DIE CODIERUNG DER 3D-STRUKTUR VON MOLEKÜLEN UND IHRE ANWENDUNG ZUR SIMULATION VON IR-SPEKTREN UND FÜR QSAR-UNTERSUCHUNGEN

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

zur

Erlangung des Doktorgrades

vorgelegt von

Jan Heinrich Schuur

aus Bonn

Als Dissertation genehmigt von den

Naturwissenschaftlichen Fakultäten der Universität Erlangen-Nürnberg.

Tag der mündlichen Prüfung:	23.Juli.1998
Vorsitzender der Promotionskomission:	Professor Dr. D. Kölzow
Erstberichterstatter:	Professor Dr. Johann Gasteiger
Zweitberichterstatter:	Privatdozent Dr. Timothy Clark

MEINEN ELTERN UND GROßELTERN

FÜR BIRGIT UND MAX W.

DANKSAGUNG

Mein Dank gebührt meinem Doktorvater, Herrn Professor Dr. Johann Gasteiger für die Aufnahme in seine Arbeitsgruppe, das interessante Thema und die vielfältige Unterstützung während der ganzen Arbeit. Ohne die Arbeitsumgebung in seiner Arbeitsgruppe mit dem Erfahrungsschatz von 3 Jahrzehnten Computereinsatz in der Chemie, der Soft- und Hardwareausstattung hätte diese Arbeit nicht so schnell und erfolgreich durchgeführt werden können.

Professor Dr. Paul von Ragué Schleyer, Professor Dr. Johann Gasteiger und Privatdozent Dr. Timothy Clark sowie der Bayerischen Staatsregierung möchte ich für die Schaffung des Computer-Chemie-Centrums und die Möglichkeit an diesem zu arbeiten, danken. Dieses europaweit größte Centrum für Computeranwendungen in der Chemie gab der Arbeit viele Impulse.

Ein besonderer Dank gilt meinen Kollegen am Computer-Chemie-Centrum, mit denen ich zusammen arbeiten durfte. Ohne sie wäre die Arbeit nicht möglich gewesen. Besonders erwähnen möchte ich (in historischer Reihenfolge):

Dr. Vera Simon und insbesondere Dr. Robert Höllering für ihre Hilfe und die Einführung in den Arbeitskreis.

Dr. Ralf Fick, Dr. Markus Wagener und Dr. Wolfgang Witzenbichler für die Einführung und die Bereitstellung der UNIX Umgebung und dafür, daß nicht jede Frage mit „Tip doch mal Apropos“ beantwortet wurde.

Dr. Jens Sadowski, Dr. Markus Wagener und Prof. Dr. Jure Zupan, ohne deren Vorarbeiten auf dem Gebiet der 3D-Strukturgenerierung und der Counterpropagation Netzwerke diese Arbeit nicht möglich gewesen wäre.

Dr. Klaus-Peter Schulz und Dr. Wolf-Dietrich Ihlenfeldt für die stete Bereitschaft zur Diskussion und die aufmunternde Atmosphäre dieser Diskussionen.

DC Paul Selzer für die stets gute Zusammenarbeit und seine Hilfsbereitschaft, sowie die Wartung der UNIX Umgebung zusammen mit DC Andreas Teckentrup.

Markus Hemmer für manch gute Idee und viele interessante Diskussionen am Ende der Arbeit.

DC Henryette Roth, Dr. Wolfgang Sauer und Dr. Nico van Eikema Hommes für die Einführung in die Welt der quantenmechanischen Berechnungen und die Durchführung der DFT-Berechnungen.

Angela Döbler für ihre Fleißarbeit als Arbeitskreissekretärin und zusammen mit ihr allen anderen Mitarbeitern des Arbeitskreises Professor Gasteiger für die Aufnahme und die Zusammenarbeit.

INHALTSVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

3D-MoRSE Code	3D-Molecular Representation of Structures based on Electron diffraction
3D	dreidimensional
IR-Spektrum	Infrarot Spektrum
NMR-Spektroskopie	Nuclear Magnetic Resonance Spektroskopie
QSAR	Quantitative Structure Activity Relationships
QSPR	Quantitative Structure Property Relationships

Leitsatz:

Die 3D-Struktur eines Moleküls legt die Moleküleigenschaften fest.

