C.C.1, Computational Chemistry“,
Bernardi, F.; Rivail, J. L. (Eds.), American Institute of Physics, Woodbury, NY, 1995, S. 520 – 525.

[117] Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.
„Data Flow Processing for Computational Chemistry Problems“
in „AIP-Conference Proceedings 330: E.C.C.C.1, Computational Chemistry“,
Bernardi, F.; Rivail, J. L. (Eds.), American Institute of Physics, Woodbury, NY, 1995, S. 514 – 519.

[118] Teckentrup, A.
„Einsatzmöglichkeiten selbstorganisierender neuronaler Netze in der Wirkstofforschung“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2000, S. 143 – 152.

[119] Sixt, S.
„Methoden zur Abschätzung umweltrelevanter physikochemischer und ökologischer Eigenschaften organischer Substanzen aus der Molekülstruktur“
Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1998.

[120] Reichardt, Ch.; Harbusch-Görnert, E.
„Erweiterung, Korrektur und Neudefinition der E_T-Lösungsmittelpolaritätsskala mit Hilfe eines lipophilen penta-*tert*-butylsubstituierten Pyridinium-*N*-phenolat-Betainfarbstoffes“
Liebigs Ann. Chem. (1983) 721 – 743.

[121] Aires-de-Sousa, J.; Gasteiger, J.
„A New Description of Molecular Chirality and its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions“
J. Chem. Inf. Comput. Sci., im Druck.

[122] Heller, S. R.
„Databases - The Journals of the 21st Century“
Internet J. Chemistry **1** (1998) 32
http://www.ijc.com/articles/1998v1/32/

A Anhang

Dieser Anhang enthält folgende Tabellen und Datensätze:

- Aufstellung der im World-Wide-Web abrufbaren zusätzlichen Information
- Aufstellung der Rechenzeiten zur Klassifizierung der Theilheimer Reaktionsdatenbank
- Tabellarische Aufstellung zueinander ähnlicher Reaktionstypen
- Ergebnis zur Validierung des Pyrazoldatensatzes nach Aufteilung in Trainings- und Testdatensatz
- Ausgangsverbindungen zum Aufbau der kombinatorischen Bibliothek I
- Tabellarische Aufstellung der vorhergesagten Regioisomere der kombinatorischen Bibliothek I
- Ausgangsverbindungen zum Aufbau der kombinatorischen Bibliothek II

A.1 Zusätzliche Information im World-Wide-Web

zu Kapitel 4.2

- Darstellung der klassifizierten Theilheimer Reaktionsdatenbank: http://www2.chemie.uni-erlangen.de/research/knowledge_acquisition/dataset/theilheimer/index.html

zu Kapitel 7.3

- Darstellung der trainierten Kohonen-Karte zur Vorhersage der Regioisomere: http://www2.chemie.uni-erlangen.de/research/knowledge_acquisition/dataset/pyrazole/index.html
- Darstellung aller Edukte des kombinatorischen Datensatzes I als Molekülstruktur mit hinterlegtem 3D-Molfile: http://www2.chemie.uni-erlangen.de/research/knowledge_acquisition/dataset/pyrazole/index.html
- Darstellung des projizierten Datensatzes aus 496 kombinatorisch erzeugten Reaktionen: http://www2.chemie.uni-erlangen.de/research/knowledge_acquisition/dataset/pyrazole/index.html
- Darstellung aller Edukte des kombinatorischen Datensatzes II als Molekülstruktur mit hinterlegtem 3D-Molfile: http://www2.chemie.uni-erlangen.de/research/knowledge_acquisition/dataset/pyrazole/index.html
- Darstellung des projizierten Datensatzes aus 10.560 kombinatorisch erzeugten Reaktionen: http://www2.chemie.uni-erlangen.de/research/knowledge_acquisition/dataset/pyrazole/index.html

A.2 Rechenzeiten

Vorgang	Computer	CPU-Zeit	Gesamtzeit
Extraktion der Reaktionen aus der MDL Datenbank	µ-VAX 3900	unbekannt	ca. 7 Tage
Codierung der Reaktionen	Pentium III, 600 MHz 256 MByte, Linux 2.2	3:26:37	3:55:13
Trainieren des Kohonen-Netzes (siehe Tab. 4-1)	Pentium III, 600 MHz 256 MByte, Linux 2.2	0:25:59	0:38:53
Projektion eines Testdaten-satzes (SPORE)	Pentium III, 600 MHz 256 MByte, Linux 2.2	0:00:02	0:00:03

A.3 Ähnliche Reaktionstypen

Reaktionstyp	Neuron	ähnliche Reaktionstypen
1,3-Addition	(36,75)	Michael-Addition
Aldol-Kondensation	(6,45)	Michael-Addition
intramolekulare Aldol-Kondensation	(70,13); (1,67)	Pummerer-Umlagerung, Wolff-Kishner-Minlon-Reduktion
Aldol-Addition	(68,2)	Reformatsky-Reaktion
Anschütz-Reaktion	(92,6)	Michael-Addition
Baeyer-Villiger-Oxidation	(70,57); (67,35); (68,35)	Dakin-Oxidation, Elbs-Reaktion, Meerwein-Ponndorf-Verley-Reduktion, Tischtschenko-Reaktion
Bamberger-Reaktion	(4,2)	Widman-Stoermer-Reaktion
Beckmann-Fragmentierung	(35,2); (28,64)	von Braun-Reaktion, Conrad-Limpach-Reaktion
Beckmann-Umlagerung	(64,30); (7,36); (33,38); (80,72)	Schmidt-Reaktion, Reformatsky-Reaktion, Hilbert-Johnson-Reaktion, Ritter-Reaktion, Curtius-Abbau
Birch-Reduktion	(14,9); (1,80)	Diels-Alder-Reaktion, Thio-Claisen-Umlagerung
Bischler-Napieralski-Reaktion	(66,47)	Friedel-Crafts-Reaktion
Bogert-Cook-Reaktion	(91,91)	Claisen-Umlagerung
Bouveault-Blanc-Reduktion	(83,74)	Meerwein-Ponndorf-Reduktion
Brown-Hydrierung	(63,2)	(anti-)Markownikoff-Hydrierung
Chapman-Umlagerung	(88,8)	Lehmstedt-Tanasescu-Reaktion
Claisen-Umlagerung	(8,8); (29,42); (90,88); (91,91)	Beckmann-Reaktion, Thio-Claisen-Umlagerung, Wagner-Meerwein-Umlagerung, Friedel-Crafts-Alkylierung, Claisen-Umlagerung, Bogert-Cook-Reaktion
Clemmensen-Reduktion	(68,81)	Leuckart-Wallach-Reaktion, Löffler-Freytag-Reaktion
Conrad-Limpach-Reaktion	(7,82)	Knorr-Anellierung
Criegee-Spaltung	(38,30)	Aldol-Kondensation
Curtius-Abbau	(79,73); (77,74)	Ritter-Reaktion, Skita's-Regel, Leuckart-Reaktion
Curtius-Synthese	(79,16)	Lossen-Umlagerung
Dakin-Oxidation	(70,57)	Baeyer-Villiger-Oxidation, Elbs-Reaktion
Dakin-West-Reaktion	(86,46)	Dieckmann-Cyclisierung
Darzens-Claisen-Reaktion	(34,68)	Wittig-Reaktion

Reaktionstyp	Neuron	ähnliche Reaktionstypen
Darzens-Synthese	(83,59)	Friedel-Crafts-Keton-Synthese
Dieckmann-Cyclisierung	(86,46); (30,74); (66,92)	Dakin-West-Reaktion, Michael-Addition, Wittig-Reaktion
Diels-Alder-Reaktion	(14,9); (13,10); (6,74); (4,79)	Birch-Reduktion, Michael-Addition, Hetero-Diels-Alder-Reaktion, Thio-Claisen-Umlagerung, 1,8-Addition
Doebner-Miller-Reaktion	(2,25)	Wittig-Reaktion, Friedländer-Synthese, Skraup-Synthese
zweifache Wagner-Meerwein-Umlagerung	(76,9)	Radzivanovsky-Synthese
Duff-Reaktion	(78,63)	Vilsmeier-Reaktion, Gattermann-Koch-Synthese, Reimer-Tiemann-Reaktion
Elbs-Reaktion	(70,57)	Baeyer-Villiger-Oxidation, Dakin-Oxidation
En-Reaktion	(5,32); (1,46)	Wagner-Umlagerung, [2,3]-sigmatrope Umlagerung
Favorskii-Umlagerung	(46,34); (13,70)	Horner-Reaktion, Wittig-Reaktion, Stobbe-Kondensation
Fischer-Indol-Ringsynthese	(20,63)	Nencki-Indol-Ringsynthese
Friedel-Crafts-Reaktion	(80,61); (81,60); (90,86); (90,88); (83,59); (66,47)	inverse Grignard-Reaktion, Fries-Umlagerung, Wurtz-Fittig-Synthese, Claisen-Umlagerung, Darzens-Reaktion, Bischler-Napieralski-Reaktion
Friedländer-Synthese	(1,1); (2,25)	Pictet-Gams-Isochinolin-Synthese, intramol.Wittig-Reaktion, Skraup-(Cohn)-Synthese, Doebner-Miller-Reaktion
Fries-Umlagerung	(81,60)	Friedel-Crafts-Reaktion
Gabriel-Synthese	(57,48); (78,73)	Hofmannscher Abbau, Schmidt-Reaktion
Gattermann-(Koch)-Reaktion	(78,63)	Vilsmeier-Reaktion, Reimer-Tiemann-Reaktion, Duff-Reaktion
Graebe-Ullmann-Reaktion	(54,60)	Stevens-Umlagerung, Scholl-Ringschlußreaktion
Grignard-Reaktion	(52,69)	Norrish-Typ-II-Reaktion, Grob-Fragmentierung, Hofmann-Reaktion
Grob-Fragmentierung	(52,69)	Norrish-Typ-II-Reaktion, Grignard-Reaktion, Hofmann-Reaktion
Hetero-Cope-Umlagerung	(1,26)	[4π+2π] Cycloaddition
Hetero-Diels-Alder-Reaktion	(6,74); (12,77); (4,78)	1,8-Addition, [4+2] Cycloaddition, Michael-Addition, [8+2] Cycloaddition
Hilbert-Johnson-Reaktion	(33,38)	Beckmann-Umlagerung, Ritter-Reaktion

Reaktionstyp	Neuron	ähnliche Reaktionstypen
Hofmann-Reaktion	(52,69)	Norrish-Typ-II-Reaktion, Grob-Fragmentierung, Grignard-Reaktion
Hofmannscher Abbau	(57,48); (68,71); (78,74)	Gabriel-Synthese, Smiles-Umlagerung, Schmidt-Reaktion, Leuckart-Reaktion, Stephen-Reduktion
Horner-Reaktion	(28,18); (46,34); (7,63)	Wittig-Reaktion, Favorskii-Umlagerung, Stobbe-Kondensation
Kendall-Reaktion	(35,90)	Oppenauer-Oxidation, Krohnke-Aldehyd-Synthese
Knoevenagel-Kondensation	(55,86)	intramolekulare Wittig-Reaktion
Knorr-Anellierung	(7,82)	Conrad-Limpach-Reaktion
Knorr-Pyrrol-Synthese	(18,40)	Michael-Addition
Kolbe-Elektrolyse	(70,68)	One-Step-Barbier-Grignard-Reaktion, Guerbet-Kondensation
Königs-Knorr-Reaktion	(76,78)	Oppenauer-Oxidation
Krohnke-Aldehyd-Synthese	(35,90)	Oppenauer-Oxidation, Kendall-Reaktion
Lehmstedt-Tanasescu-Reaktion	(88,8)	Chapman-Umlagerung
Leuckart-Reaktion	(79,73); (78,74)	Ritter-Reaktion, Curtius-Abbau, Skita's-Regel, Stephen-Reduktion, Hofmannscher Abbau
Leuckart-Wallach-Reaktion	(68,81)	Löffler-Freytag-Reaktion, Clemmensen-Reduktion
Löffler-Freytag-Reaktion	(68,81)	Clemmensen-Reduktion, Leuckart-Wallach-Reaktion
Lossen-Umlagerung	(79,16); (83,16)	Curtius-Umlagerung, Schmidt-Reaktion
Mannich-Reaktion	(81,3); (71,10); (60,7)	Pictet-Spengler-Ringschlußreaktion, Michael-Addition
Markownikoff-Hydrierung	(63,2)	Brown-Hydrierung, anti-Markownikoff-Hydrierung
Meerwein-Ponndorf-Verley-Reduktion	(83,74); (67,35)	Bouveault-Blanc-Reduktion, Baeyer-Villiger-Oxidation
Michael-Addition	(92,5); (60,7); (13,10); (82,11); (8,42); (6,45); (40,57); (30,74); (36,75); (18,40); (12,77)	Anschütz-Reaktion, Mannich-Reaktion, Diels-Alder-Reaktion, Vilsmeier-Reaktion, [2+2]-Cycloaddition, Aldol-Kondensation, Robinson-Anellierung, Dieckmann-Cyclisierung, 1,3-Addition, Knorr-Pyrrol-Synthese, Hetero-Diels-Alder-Cycloaddition
Nef-Reaktion	(29,91)	Oppenauer-Oxidation, Serini-Reaktion
Nencki-Indol-Ringsynthese	(20,63)	Fischer-Indol-Ringsynthese

Reaktionstyp	Neuron	ähnliche Reaktionstypen
Norrish-Typ-II-Reaktion	(52,69)	Grob-Fragmentierung, Grignard-Reaktion, Hofmann-Reaktion
One-Step-Barbier-Grignard-Reaktion	(70,68)	Kolbe-Elektrolyse, Guerbet-Kondensation
Oppenauer-Oxidation	(76,78); (29,91); (35,90)	Königs-Knorr-Reaktion, Nef-Reaktion, Serini-Reaktion, Krohnke-Aldehyd-Synthese, Kendall-Reaktion
Pictet-Gams-Isochinolin-Synthese	(1,1); (1,90)	Friedländer-Synthese, Richter-Reaktion
Pictet-Spengler-Ringschlußreaktion	(71,10); (81,3)	Mannich-Reaktion
Pinakol-Umlagerung	(38,22)	Wittig-Reaktion
Pomeranz-Fritsch-Reaktion	(8,90)	Smiles-Umlagerung
Pummerer-Umlagerung	(70,13)	intramolekulare Aldol-Kondensation
Radzivanovsky-Synthese	(76,9)	zweifache Wagner-Meerwein-Umlagerung
Reformatsky-Reaktion	(68,2); (60,75); (64,4); (7,36); (22,73); (54,78)	Aldol-Addition, Wittig-Reaktion, Beckmann-Umlagerung, Stobbe-Kondensation
Reichert-Nieuwland-Reaktion	(70,18)	Sommelet-Umlagerung, Wagner-Meerwein-Umlagerung
Reimer-Tiemann	(78,63)	Vilsmeier-Reaktion, Gattermann-(Koch)-Reaktion, Duff-Reaktion
Richter-Reaktion	(1,90)	Pictet-Gams-Isochinolin-Synthese
Ritter-Reaktion	(33,38); (77,74); (79,73)	Hilbert-Johnson-Reaktion, Beckmann-Umlagerung, Curtius-Umlagerung, Skita's-Regel, Leuckart-Reaktion
Robinson-Anellierung	(40,57)	Michael-Addition
Schmidt-Reaktion	(78,73); (64,30); (83,16); (68,71)	Gabriel-Synthese, Beckmann-Umlagerung, Lossen-Umlagerung, Smiles-Umlagerung, Hofmannscher Abbau
Scholl-Ringschlußreaktion	(54,60)	Graebe-Ullmann-Reaktion, Stevens-Umlagerung
Semmler-Wolff-Reaktion	(30,18)	Wittig-Reaktion
Serini-Reaktion	(29,91)	Nef-Reaktion, Oppenauer-Oxidation
Shapiro-Reaktion	(51,32)	Wittig-Reaktion
Skita's Regel	(79,73)	Ritter-Reaktion, Curtius-Umlagerung, Leuckart-Reaktion
Skraup-(Cohn)-Synthese	(2,25)	intramol. Wittig-Reaktion, Friedländer-Synthese, Doebner-Miller-Reaktion
Smiles-Umlagerung	(68,71); (8,90)	Hofmannscher Abbau, Schmidt-Reaktion, Pomeranz-Fritsch-Reaktion

Reaktionstyp	Neuron	ähnliche Reaktionstypen
Sommelet-(Hauser)-Umlagerung	(70,18); (27,60); (92,88)	Wagner-Meerwein-Umlagerung, Reichert-Nieuwland-Reaktion, [2,3]-sigmatrope Umlagerung, Claisen-Umlagerung
Stephen-Reduktion	(78,74)	Leuckart-Reaktion, Hofmannscher Abbau
Stevens-Umlagerung	(54,60); (52,17)	Graebe-Ullmann-Reaktion, Scholl-Ringschlußreaktion, inter. Michael-Addition
therm. Stevens-Umlagerung	(73,89)	Wawzonek-Umlagerung
Stobbe-Kondensation	(13,70); (22,73); (59,80); (7,63)	Favorskii-Ringverengung, Reformatsky-Reaktion, Wittig-Reaktion, Horner-Reaktion
Thio-Claisen-Umlagerung	(29,42); (4,79); (1,80)	Wagner-Meerwein-Umlagerung, Diels-Alder-Reaktion, Birch-Reduktion,
Tischtschenko-Reaktion	(68,35)	Baeyer-Villiger-Oxidation
Vilsmeier-Reaktion	(78,63); (82,11)	Gattermann-(Koch)-Reaktion, Reimer-Tiemann-Reaktion, Duff-Reaktion, Michael-Addition
vinyloge Dieckmann Cyclisierung	(48,73)	Tschugaeff-Reaktion
von Braun-Reaktion	(35,2)	Beckmann-Fragmentierung
Wagner-Umlagerung	(5,32)	En-Reaktion
Wagner-Meerwein-Umlagerung	(70,18); (29,42)	Sommelet-Umlagerung, Reichert-Nieuwland-Reaktion, (Thio-)Claisen-Umlagerung
Wawzonek-Umlagerung	(73,89)	therm. Stevens-Umlagerung
Widman-Stoermer-Reaktion	(4,2)	Bamberger-Reaktion
Wittig-Reaktion	(38,22); (55,78); (28,18); (30,18); (51,32); (46,34); (27,61); (34,68); (54,78); (55,78); (59,80); (66,92)	Pinakol-Umlagerung, Doebner-Reaktion, Horner-Reaktion, Semmler-Wolff-Reaktion, Shapiro-Reaktion, Favorskii-Umlagerung, Darzens-(Claisen-)Reaction, Reformatsky-Reaktion, Stobbe-Kondensation, Dieckmann-Cyclisierung
intramolekulare Wittig-Reaktion	(60,75); (55,86)	Knoevenagel-Kondensation, Friedländer-Synthese, Skraup-(Cohn)-Synthese, Doebner-Miller-Reaktion
Wolff-Kishner-Minlon-Reduktion	(1,67)	Aldol-Kondensation
Wurtz-Fittig-Synthese	(90,86)	Friedel-Crafts-Reaktion
[1,5]-sigmatrope Umlagerung	(32,48)	Pinakol-Umlagerung
[2+2]-Cycloaddition	(8,42)	Michael-Addition
[2,3]-sigmatrope Umlagerung	(1,46); (27,60)	En-Reaktion, Sommelet-(Hauser)-Umlagerung

Reaktionstyp	Neuron	ähnliche Reaktionstypen
[4+2] Cycloaddition	(12,77)	Hetero-Diels-Alder-Reaktion, Michael-Addition
[4π+2π] Cycloaddition	(1,26)	Hetero-Cope-Umlagerung
[8+2] Cycloaddition	(4,78)	Hetero-Diels-Alder-Reaktion

A.4 Ergebnis zur Validierung des Pyrazoldatensatzes

Projektion des Datensatzes DS2 in das mit DS1 trainierte Netz

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
3	1	2	(10,4)	317
6	1	1	(9,6)	21
7	1	1	(8,11)	
8	1	1	(13,7)	105, 106, 108, 229, 230, 235
11	1	1	(9,7)	243, 244
12	1	1	(10,5)	255
13	1	1	(10,7)	116, 269
14	1	1	(10,6)	17, 18
15	1	1	(11,6)	19, 67, 83, 125, 148, 149
16	1	1	(11,6)	19, 67, 83, 125, 148, 149
20	1	1	(9,6)	21
22	1	1	(10,6)	17, 18
23	1	1	(7,13)	32, 33, 41, 45, 46
24	1	1	(5,13)	27, 30
25	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
26	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
28	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
31	1	1	(5,13)	27, 30
36	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
37	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
40	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
42	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
44	1	1	(7,13)	32, 33, 41, 45, 46
47	1	1	(7,12)	29, 34, 35, 38, 39, 43, 352
50	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
51	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
52	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
53	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
57	1	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
58	1	-1	(12,9)	302, 615
62	1	1	(11,7)	85, 118, 143, 151, 152, 192, 248, 249
63	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
65	1	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
66	1	1	(10,3)	147, 283
68	1	1	(11,6)	19, 67, 83, 125, 148, 149
69	1	1	(11,7)	85, 118, 143, 151, 152, 192, 248, 249