1 Leitsatz

Dieser Arbeit liegt ein Leitsatz zugrunde: „Die 3D-Struktur eines Moleküls legt die Moleküleigenschaften fest.“. Dies ist ebenso wahr wie problematisch. Denn wie Abbildung 1 zeigt, ist die 3D-Struktur eines Moleküls keineswegs eindeutig. Trotz des rigiden Gerüsts ist die Lage der Atomkerne aufgrund der Schwingungen des Moleküls nicht eindeutig definiert. Es ist lediglich möglich, Bereiche mit einer Aufenthaltswahrscheinlichkeit der Atomkerne anzugeben. Schon hier ist vereinfacht worden, denn die Elektronen, die die Atomkerne umgeben, sind gar nicht erst erwähnt.

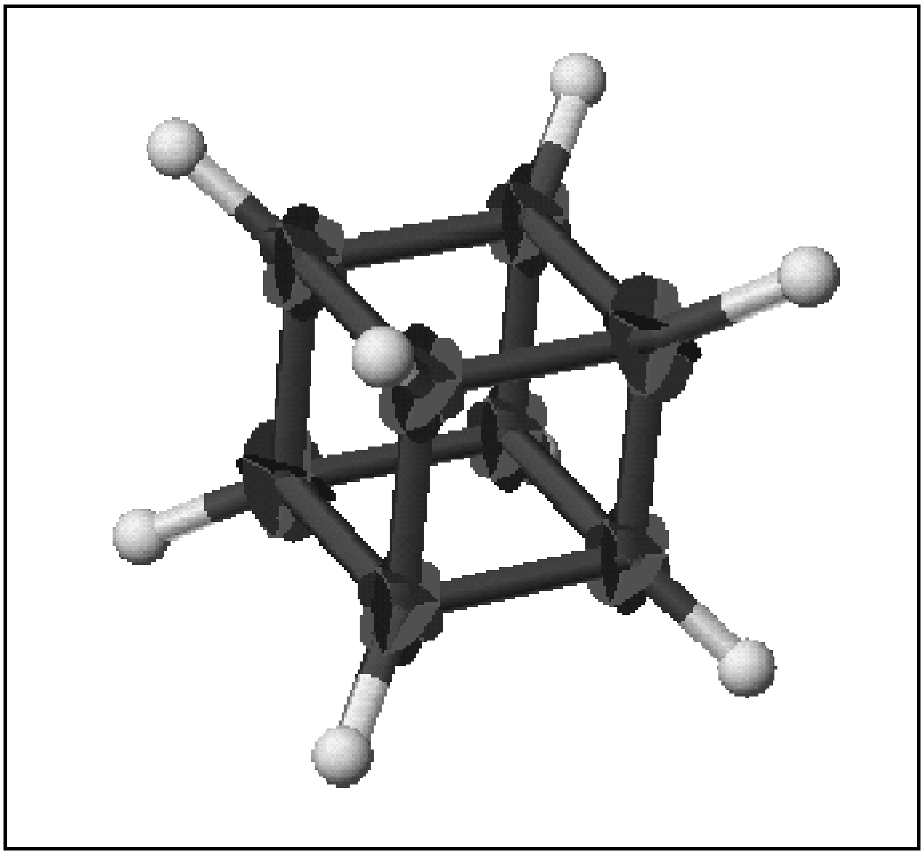


Abbildung 1: Die Kristallstruktur von Cuban mit den Aufenthaltswahrscheinlichkeiten der Kohlenstoffatome

Elektronen sind nur als 3D-Elektronendichteverteilung faßbar, da sie aufgrund des Welle - Teilchen Dualismus am besten als stehende Welle beschrieben werden (Abbildung 2).