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
71	1	1	(7,1)	73, 139, 140, 141, 203
72	1	1	(7,1)	73, 139, 140, 141, 203
75	1	1	(12,7)	226, 227, 228, 238, 239, 241
77	1	2	(13,10)	570
78	1	2	(6,6)	157, 391, 619
79	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
80	1	2	(10,9)	56, 64, 578, 580, 583
82	1	1	(10,3)	147, 283
84	1	1	(11,6)	19, 67, 83, 125, 148, 149
87	1	1	(7,11)	88, 136, 263
89	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
91	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
92	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
96	1	2	(13,8)	584, 585
97	1	2	(5,11)	410
99	1	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
100	1	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
102	1	1	(11,3)	101, 103
104	1	1	(11,3)	101, 103
107	1	1	(13,7)	105, 106, 108, 229, 230, 235
109	1	1	(13,7)	105, 106, 108, 229, 230, 235
110	1	1	(13,7)	105, 106, 108, 229, 230, 235
111	1	1	(13,7)	105, 106, 108, 229, 230, 235
113	1	1	(10,6)	17, 18
114	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
115	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
119	1	1	(11,7)	85, 118, 143, 151, 152, 192, 248, 249
120	1	2	(2,1)	261, 433, 434, 573, 574
121	1	2	(2,1)	261, 433, 434, 573, 574
123	1	1	(13,6)	124, 232, 560
126	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
127	1	2	(10,9)	56, 64, 578, 580, 583
129	1	1	(10,3)	147, 283
130	1	1	(11,6)	19, 67, 83, 125, 148, 149
131	1	1	(11,6)	19, 67, 83, 125, 148, 149
132	1	1	(11,7)	85, 118, 143, 151, 152, 192, 248, 249
142	1	1	(8,1)	74, 191, 282, 289, 293
144	1	1	(13,5)	137, 138, 222, 225
145	1	2	(10,9)	56, 64, 578, 580, 583

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
150	1	1	(6,9)	168
153	1	1	(9,6)	21
154	1	1	(10,12)	161, 212
155	1	-1	(9,13)	156, 471
158	1	-1	(9,13)	156, 471
159	1	1	(5,8)	194
160	1	1	(10,5)	255
163	1	2	(6,6)	157, 391, 619
164	1	1	(6,8)	
165	1	1	(9,5)	307, 308
166	1	-1	(9,13)	156, 471
167	1	-1	(9,13)	156, 471
170	1	1	(10,13)	133, 134
171	1	1	(10,13)	133, 134
173	1	1	(7,3)	172
174	1	1	(7,3)	172
175	1	1	(8,2)	176, 178, 179
181	1	2	(3,6)	
186	1	1	(7,10)	201, 221, 254
188	1	1	(7,1)	73, 139, 140, 141, 203
189	1	1	(7,1)	73, 139, 140, 141, 203
190	1	1	(7,1)	73, 139, 140, 141, 203
195	1	1	(11,13)	196
199	1	-1	(7,7)	162, 259, 512, 571
202	1	1	(7,10)	201, 221, 254
204	1	1	(7,1)	73, 139, 140, 141, 203
205	1	1	(7,1)	73, 139, 140, 141, 203
207	1	1	(12,7)	226, 227, 228, 238, 239, 241
208	1	1	(5,8)	194
209	1	1	(9,12)	2
210	1	1	(10,12)	161, 212
211	1	1	(10,12)	161, 212
215	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
216	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
217	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
218	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
219	1	1	(13,3)	48, 49, 54, 55, 213, 214, 220
223	1	1	(11,7)	85, 118, 143, 151, 152, 192, 248, 249
224	1	1	(11,7)	85, 118, 143, 151, 152, 192, 248, 249

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
231	1	1	(12,7)	226, 227, 228, 238, 239, 241
233	1	1	(12,7)	226, 227, 228, 238, 239, 241
234	1	1	(13,7)	105, 106, 108, 229, 230, 235
236	1	1	(12,7)	226, 227, 228, 238, 239, 241
240	1	1	(12,7)	226, 227, 228, 238, 239, 241
242	1	1	(9,7)	243, 244
246	1	2	(1,5)	245, 437, 451, 457, 538, 545, 589
247	1	2	(1,5)	245, 437, 451, 457, 538, 545, 589
250	1	1	(7,10)	201, 221, 254
251	1	1	(7,10)	201, 221, 254
252	1	1	(9,2)	253
257	1	2	(5,1)	390
258	1	-1	(7,7)	162, 259, 512, 571
260	1	2	(2,1)	261, 433, 434, 573, 574
262	1	2	(1,1)	575
264	1	1	(7,11)	88, 136, 263
265	1	1	(4,3)	185, 267, 268, 270
266	1	1	(4,3)	185, 267, 268, 270
271	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
272	1	1	(4,1)	86, 90, 93, 95, 256, 373, 407
273	1	1	(9,5)	307, 308
275	1	1	(8,11)	
277	1	1	(10,3)	147, 283
279	1	1	(10,3)	147, 283
281	1	1	(9,1)	280, 285, 286, 288, 294, 295
284	1	1	(10,3)	147, 283
287	1	1	(9,1)	280, 285, 286, 288, 294, 295
290	1	1	(9,1)	280, 285, 286, 288, 294, 295
291	1	1	(9,1)	280, 285, 286, 288, 294, 295
292	1	1	(8,1)	74, 191, 282, 289, 293
297	1	1	(6,3)	296, 299, 300
298	1	1	(6,1)	135, 206
301	1	2	(13,9)	60, 94, 376, 392, 405, 614
303	1	1	(13,5)	137, 138, 222, 225
305	1	1	(6,7)	
306	1	2	(6,6)	157, 391, 619
309	1	2	(6,6)	157, 391, 619
310	1	1	(10,7)	116, 269
312	1	1	(10,7)	116, 269

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
313	1	1	(10,7)	116, 269
314	2	2	(12,11)	519, 609, 611, 612, 613
316	2	1	(2,8)	4, 5
318	2	1	(2,8)	4, 5
319	2	2	(3,6)	
320	2	2	(10,11)	
321	2	2	(1,3)	420, 421, 424
322	2	2	(3,12)	323, 348, 350, 351, 356
326	2	2	(3,5)	429
327	2	2	(2,6)	330, 335, 426
328	2	2	(1,6)	380, 381, 396, 397, 438, 443, 462
329	2	2	(1,6)	380, 381, 396, 397, 438, 443, 462
331	2	2	(2,6)	330, 335, 426
332	2	2	(1,6)	380, 381, 396, 397, 438, 443, 462
333	2	2	(3,6)	
334	2	2	(3,6)	
337	2	2	(3,12)	323, 348, 350, 351, 356
338	2	2	(3,12)	323, 348, 350, 351, 356
339	2	2	(3,12)	323, 348, 350, 351, 356
340	2	2	(3,12)	323, 348, 350, 351, 356
341	2	2	(3,12)	323, 348, 350, 351, 356
342	2	2	(3,12)	323, 348, 350, 351, 356
343	2	2	(3,12)	323, 348, 350, 351, 356
344	2	2	(3,12)	323, 348, 350, 351, 356
347	2	2	(3,12)	323, 348, 350, 351, 356
349	2	2	(3,12)	323, 348, 350, 351, 356
353	2	2	(3,12)	323, 348, 350, 351, 356
355	2	2	(3,12)	323, 348, 350, 351, 356
358	2	2	(3,13)	336, 345, 346, 354, 357, 359
360	2	2	(3,12)	323, 348, 350, 351, 356
361	2	2	(1,13)	362, 363, 366, 367
364	2	2	(1,13)	362, 363, 366, 367
365	2	2	(1,13)	362, 363, 366, 367
368	2	2	(1,13)	362, 363, 366, 367
370	2	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
371	2	1	(4,1)	86, 90, 93, 95, 256, 373, 407
374	2	1	(10,8)	
377	2	2	(4,4)	117, 369, 393, 458
378	2	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
382	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
388	2	2	(2,4)	520, 546, 551, 553
395	2	2	(2,9)	379, 460
398	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
399	2	2	(12,10)	403, 404, 427, 428
401	2	2	(3,11)	400, 577
402	2	2	(12,10)	403, 404, 427, 428
406	2	2	(13,9)	60, 94, 376, 392, 405, 614
408	2	2	(12,10)	403, 404, 427, 428
409	2	1	(6,2)	1
411	2	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
412	2	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
418	2	2	(1,3)	420, 421, 424
419	2	2	(1,3)	420, 421, 424
422	2	2	(1,3)	420, 421, 424
423	2	2	(1,3)	420, 421, 424
425	2	1	(9,4)	112
431	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
432	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
436	2	2	(1,5)	245, 437, 451, 457, 538, 545, 589
439	2	2	(13,9)	60, 94, 376, 392, 405, 614
440	2	2	(4,4)	117, 369, 393, 458
441	2	1	(6,5)	76, 81, 98, 128, 146, 389, 394, 413, 459
442	2	2	(2,9)	379, 460
444	2	2	(1,6)	380, 381, 396, 397, 438, 443, 462
446	2	2	(13,2)	447
448	2	2	(11,11)	485
449	2	2	(3,11)	400, 577
450	2	2	(1,5)	245, 437, 451, 457, 538, 545, 589
452	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
453	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
454	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
455	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
456	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
461	2	2	(1,6)	380, 381, 396, 397, 438, 443, 462
463	2	2	(8,10)	435
464	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
466	2	2	(4,6)	
468	2	1	(8,5)	

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
469	2	1	(8,5)	
472	2	-1	(9,9)	278, 591
473	2	2	(2,7)	325, 568
475	2	-1	(9,9)	278, 591
479	2	-1	(9,13)	156, 471
480	2	-1	(9,13)	156, 471
481	2	2	(8,10)	435
482	2	1	(10,13)	133, 134
483	2	2	(13,2)	447
484	2	2	(13,2)	447
486	2	2	(11,11)	485
487	2	2	(11,11)	485
488	2	2	(13,13)	489, 491
492	2	2	(13,13)	489, 491
494	2	1	(9,6)	21
495	2	2	(4,11)	496
497	2	1	(9,12)	2
498	2	1	(4,3)	185, 267, 268, 270
499	2	2	(9,11)	563, 564
500	2	1	(4,13)	187, 197
502	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
503	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
504	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
505	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
507	2	1	(5,8)	194
509	2	2	(11,1)	467, 474, 508, 523, 525
510	2	1	(4,13)	187, 197
513	2	1	(9,7)	243, 244
514	2	2	(10,11)	
515	2	2	(9,11)	563, 564
521	2	1	(5,8)	194
522	2	-1	(9,3)	184, 315
527	2	2	(1,12)	526, 528, 531, 533
529	2	2	(1,12)	526, 528, 531, 533
530	2	2	(1,12)	526, 528, 531, 533
532	2	2	(1,12)	526, 528, 531, 533
534	2	2	(9,11)	563, 564
535	2	2	(1,5)	245, 437, 451, 457, 538, 545, 589
539	2	2	(2,4)	520, 546, 551, 553

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
540	2	2	(2,4)	520, 546, 551, 553
541	2	2	(2,4)	520, 546, 551, 553
543	2	2	(2,3)	542
544	2	2	(2,4)	520, 546, 551, 553
547	2	2	(2,3)	542
548	2	2	(2,3)	542
549	2	2	(2,4)	520, 546, 551, 553
550	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
552	2	2	(2,4)	520, 546, 551, 553
554	2	2	(2,4)	520, 546, 551, 553
555	2	2	(4,7)	556, 557
558	2	1	(13,6)	124, 232, 560
559	2	1	(13,6)	124, 232, 560
561	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
566	2	2	(12,13)	565, 594, 600
567	2	2	(9,11)	563, 564
569	2	2	(13,9)	60, 94, 376, 392, 405, 614
572	2	-1	(7,7)	162, 259, 512, 571
576	2	2	(3,11)	400, 577
579	2	2	(10,9)	56, 64, 578, 580, 583
581	2	2	(10,9)	56, 64, 578, 580, 583
582	2	2	(2,5)	237, 375, 445, 465, 536, 537, 562
588	2	2	(10,11)	
590	2	2	(2,9)	379, 460
593	2	2	(12,13)	565, 594, 600
598	2	2	(12,13)	565, 594, 600
599	2	2	(12,13)	565, 594, 600
601	2	2	(12,13)	565, 594, 600
602	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
604	2	2	(13,12)	603
605	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
606	2	2	(12,12)	384, 385, 386, 387, 501, 516, 517, 518, 595
607	2	2	(12,13)	565, 594, 600
608	2	2	(12,13)	565, 594, 600
610	2	2	(12,11)	519, 609, 611, 612, 613
616	2	1	(13,5)	137, 138, 222, 225
617	2	1	(7,6)	304
618	2	1	(6,7)	
620	2	2	(3,9)	

Reaktions-Nr. DS2	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
621	2	2	(3,9)	
622	2	2	(6,6)	157, 391, 619
624	2	2	(4,9)	587, 623, 625
626	2	2	(4,9)	587, 623, 625

Projektion des Datensatzes DS1 in das mit DS2 trainierte Netz

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
1	1	1	(2,7)	257, 297, 298
2	1	-1	(11,5)	209, 497
4	1	2	(8,2)	316, 318, 395, 442, 590
5	1	2	(8,2)	316, 318, 395, 442, 590
9	1	1	(8,9)	36, 37, 40, 47
10	1	1	(8,9)	36, 37, 40, 47
17	1	1	(5,7)	14, 22, 113
18	1	1	(5,7)	14, 22, 113
19	1	1	(8,6)	15, 16
21	1	1	(4,6)	6, 20, 494
27	1	1	(7,9)	24, 31
29	1	1	(9,9)	25, 26, 28, 42
30	1	1	(7,9)	24, 31
32	1	1	(9,9)	25, 26, 28, 42
33	1	1	(9,10)	23, 44
34	1	1	(8,9)	36, 37, 40, 47
35	1	1	(8,9)	36, 37, 40, 47
38	1	1	(8,9)	36, 37, 40, 47
39	1	1	(8,9)	36, 37, 40, 47
41	1	1	(9,10)	23, 44
43	1	1	(9,9)	25, 26, 28, 42
45	1	1	(9,10)	23, 44
46	1	1	(9,10)	23, 44
48	1	1	(13,13)	50, 51, 52, 53
49	1	1	(13,13)	50, 51, 52, 53
54	1	1	(13,13)	50, 51, 52, 53
55	1	1	(13,13)	50, 51, 52, 53
56	1	-1	(13,1)	80, 127, 145, 374, 579, 581
59	1	1	(8,5)	

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
60	1	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
61	1	2	(1,9)	265, 266, 377, 440, 498
64	1	-1	(13,1)	80, 127, 145, 374, 579, 581
67	1	1	(6,7)	68, 84, 130, 131
70	1	-1	(13,1)	80, 127, 145, 374, 579, 581
73	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
74	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
76	1	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
81	1	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
83	1	1	(6,7)	68, 84, 130, 131
85	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
86	1	1	(1,7)	63, 79, 89, 91, 92, 114, 126
88	1	1	(9,8)	87, 264
90	1	1	(1,7)	63, 79, 89, 91, 92, 114, 126
93	1	1	(1,7)	63, 79, 89, 91, 92, 114, 126
94	1	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
95	1	1	(1,7)	63, 79, 89, 91, 92, 114, 126
98	1	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
101	1	1	(8,4)	102, 104
103	1	1	(8,4)	102, 104
105	1	1	(6,4)	8, 107, 109, 110, 111, 234
106	1	1	(6,4)	8, 107, 109, 110, 111, 234
108	1	1	(6,4)	8, 107, 109, 110, 111, 234
112	1	2	(8,3)	
116	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
117	1	2	(1,9)	265, 266, 377, 440, 498
118	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
122	1	1	(4,4)	
124	1	-1	(7,6)	123, 144, 558, 559
125	1	1	(8,6)	15, 16
128	1	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
133	1	1	(13,7)	170, 171, 211
134	1	1	(13,7)	170, 171, 211
135	1	1	(7,13)	173, 174
136	1	1	(9,8)	87, 264
137	1	-1	(7,6)	123, 144, 558, 559
138	1	-1	(7,6)	123, 144, 558, 559
139	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
140	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
141	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
143	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
146	1	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
147	1	1	(9,5)	3, 66, 82, 129, 277, 279, 284
148	1	1	(6,7)	68, 84, 130, 131
149	1	1	(6,7)	68, 84, 130, 131
151	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
152	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
156	1	1	(11,9)	155
157	1	1	(3,5)	163
161	1	1	(13,6)	154, 195, 210
162	1	1	(1,5)	199, 258, 572
168	1	1	(2,11)	150
169	1	2	(11,13)	446, 483, 484
172	1	1	(7,13)	173, 174
176	1	1	(9,12)	175, 252
177	1	0	(6,12)	
178	1	1	(9,12)	175, 252
179	1	1	(9,12)	175, 252
180	1	2	(10,11)	468, 469
182	1	1	(7,11)	
183	1	1	(7,11)	
184	1	2	(10,5)	522
185	1	2	(1,9)	265, 266, 377, 440, 498
187	1	2	(6,11)	500, 510
191	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
192	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
193	1	1	(8,5)	
194	1	-1	(1,11)	159, 208, 507, 521
196	1	1	(13,6)	154, 195, 210
197	1	2	(6,11)	500, 510
198	1	2	(1,3)	555, 624
200	1	2	(1,3)	555, 624
201	1	1	(7,12)	
203	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
206	1	1	(2,7)	257, 297, 298
212	1	1	(13,6)	154, 195, 210
213	1	1	(13,12)	215, 216, 217, 218, 219
214	1	1	(13,12)	215, 216, 217, 218, 219

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
220	1	1	(13,12)	215, 216, 217, 218, 219
221	1	1	(6,13)	186, 202, 250, 251
222	1	-1	(7,6)	123, 144, 558, 559
225	1	-1	(7,6)	123, 144, 558, 559
226	1	1	(6,5)	75, 231, 233, 240
227	1	1	(6,5)	75, 231, 233, 240
228	1	1	(6,5)	75, 231, 233, 240
229	1	1	(6,4)	8, 107, 109, 110, 111, 234
230	1	1	(6,4)	8, 107, 109, 110, 111, 234
232	1	-1	(7,6)	123, 144, 558, 559
235	1	1	(6,4)	8, 107, 109, 110, 111, 234
237	1	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561
238	1	1	(6,5)	75, 231, 233, 240
239	1	1	(6,5)	75, 231, 233, 240
241	1	1	(6,5)	75, 231, 233, 240
243	1	1	(4,7)	11, 242, 513
244	1	1	(4,7)	11, 242, 513
245	1	2	(4,1)	246, 247, 436, 450, 535
248	1	1	(6,5)	75, 231, 233, 240
249	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
253	1	1	(9,12)	175, 252
254	1	1	(6,13)	186, 202, 250, 251
255	1	1	(5,6)	12, 160
256	1	1	(1,7)	63, 79, 89, 91, 92, 114, 126
259	1	1	(1,5)	199, 258, 572
261	1	1	(2,13)	120, 121, 260
263	1	1	(9,8)	87, 264
267	1	2	(1,9)	265, 266, 377, 440, 498
268	1	2	(1,9)	265, 266, 377, 440, 498
269	1	1	(6,6)	13, 62, 69, 119, 132, 223, 224
270	1	2	(1,9)	265, 266, 377, 440, 498
274	1	1	(6,8)	310, 312, 313
276	1	1	(7,7)	
278	1	2	(12,1)	475
280	1	1	(8,13)	281, 287
282	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
283	1	1	(9,5)	3, 66, 82, 129, 277, 279, 284
285	1	1	(8,13)	281, 287
286	1	1	(8,13)	281, 287