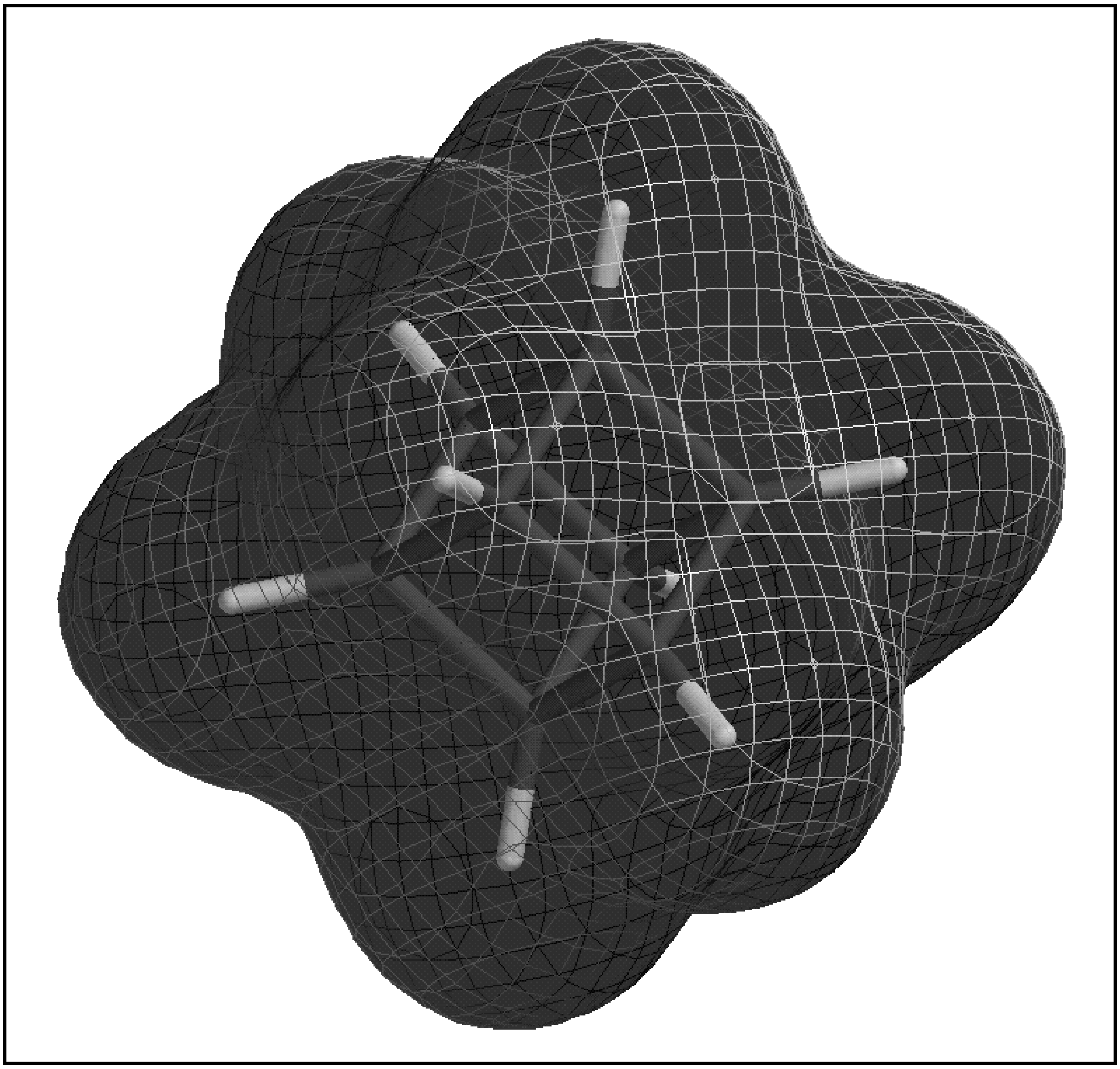


Abbildung 2: Cuban und seine Elektronenwolke (Spartan/AM1)

Für den Fall der Schwingungen und der Elektronendichteverteilung ist die Vereinfachung durch die Angabe der Koordinaten für Atomkerne bzw. Atomschwerpunkte jedoch gut nachvollziehbar, denn die Elektronendichte ist überwiegend radialsymmetrisch und die Beschreibung einer Atomposition durch die hypothetische Ruhelage eines Atoms scheint logisch. Ein deutlich schwerwiegenderes Problem ergibt sich allerdings bei der Betrachtung konformativ flexibler Moleküle, wie Abbildung 3 mit den zwei im Kristall existierenden Konformationen von 2-Acetylamino-4-oxo-4-(2-aminophenyl)-butansäure zeigt.