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
288	1	1	(8,13)	281, 287
289	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
293	1	1	(9,13)	71, 72, 142, 188, 189, 190, 204, 205, 292
294	1	1	(8,13)	281, 287
295	1	1	(8,13)	281, 287
296	1	1	(2,7)	257, 297, 298
299	1	1	(2,7)	257, 297, 298
300	1	1	(2,7)	257, 297, 298
302	1	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
304	1	1	(2,5)	78, 306, 309, 617, 622
307	1	1	(9,6)	165
308	1	1	(9,6)	165
311	1	1	(6,8)	310, 312, 313
315	2	2	(10,5)	522
317	2	1	(9,5)	3, 66, 82, 129, 277, 279, 284
323	2	2	(4,12)	322, 347, 349, 353, 360
324	2	2	(1,2)	181, 319, 333, 334, 473
325	2	2	(1,1)	327, 331
330	2	2	(1,1)	327, 331
335	2	2	(1,1)	327, 331
336	2	2	(4,11)	338, 339, 341, 342, 355, 358
345	2	2	(4,11)	338, 339, 341, 342, 355, 358
346	2	2	(4,11)	338, 339, 341, 342, 355, 358
348	2	2	(4,12)	322, 347, 349, 353, 360
350	2	2	(4,12)	322, 347, 349, 353, 360
351	2	2	(4,12)	322, 347, 349, 353, 360
352	2	1	(8,9)	36, 37, 40, 47
354	2	2	(4,11)	338, 339, 341, 342, 355, 358
356	2	2	(4,11)	338, 339, 341, 342, 355, 358
357	2	2	(4,11)	338, 339, 341, 342, 355, 358
359	2	2	(4,11)	338, 339, 341, 342, 355, 358
362	2	2	(13,10)	361, 364, 365, 368
363	2	2	(13,10)	361, 364, 365, 368
366	2	2	(13,10)	361, 364, 365, 368
367	2	2	(13,10)	361, 364, 365, 368
369	2	2	(1,9)	265, 266, 377, 440, 498
372	2	2	(3,1)	444, 461
373	2	1	(1,7)	63, 79, 89, 91, 92, 114, 126
375	2	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
376	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
379	2	2	(8,2)	316, 318, 395, 442, 590
380	2	2	(3,1)	444, 461
381	2	2	(3,1)	444, 461
383	2	2	(1,9)	265, 266, 377, 440, 498
384	2	2	(13,3)	452, 453, 454, 502, 503
385	2	2	(13,3)	452, 453, 454, 502, 503
386	2	2	(13,3)	452, 453, 454, 502, 503
387	2	2	(12,3)	455, 504, 602, 605, 606
389	2	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
390	2	1	(2,7)	257, 297, 298
391	2	1	(2,5)	78, 306, 309, 617, 622
392	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
393	2	2	(1,9)	265, 266, 377, 440, 498
394	2	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
396	2	2	(3,1)	444, 461
397	2	2	(3,1)	444, 461
400	2	2	(3,10)	401, 449, 576
403	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
404	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
405	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
407	2	1	(1,7)	63, 79, 89, 91, 92, 114, 126
410	2	1	(6,10)	97
413	2	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
414	2	2	(8,2)	316, 318, 395, 442, 590
415	2	2	(8,2)	316, 318, 395, 442, 590
416	2	2	(8,2)	316, 318, 395, 442, 590
417	2	2	(8,2)	316, 318, 395, 442, 590
420	2	2	(6,1)	418, 419, 422, 423
421	2	2	(6,1)	418, 419, 422, 423
424	2	2	(6,1)	418, 419, 422, 423
426	2	2	(1,1)	327, 331
427	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
428	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
429	2	2	(2,2)	326, 582
430	2	-1	(13,1)	80, 127, 145, 374, 579, 581
433	2	1	(2,13)	120, 121, 260
434	2	1	(2,13)	120, 121, 260
435	2	2	(6,9)	463, 481

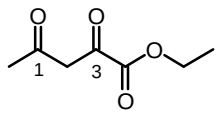
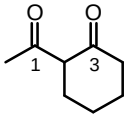
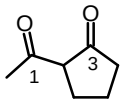
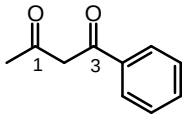
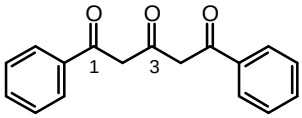
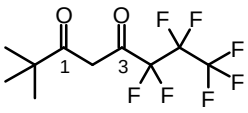
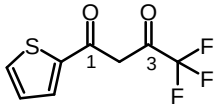
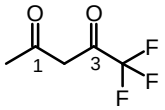
Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
437	2	2	(4,1)	246, 247, 436, 450, 535
438	2	2	(2,1)	328, 329, 332
443	2	2	(3,1)	444, 461
445	2	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561
447	2	2	(11,13)	446, 483, 484
451	2	2	(4,1)	246, 247, 436, 450, 535
457	2	2	(4,1)	246, 247, 436, 450, 535
458	2	2	(1,9)	265, 266, 377, 440, 498
459	2	2	(3,4)	57, 65, 99, 100, 370, 378, 411, 412, 441
460	2	2	(8,2)	316, 318, 395, 442, 590
462	2	2	(3,1)	444, 461
465	2	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561
467	2	2	(10,12)	509
470	2	1	(2,5)	78, 306, 309, 617, 622
471	2	1	(11,10)	158, 166, 167, 479, 480
474	2	2	(10,12)	509
476	2	1	(2,5)	78, 306, 309, 617, 622
477	2	1	(5,8)	153
478	2	2	(8,1)	620, 621
485	2	2	(11,3)	448, 486, 487, 610
489	2	2	(12,4)	488, 492
490	2	2	(11,2)	472
491	2	2	(12,4)	488, 492
493	2	2	(10,11)	468, 469
496	2	2	(5,11)	495
501	2	2	(13,3)	452, 453, 454, 502, 503
506	2	2	(3,1)	444, 461
508	2	2	(10,12)	509
511	2	1	(3,5)	163
512	2	1	(1,5)	199, 258, 572
516	2	2	(13,3)	452, 453, 454, 502, 503
517	2	2	(13,3)	452, 453, 454, 502, 503
518	2	2	(13,3)	452, 453, 454, 502, 503
519	2	2	(11,3)	448, 486, 487, 610
520	2	2	(4,2)	388, 539, 540, 541, 544, 549, 552, 554
523	2	2	(10,12)	509
524	2	2	(11,12)	
525	2	2	(10,12)	509
526	2	2	(13,9)	527, 529, 530, 532

Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
528	2	2	(13,9)	527, 529, 530, 532
531	2	2	(13,9)	527, 529, 530, 532
533	2	2	(13,9)	527, 529, 530, 532
536	2	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561
537	2	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561
538	2	2	(4,1)	246, 247, 436, 450, 535
542	2	2	(6,2)	543, 547, 548
545	2	2	(4,1)	246, 247, 436, 450, 535
546	2	2	(4,2)	388, 539, 540, 541, 544, 549, 552, 554
551	2	2	(4,2)	388, 539, 540, 541, 544, 549, 552, 554
553	2	2	(4,2)	388, 539, 540, 541, 544, 549, 552, 554
556	2	2	(1,3)	555, 624
557	2	2	(1,3)	555, 624
560	2	-1	(7,6)	123, 144, 558, 559
562	2	2	(3,2)	382, 398, 431, 432, 456, 464, 505, 550, 561
563	2	2	(11,8)	499, 514, 515, 534, 567
564	2	2	(11,8)	499, 514, 515, 534, 567
565	2	2	(13,4)	566, 593, 598, 599, 601, 607, 608
568	2	2	(1,2)	181, 319, 333, 334, 473
570	2	1	(9,2)	77
571	2	1	(1,5)	199, 258, 572
573	2	1	(2,13)	120, 121, 260
574	2	1	(2,13)	120, 121, 260
575	2	1	(1,13)	262
577	2	2	(3,10)	401, 449, 576
578	2	-1	(13,1)	80, 127, 145, 374, 579, 581
580	2	-1	(13,1)	80, 127, 145, 374, 579, 581
583	2	-1	(13,1)	80, 127, 145, 374, 579, 581
584	2	-1	(10,3)	96, 314
585	2	-1	(10,3)	96, 314
586	2	2	(3,10)	401, 449, 576
587	2	2	(1,3)	555, 624
589	2	2	(4,1)	246, 247, 436, 450, 535
591	2	2	(12,1)	475
592	2	2	(8,2)	316, 318, 395, 442, 590
594	2	2	(13,4)	566, 593, 598, 599, 601, 607, 608
595	2	2	(12,3)	455, 504, 602, 605, 606
596	2	2	(8,2)	316, 318, 395, 442, 590
597	2	2	(8,2)	316, 318, 395, 442, 590

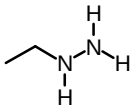
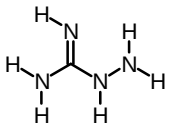
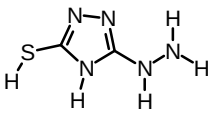
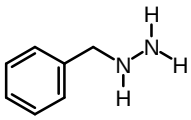
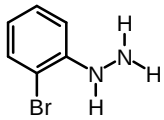
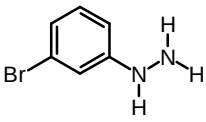
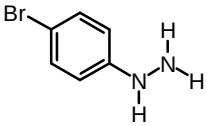
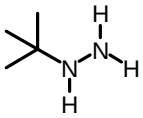
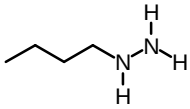
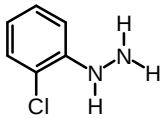
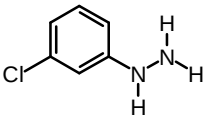
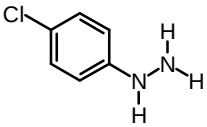
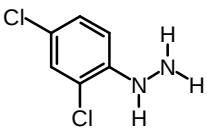
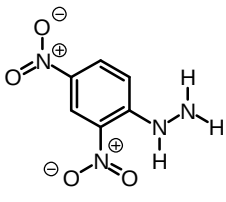
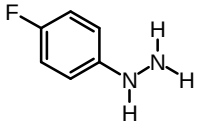
Reaktions-Nr. DS1	Klasse	vorherges. Klasse	Gewinner-neuron	Gewinnerreaktion
600	2	2	(13,4)	566, 593, 598, 599, 601, 607, 608
603	2	2	(13,5)	604
609	2	2	(11,3)	448, 486, 487, 610
611	2	2	(13,2)	
612	2	2	(11,3)	448, 486, 487, 610
613	2	2	(11,3)	448, 486, 487, 610
614	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
615	2	2	(10,1)	58, 301, 399, 402, 406, 408, 439, 569
619	2	1	(2,5)	78, 306, 309, 617, 622
623	2	2	(2,3)	466, 626
625	2	2	(2,3)	466, 626

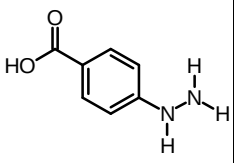
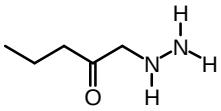
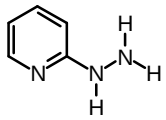
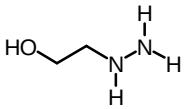
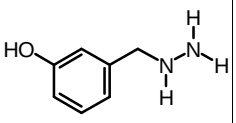
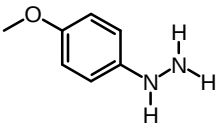
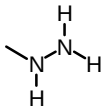
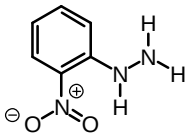
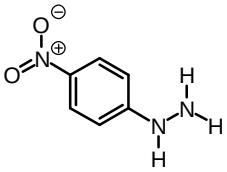
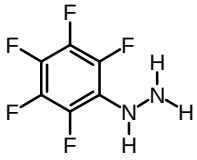
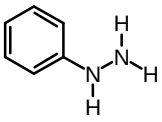
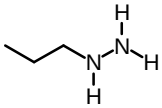
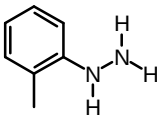
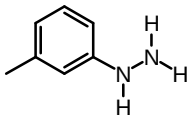
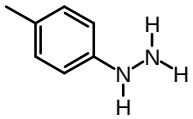
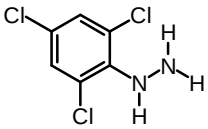
A.5 Ausgangsverbindungen zur kombinatorischen Bibliothek I

8 1,3-Dicarbonylverbindungen

 **1**	 **2**	 **3**	 **4**
 **5**	 **6**	 **7**	 **8**

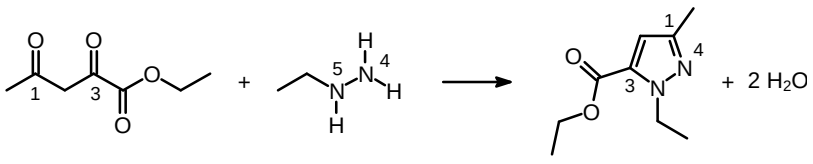
31 Hydrazinderivate

 **1**	 **2**	 **3**	 **4**	 **5**
 **6**	 **7**	 **8**	 **9**	 **10**
 **11**	 **12**	 **13**	 **14**	 **15**

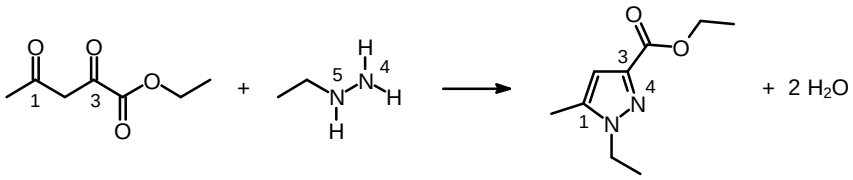
 **16**	 **17**	 **18**	 **19**	 **20**
 **21**	 **22**	 **23**	 **24**	 **25**
 **26**	 **27**	 **28**	 **29**	 **30**
 **31**				

Um eindeutig zwischen den Regioisomeren unterscheiden zu können, wurden die beiden Kohlenstoffatome der Dicarbonylgruppe mit den Ziffern 1 und 3 gekennzeichnet.

Reagiert das Atom mit der Nummer 1 der ersten 1,3-Dicarbonylverbindung mit Atom Nummer 4 der ersten Hydrazinverbindung, so wird diese Reaktion mit „$^{1}1 + 1$“ abgekürzt.



Das andere Regioisomere entsteht, wenn das gekennzeichnete Atom Nummer 3 der 1,3-Dicarbonylverbindung mit Atom 4 der Hydrazinverbindung reagiert. In diesem Fall lautet die Abkürzung „$^{3}1 + 1$“.



A.6 Vorhergesagte Regioisomere der kombinatorischen Bibliothek I

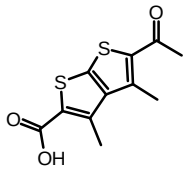
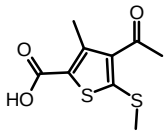
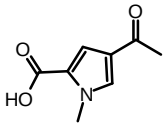
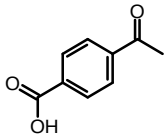
Neuron	Reaktionsnummer	Edukte
(9,1)	8, 10, 56, 58, 120, 122, 136, 138, 264, 266, 296, 298, 312, 314, 344, 346, 424, 426	^{1}4 + 1, ^{3}5 + 1, ^{1}4 + 4, ^{3}5 + 4, ^{1}4 + 8, ^{3}5 + 8, ^{1}4 + 9, ^{3}5 + 9, ^{1}4 + 17, ^{3}5 + 17, ^{1}4 + 19, ^{3}5 + 19, ^{1}4 + 20, ^{3}5 + 20, ^{1}4 + 22, ^{3}5 + 22, ^{1}4 + 27, ^{3}5 + 27
(8,2)	72, 74, 88, 90, 104, 106, 152, 154, 168, 170, 184, 186, 200, 202, 232, 234, 248, 250, 280, 282, 328, 330, 392, 394, 408, 410, 440, 442, 456, 458, 472, 474, 488, 490	^{1}4 + 5, ^{3}5 + 5, ^{1}4 + 6, ^{3}5 + 6, ^{1}4 + 7, ^{3}5 + 7, ^{1}4 + 10, ^{3}5 + 10, ^{1}4 + 11, ^{3}5 + 11, ^{1}4 + 12, ^{3}5 + 12, ^{1}4 + 13, ^{3}5 + 13, ^{1}4 + 15, ^{3}5 + 15, ^{1}4 + 16, ^{3}5 + 16, ^{1}4 + 18, ^{3}5 + 18, ^{1}4 + 21, ^{3}5 + 21, ^{1}4 + 25, ^{3}5 + 25, ^{1}4 + 26, ^{3}5 + 26, ^{1}4 + 28, ^{3}5 + 28, ^{1}4 + 29, ^{3}5 + 29, ^{1}4 + 30, ^{3}5 + 30, ^{1}4 + 31, ^{3}5 + 31
(8,3)	24, 26	^{1}4 + 2, ^{3}5 + 2
(9,3)	216, 218, 360, 362, 376, 378	^{1}4 + 14, ^{3}5 + 14, ^{1}4 + 23, ^{3}5 + 23, ^{1}4 + 24, ^{3}5 + 24
X(7,4)X	35, 36, 37, 38, 45, 46	^{3}2 + 3, ^{1}2 + 3, ^{3}3 + 3, ^{1}3 + 3, ^{3}7 + 3, ^{1}7 + 3
(6,5)	18, 66, 82, 98, 146, 162, 178, 194, 210, 226, 242, 274, 322, 354, 370, 386, 402, 434, 450, 466, 482	^{1}1 + 2, ^{1}1 + 5, ^{1}1 + 6, ^{1}1 + 7, ^{1}1 + 10, ^{1}1 + 11, ^{1}1 + 12, ^{1}1 + 13, ^{1}1 + 14, ^{1}1 + 15, ^{1}1 + 16, ^{1}1 + 18, ^{1}1 + 21, ^{1}1 + 23, ^{1}1 + 24, ^{1}1 + 25, ^{1}1 + 26, ^{1}1 + 28, ^{1}1 + 29, ^{1}1 + 30, ^{1}1 + 31
(1,6)	13, 61, 125, 141, 269, 301, 317, 349, 429	^{3}7 + 1, ^{3}7 + 4, ^{3}7 + 8, ^{3}7 + 9, ^{3}7 + 17, ^{3}7 + 19, ^{3}7 + 20, ^{3}7 + 22, ^{3}7 + 27
(2,6)	29, 77, 93, 109, 157, 173, 189, 205, 221, 237, 253, 285, 333, 365, 381, 397, 413, 445, 461, 477, 493	^{3}7 + 2, ^{3}7 + 5, ^{3}7 + 6, ^{3}7 + 7, ^{3}7 + 10, ^{3}7 + 11, ^{3}7 + 12, ^{3}7 + 13, ^{3}7 + 14, ^{3}7 + 15, ^{3}7 + 16, ^{3}7 + 18, ^{3}7 + 21, ^{3}7 + 23, ^{3}7 + 24, ^{3}7 + 25, ^{3}7 + 26, ^{3}7 + 28, ^{3}7 + 29, ^{3}7 + 30, ^{3}7 + 31
(5,6)	258	^{1}1 + 17
(6,6)	28, 32, 220, 224, 256, 368, 384	^{1}6 + 2, ^{1}8 + 2, ^{1}6 + 14, ^{1}8 + 14, ^{1}8 + 16, ^{1}8 + 23, ^{1}8 + 24
(8,6)	14, 62, 78, 94, 110, 126, 142, 158, 174, 190, 206, 238, 270, 286, 302, 318, 334, 350, 398, 414, 430, 446, 462, 478, 494	^{1}7 + 1, ^{1}7 + 4, ^{1}7 + 5, ^{1}7 + 6, ^{1}7 + 7, ^{1}7 + 8, ^{1}7 + 9, ^{1}7 + 10, ^{1}7 + 11, ^{1}7 + 12, ^{1}7 + 13, ^{1}7 + 15, ^{1}7 + 17, ^{1}7 + 18, ^{1}7 + 19, ^{1}7 + 20, ^{1}7 + 21, ^{1}7 + 22, ^{1}7 + 25, ^{1}7 + 26, ^{1}7 + 27, ^{1}7 + 28, ^{1}7 + 29, ^{1}7 + 30, ^{1}7 + 31