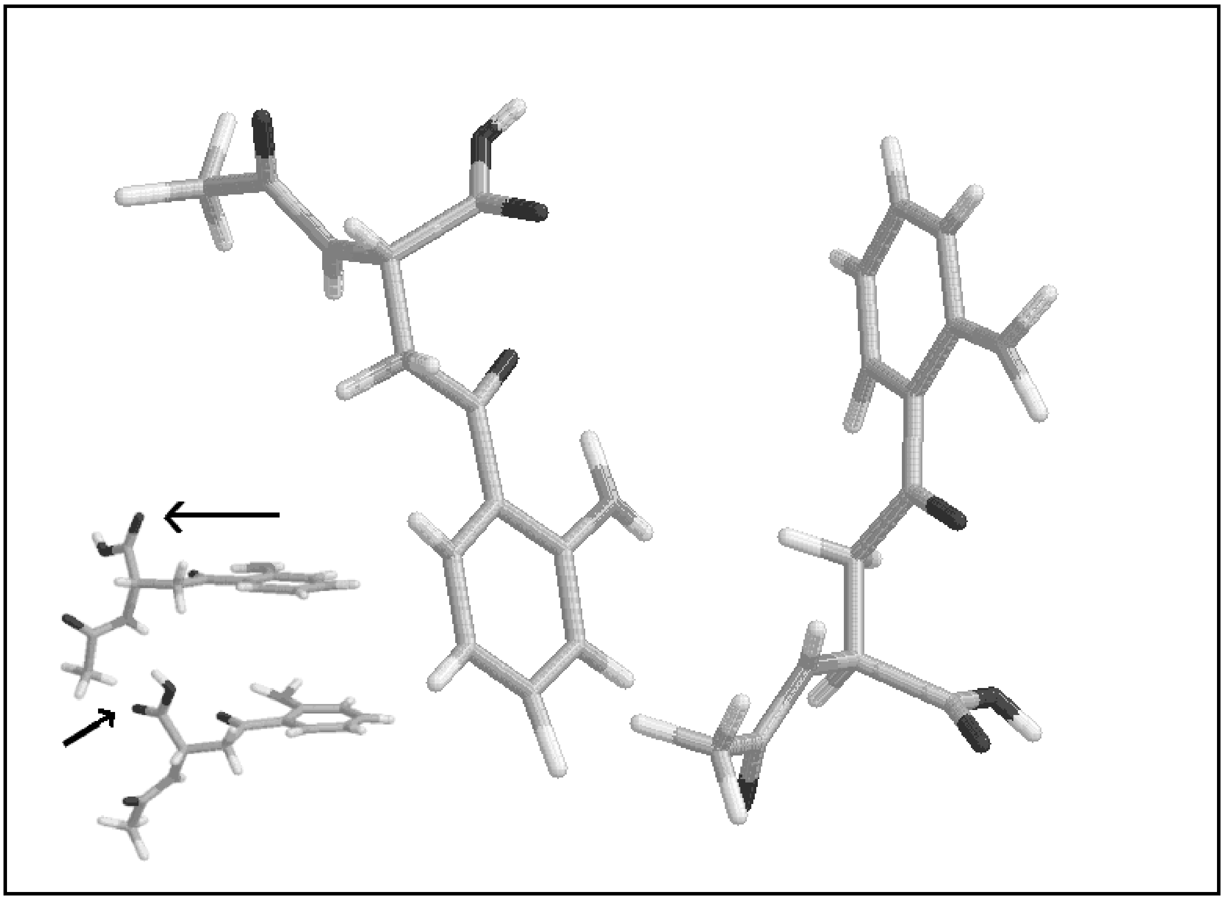


Abbildung 3: Die Konformere von 2-Acetylamino-4-oxo-4-(2-aminophenyl)-butansäure, wie sie in der Kristallzelle vorliegen (groß).[1] Am deutlichsten sind die Unterschiede in der Stellung der Carboxylgruppe zum Benzolring zu erkennen. Der Torsionswinkel ausgehend vom markierten Carbonyl-Sauerstoff zum Carbonyl-Kohlenstoff und weiter über den benachbarten tertiären Kohlenstoff zum Wasserstoffatom am tertiären Kohlenstoffatom unterscheidet sich zwischen beiden Konformeren um 183.8° (links unten, Moleküle am Benzolring ausgerichtet).

Damit ist klar: konformativ flexible Moleküle können selbst im kristallinen Zustand, der häufig die energieärmste Zustandsform ist, in verschiedenen Konformationen existieren. So kann die Frage nach der zu codierenden 3D-Struktur eines Moleküls nicht eindeutig mit einer energieärmsten Konformation beantwortet werden. Erschwerend kommt noch hinzu, daß die energieärmste Konformation eines Moleküls im Vakuum oder im Kristall nicht die wirksame Konformation des Moleküls an einem Substrat sein muß, wie das folgende Beispiel des Inhibitors A79285 der HIV-1 Protease zeigt. In der mittels AM1 berechneten Vakuumstruktur ist die Stellung des mittleren Benzolrings des Inhibitors deutlich anders als in der aktiven Konformation in der Rezeptortasche. Dies ist insofern von besonderer Bedeutung, da sich dieser Benzolring mutmaßlich in der Nähe des aktiven Zentrums befindet, da hier die polaren Gruppen des

Inhibitors konzentriert sind (Acetalstruktur und zwei geminale Fluoratome). Abbildung 4 zeigt die Struktur am Rezeptor und eine mit AM1 aus der Rezeptorkonformation optimierte Konformation. Die Änderung der Konformation ist am deutlichsten an der Stellung des Benzolrings rechts der Mitte zu sehen.

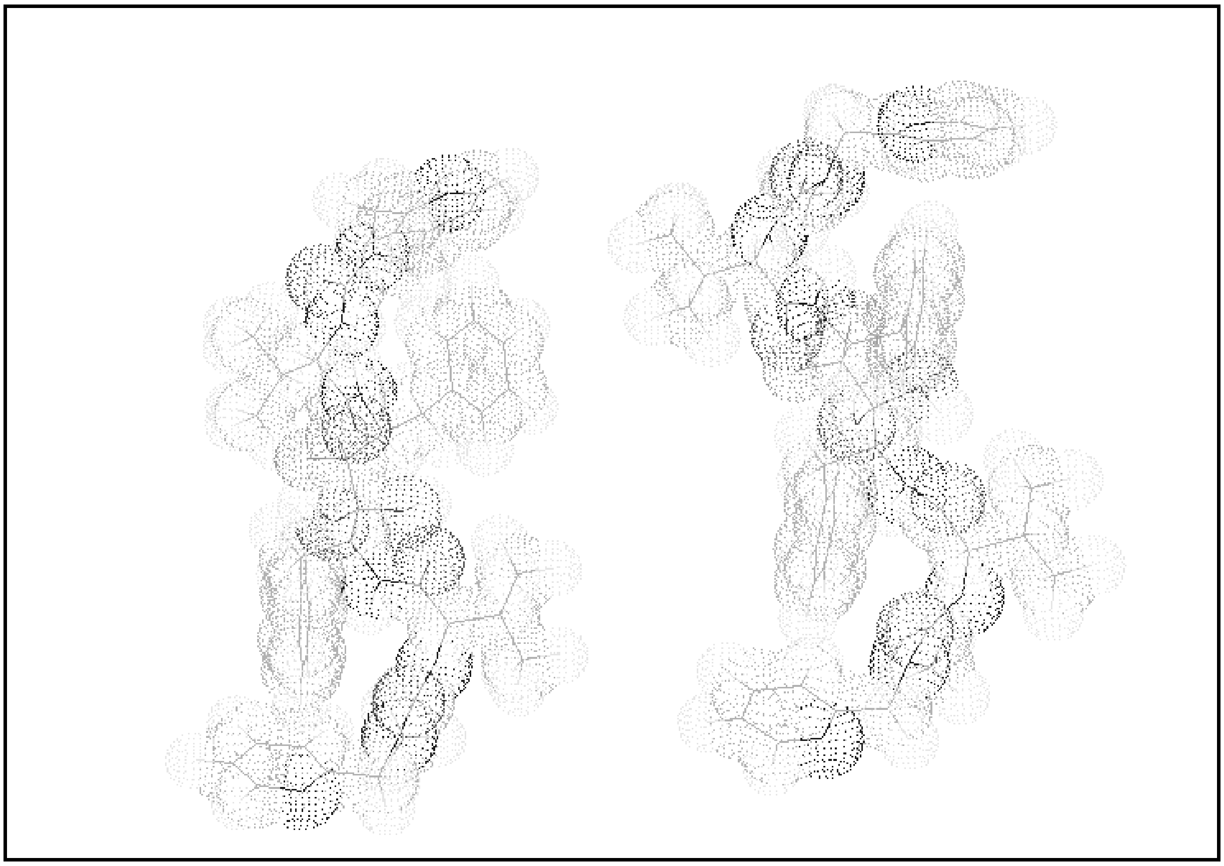


Abbildung 4: Die Konformation des HIV-1 Proteaseinhibitors A79285 am Rezeptor (links) und die aus der Rezeptorkonformation mit AM1 optimierte Vakuumkonformation (rechts).