Neuron	Reaktionsnummer	Edukte
(1,7)	17, 27, 31, 65, 75, 79, 81, 91, 95, 97, 107, 111, 145, 155, 159, 161, 171, 175, 177, 187, 191, 193, 203, 207, 225, 235, 239, 241, 251, 255, 273, 283, 287, 321, 331, 335, 353, 363, 367, 369, 379, 383, 385, 395, 399, 401, 411, 415, 433, 443, 447, 449, 459, 463, 465, 475, 479, 481, 491, 495	$^{3}1 + 2$, $^{3}6 + 2$, $^{3}8 + 2$, $^{3}1 + 5$, $^{3}6 + 5$, $^{3}8 + 5$, $^{3}1 + 6$, $^{3}6 + 6$, $^{3}8 + 6$, $^{3}1 + 7$, $^{3}6 + 7$, $^{3}8 + 7$, $^{3}1 + 10$, $^{3}6 + 10$, $^{3}8 + 10$, $^{3}1 + 11$, $^{3}6 + 11$, $^{3}8 + 11$, $^{3}1 + 12$, $^{3}6 + 12$, $^{3}8 + 12$, $^{3}1 + 13$, $^{3}6 + 13$, $^{3}8 + 13$, $^{3}1 + 15$, $^{3}6 + 15$, $^{3}8 + 15$, $^{3}1 + 16$, $^{3}6 + 16$, $^{3}8 + 16$, $^{3}1 + 18$, $^{3}6 + 18$, $^{3}8 + 18$, $^{3}1 + 21$, $^{3}6 + 21$, $^{3}8 + 21$, $^{3}1 + 23$, $^{3}6 + 23$, $^{3}8 + 23$, $^{3}1 + 24$, $^{3}6 + 24$, $^{3}8 + 24$, $^{3}1 + 25$, $^{3}6 + 25$, $^{3}8 + 25$, $^{3}1 + 26$, $^{3}6 + 26$, $^{3}8 + 26$, $^{3}1 + 28$, $^{3}6 + 28$, $^{3}8 + 28$, $^{3}1 + 29$, $^{3}6 + 29$, $^{3}8 + 29$, $^{3}1 + 30$, $^{3}6 + 30$, $^{3}8 + 30$, $^{3}1 + 31$, $^{3}6 + 31$, $^{3}8 + 31$
(6,7)	76, 80, 92, 96, 108, 112, 156, 160, 172, 176, 188, 192, 204, 208, 236, 240, 252, 284, 288, 332, 336, 364, 380, 396, 400, 412, 416, 444, 448, 460, 464, 476, 480, 492, 496	$^{1}6 + 5$, $^{1}8 + 5$, $^{1}6 + 6$, $^{1}8 + 6$, $^{1}6 + 7$, $^{1}8 + 7$, $^{1}6 + 10$, $^{1}8 + 10$, $^{1}6 + 11$, $^{1}8 + 11$, $^{1}6 + 12$, $^{1}8 + 12$, $^{1}6 + 13$, $^{1}8 + 13$, $^{1}6 + 15$, $^{1}8 + 15$, $^{1}6 + 16$, $^{1}6 + 18$, $^{1}8 + 18$, $^{1}6 + 21$, $^{1}8 + 21$, $^{1}6 + 23$, $^{1}6 + 24$, $^{1}6 + 25$, $^{1}8 + 25$, $^{1}6 + 26$, $^{1}8 + 26$, $^{1}6 + 28$, $^{1}8 + 28$, $^{1}6 + 29$, $^{1}8 + 29$, $^{1}6 + 30$, $^{1}8 + 30$, $^{1}6 + 31$, $^{1}8 + 31$
(7,7)	30, 222, 254, 366, 382	$^{1}7 + 2$, $^{1}7 + 14$, $^{1}7 + 16$, $^{1}7 + 23$, $^{1}7 + 24$
X(1,8)X	33, 34, 43, 44, 47, 48, 209, 219, 223	$^{3}1 + 3$, $^{1}1 + 3$, $^{3}6 + 3$, $^{1}6 + 3$, $^{3}8 + 3$, $^{1}8 + 3$, $^{3}1 + 14$, $^{3}6 + 14$, $^{3}8 + 14$
X(3,8)X	3, 4, 5, 6, 51, 52, 53, 54, 131, 132, 133, 134, 291, 292, 293, 294, 307, 308, 309, 310, 339, 340, 341, 342, 419, 420, 421, 422	$^{3}2 + 1$, $^{1}2 + 1$, $^{3}3 + 1$, $^{1}3 + 1$, $^{3}2 + 4$, $^{1}2 + 4$, $^{3}3 + 4$, $^{1}3 + 4$, $^{3}2 + 9$, $^{1}2 + 9$, $^{3}3 + 9$, $^{1}3 + 9$, $^{3}2 + 19$, $^{1}2 + 19$, $^{3}3 + 19$, $^{1}3 + 19$, $^{3}2 + 20$, $^{1}2 + 20$, $^{3}3 + 20$, $^{1}3 + 20$, $^{3}2 + 22$, $^{1}2 + 22$, $^{3}3 + 22$, $^{1}3 + 22$, $^{3}2 + 27$, $^{1}2 + 27$, $^{3}3 + 27$, $^{1}3 + 27$
(5,8)	2, 12, 16, 50, 60, 64, 114, 124, 128, 130, 140, 144, 268, 272, 290, 300, 304, 306, 316, 320, 338, 348, 352, 418, 428, 432	$^{1}1 + 1$, $^{1}6 + 1$, $^{1}8 + 1$, $^{1}1 + 4$, $^{1}6 + 4$, $^{1}8 + 4$, $^{1}1 + 8$, $^{1}6 + 8$, $^{1}8 + 8$, $^{1}1 + 9$, $^{1}6 + 9$, $^{1}8 + 9$, $^{1}6 + 17$, $^{1}8 + 17$, $^{1}1 + 19$, $^{1}6 + 19$, $^{1}8 + 19$, $^{1}1 + 20$, $^{1}6 + 20$, $^{1}8 + 20$, $^{1}1 + 22$, $^{1}6 + 22$, $^{1}8 + 22$, $^{1}1 + 27$, $^{1}6 + 27$, $^{1}8 + 27$
(1,9)	271	$^{3}8 + 17$
(2,9)	1, 11, 15, 49, 59, 63, 113, 123, 127, 129, 139, 143, 257, 267, 289, 299, 303, 305, 315, 319, 337, 347, 351, 417, 427, 431	$^{3}1 + 1$, $^{3}6 + 1$, $^{3}8 + 1$, $^{3}1 + 4$, $^{3}6 + 4$, $^{3}8 + 4$, $^{3}1 + 8$, $^{3}6 + 8$, $^{3}8 + 8$, $^{3}1 + 9$, $^{3}6 + 9$, $^{3}8 + 9$, $^{3}1 + 17$, $^{3}6 + 17$, $^{3}1 + 19$, $^{3}6 + 19$, $^{3}8 + 19$, $^{3}1 + 20$, $^{3}6 + 20$, $^{3}8 + 20$, $^{3}1 + 22$, $^{3}6 + 22$, $^{3}8 + 22$, $^{3}1 + 27$, $^{3}6 + 27$, $^{3}8 + 27$

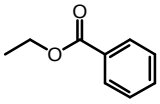
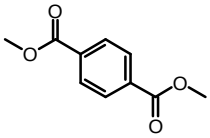
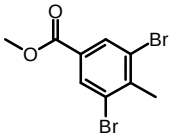
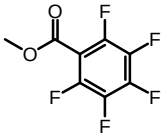
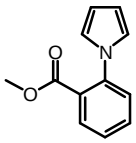
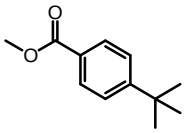
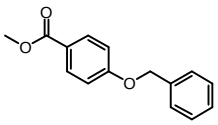
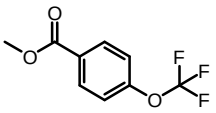
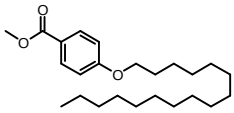
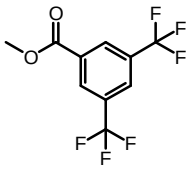
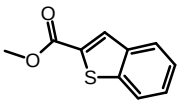
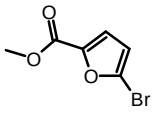
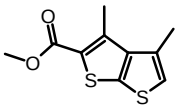
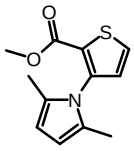
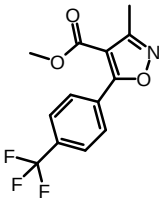
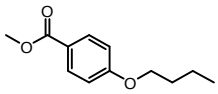
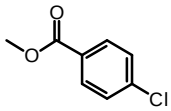
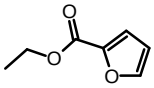
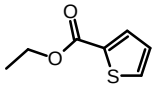
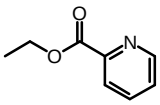
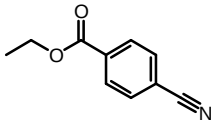
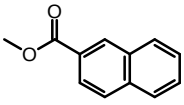
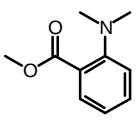
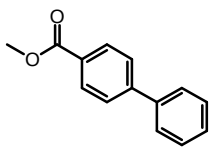
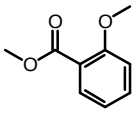
Neuron	Reaktionsnummer	Edukte
X(5,9)X	67, 68, 69, 70, 83, 84, 85, 86, 99, 100, 101, 102, 115, 116, 117, 118, 147, 148, 149, 150, 163, 164, 165, 166, 179, 180, 181, 182, 195, 196, 197, 198, 227, 228, 229, 230, 259, 260, 261, 262, 323, 324, 325, 326, 387, 388, 389, 390, 403, 404, 405, 406, 435, 436, 437, 438, 451, 452, 453, 454, 467, 468, 469, 470, 483, 484, 485, 486	$^{3}2 + 5$, $^{1}2 + 5$, $^{3}3 + 5$, $^{1}3 + 5$, $^{3}2 + 6$, $^{1}2 + 6$, $^{3}3 + 6$, $^{1}3 + 6$, $^{3}2 + 7$, $^{1}2 + 7$, $^{3}3 + 7$, $^{1}3 + 7$, $^{3}2 + 8$, $^{1}2 + 8$, $^{3}3 + 8$, $^{1}3 + 8$, $^{3}2 + 10$, $^{1}2 + 10$, $^{3}3 + 10$, $^{1}3 + 10$, $^{3}2 + 11$, $^{1}2 + 11$, $^{3}3 + 11$, $^{1}3 + 11$, $^{3}2 + 12$, $^{1}2 + 12$, $^{3}3 + 12$, $^{1}3 + 12$, $^{3}2 + 13$, $^{1}2 + 13$, $^{3}3 + 13$, $^{1}3 + 13$, $^{3}2 + 15$, $^{1}2 + 15$, $^{3}3 + 15$, $^{1}3 + 15$, $^{3}2 + 17$, $^{1}2 + 17$, $^{3}3 + 17$, $^{1}3 + 17$, $^{3}2 + 21$, $^{1}2 + 21$, $^{3}3 + 21$, $^{1}3 + 21$, $^{3}2 + 25$, $^{1}2 + 25$, $^{3}3 + 25$, $^{1}3 + 25$, $^{3}2 + 26$, $^{1}2 + 26$, $^{3}3 + 26$, $^{1}3 + 26$, $^{3}2 + 28$, $^{1}2 + 28$, $^{3}3 + 28$, $^{1}3 + 28$, $^{3}2 + 29$, $^{1}2 + 29$, $^{3}3 + 29$, $^{1}3 + 29$, $^{3}2 + 30$, $^{1}2 + 30$, $^{3}3 + 30$, $^{1}3 + 30$, $^{3}2 + 31$, $^{1}2 + 31$, $^{3}3 + 31$, $^{1}3 + 31$
(2,10)	7, 9, 55, 57, 119, 121, 135, 137, 263, 265, 295, 297, 311, 313, 343, 345, 423, 425	$^{3}4 + 1$, $^{1}5 + 1$, $^{3}4 + 4$, $^{1}5 + 4$, $^{3}4 + 8$, $^{1}5 + 8$, $^{3}4 + 9$, $^{1}5 + 9$, $^{3}4 + 17$, $^{1}5 + 17$, $^{3}4 + 19$, $^{1}5 + 19$, $^{3}4 + 20$, $^{1}5 + 20$, $^{3}4 + 22$, $^{1}5 + 22$, $^{3}4 + 27$, $^{1}5 + 27$
X(6,10)X	19, 20, 21, 22, 211, 212, 213, 214, 243, 244, 245, 246, 275, 276, 277, 278, 355, 356, 357, 358, 371, 372, 373, 374	$^{3}2 + 2$, $^{1}2 + 2$, $^{3}3 + 2$, $^{1}3 + 2$, $^{3}2 + 14$, $^{1}2 + 14$, $^{3}3 + 14$, $^{1}3 + 14$, $^{3}2 + 16$, $^{1}2 + 16$, $^{3}3 + 16$, $^{1}3 + 16$, $^{3}2 + 18$, $^{1}2 + 18$, $^{3}3 + 18$, $^{1}3 + 18$, $^{3}2 + 23$, $^{1}2 + 23$, $^{3}3 + 23$, $^{1}3 + 23$, $^{3}2 + 24$, $^{1}2 + 24$, $^{3}3 + 24$, $^{1}3 + 24$
(3,11)	23, 25, 247, 249	$^{3}4 + 2$, $^{1}5 + 2$, $^{3}4 + 16$, $^{1}5 + 16$
(4,11)	71, 73, 87, 89, 103, 105, 151, 153, 167, 169, 183, 185, 199, 201, 231, 233, 279, 281, 327, 329, 391, 393, 407, 409, 439, 441, 455, 457, 471, 473, 487, 489	$^{3}4 + 5$, $^{1}5 + 5$, $^{3}4 + 6$, $^{1}5 + 6$, $^{3}4 + 7$, $^{1}5 + 7$, $^{3}4 + 10$, $^{1}5 + 10$, $^{3}4 + 11$, $^{1}5 + 11$, $^{3}4 + 12$, $^{1}5 + 12$, $^{3}4 + 13$, $^{1}5 + 13$, $^{3}4 + 15$, $^{1}5 + 15$, $^{3}4 + 18$, $^{1}5 + 18$, $^{3}4 + 21$, $^{1}5 + 21$, $^{3}4 + 25$, $^{1}5 + 25$, $^{3}4 + 26$, $^{1}5 + 26$, $^{3}4 + 28$, $^{1}5 + 28$, $^{3}4 + 29$, $^{1}5 + 29$, $^{3}4 + 30$, $^{1}5 + 30$, $^{3}4 + 31$, $^{1}5 + 31$
(4,12)	215, 217, 359, 361, 375, 377	$^{3}4 + 14$, $^{1}5 + 14$, $^{3}4 + 23$, $^{1}5 + 23$, $^{3}4 + 24$, $^{1}5 + 24$
(4,13)	39, 40, 41, 42	$^{3}4 + 3$, $^{1}4 + 3$, $^{1}5 + 3$, $^{3}5 + 3$

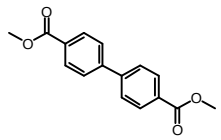
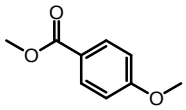
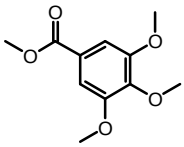
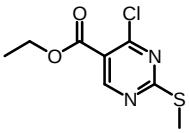
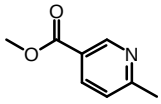
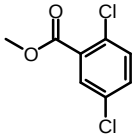
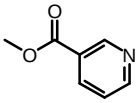
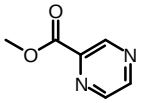
A.7 Ausgangsverbindungen zur kombinatorischen Bibliothek II

4 Ausgangsverbindungen vom Typ R1:

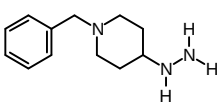
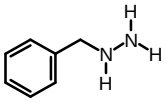
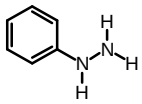
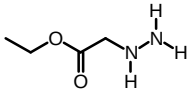
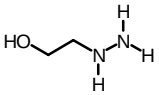
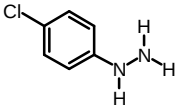
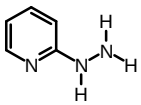
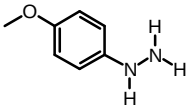
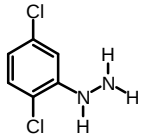
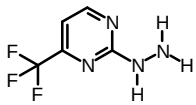
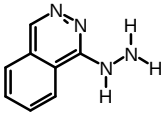
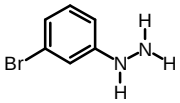
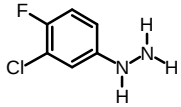
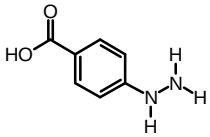
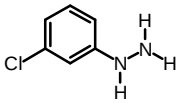
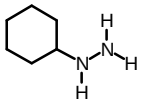
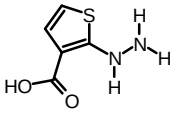
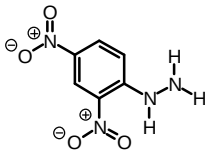
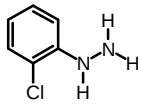
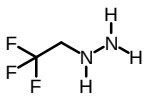
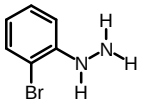
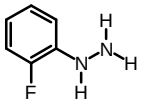
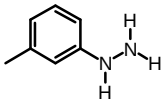
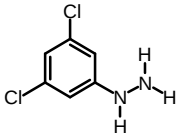
 **1**	 **2**	 **3**	 **4**

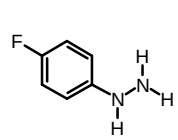
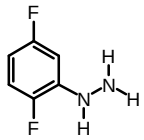
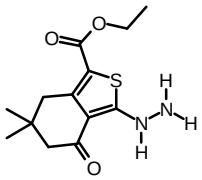
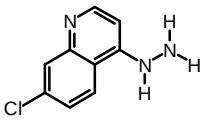
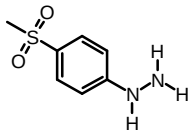
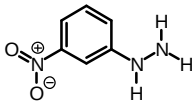
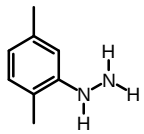
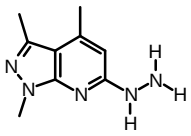
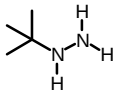
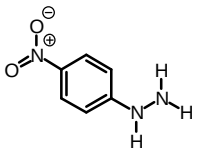
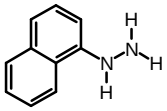
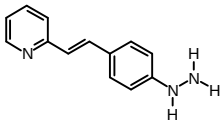
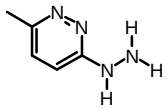
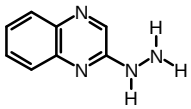
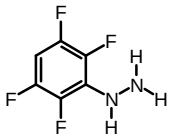
33 Ausgangsverbindungen vom Typ R2:

 **1**	 **2**	 **3**	 **4**	 **5**
 **6**	 **7**	 **8**	 **9**	 **10**
 **11**	 **12**	 **13**	 **14**	 **15**
 **16**	 **17**	 **18**	 **19**	 **20**
 **21**	 **22**	 **23**	 **24**	 **25**

 **26**	 **27**	 **28**	 **29**	 **30**
 **31**	 **32**	 **33**		

40 Ausgangsverbindungen vom Typ R4:

 **1**	 **2**	 **3**	 **4**	 **5**
 **6**	 **7**	 **8**	 **9**	 **10**
 **11**	 **12**	 **13**	 **14**	 **15**
 **16**	 **17**	 **18**	 **19**	 **20**
 **21**	 **22**	 **23**	 **24**	 **25**

 **26**	 **27**	 **28**	 **29**	 **30**
 **31**	 **32**	 **33**	 **34**	 **35**
 **36**	 **37**	 **38**	 **39**	 **40**

Publikationen

1. Sacher, O.
   „Von Daten zu Kurven“
   Nachr. Chem. Tech. Lab., **1997**, *45*, 1007 – 1009.

2. Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K.
   „Classification of Organic Reactions: Similarity of Reactions Based on Changes in the Electronic Features of Oxygen Atoms at the Reaction Sites“
   J. Chem. Inf. Comput. Sci., **1998**, *38*, 210 – 219.

3. Gasteiger, J.; Pförtner, M.; Sitzmann, M.; Höllering, R.; Sacher, O.; Kostka, T.; Karg, N.
   „Computer-Assisted Synthesis and Reaction Planning in Combinatorial Chemistry“
   Persp. Drug Discov. Design, **2000**, *20*, 245 – 264.

Lebenslauf

Name	Oliver Sacher
Geburtsdatum und -ort	02. August 1970 in Nürnberg
Eltern	Rainer und Monika Sacher, geb. Cermak
Staatsangehörigkeit	deutsch
Familienstand	ledig, keine Kinder

Schulbildung

09/1977 - 07/1981	Grundschule Lauf a. d. Pegnitz
09/1981 - 07/1990	Christoph-Jacob-Treu-Gymnasium Lauf a. d. Pegnitz

Hochschulausbildung

09/1990 - 12/1995	Studium der Chemie an der Friedrich-Alexander-Universität Erlangen-Nürnberg
12/1995 - 06/1996	Diplomarbeit bei Prof. Gasteiger am Computer-Chemie-Centrum des Instituts für Organische Chemie der Universität Erlangen-Nürnberg zu dem Thema „Klassifizierung organischer Reaktionen mittels Kohonen-Netze“
seit 07/1996	Promotionsarbeit bei Prof. Gasteiger

Erlangen, 13.11.2000