1.1 Fazit:

Bei der Codierung der 3D-Struktur von Molekülen sind die gemachten Vereinfachungen zu bedenken. Statt der Position eines Atomkerns kann nur seine Aufenthaltswahrscheinlichkeit bestimmt werden und die Positionen der Elektronen können eigentlich nur mit einer Elektronenwolke beschrieben werden. Dennoch von einem punktförmigen Atom auszugehen ist mitunter bereits eine drastische Vereinfachung, aber eine derartige Beschreibung eines Moleküls ist in bezug auf viele makroskopische Moleküleigenschaften eindeutig. Für die verschiedenen Konformationen eines Moleküls gilt dies nicht mehr, insofern sollte ein 3D-Strukturcode von der Konformation eines Moleküls abhängen. Allerdings ist zu fragen, wieweit der Einfluß der

Konformation auf den 3D-Strukturcode gehen sollte. Die Rotationsfrequenz einer Methylgruppe liegt in der Größenordnung von 30 MHz bei Raumtemperatur. Langsame Meßmethoden werden damit meistens nur das thermische Gleichgewicht der verschiedenen Konformationsisomere sehen und die aktive Konformation biologisch aktiver Verbindungen in der sie an den Rezeptor binden, die oft nicht bekannt ist, wird umgehend aus dem Gleichgewicht nachgebildet werden können. Womit sich die Frage nach einer repräsentativen Darstellung der 3D-Struktur stellt, die weitgehend unabhängig von der codierten Konformation ist, solange eine Konformationsänderung keine drastische Änderung der Moleküleigenschaften erwarten läßt.

2 *Aufgabengebiete der Arbeit*

Die Aufgabenstellung dieser Arbeit war die Entwicklung einer 3D-Strukturcodierung im Rahmen von Projekten zur Interpretation von Infrarotspektren. Allerdings war schon von Anfang an klar, daß sich eine 3D-Strukturcodierung auch im Bereich der Quantitativen Struktur-Aktivitäts-Beziehungen (QSAR) zur Vorhersage biologischer Wirkungen von Stoffen einsetzen lassen müßte.

2.1 Infrarotspektroskopie

Die Infrarotspektroskopie, oder auch kurz IR-Spektroskopie genannt, wurde vielfach von der NMR-Spektroskopie als Standardmethode zur Strukturaufklärung in der organischen Chemie abgelöst, obwohl IR-Spektren noch in den fünfziger Jahren das wesentliche Instrument zur Strukturaufklärung in der organischen Chemie waren. Wie war das möglich? Ganz einfach, die NMR-Spektroskopie entsprach den Anforderungen und der Denkweise des Chemikers im Hinblick auf die Strukturaufklärung besser. Die in der NMR-Spektroskopie mögliche selektive Anregung der einzelnen Kerne und die lokaleren Zusammenhänge der NMR-Absorptionslinien mit der chemischen Umgebung der einzelnen Atome, entsprechen dem üblichen Forschungsgrundsatz *teile und untersuche* besser als eine übliche IR-Absorptionslinie, die auf der Schwingung eines ganzen Moleküls beruht. Zudem boten die lokaleren Zusammenhänge in der NMR-Spektroskopie die Möglichkeit, einfache Inkrementsysteme zur Bestimmung der Absorptionslage der einzelnen Atome im Molekül zu entwickeln, mit denen jeder Laborchemiker arbeiten konnte. Ähnliche Versuche für die IR-Spektroskopie scheiterten jedoch, da selbst die Lage der Carbonylbande vom gesamten Rest des Moleküls mit beeinflußt wird.

Auch die Massenspektroskopie nahm der Infrarotspektroskopie Marktanteile weg. Der direkte Wechsel für eine Fragestellung von der Infrarotspektroskopie zur Massenspektroskopie ist zwar unwahrscheinlich, da die gelieferten Informationen zu verschieden sind. Für neue analytische Fragestellungen wurde aber die Massenspektroskopie bevorzugt vor der Infrarotspektroskopie verwendet. Die höhere Empfindlichkeit und die klarere Aussage des Molpeaks waren die Argumente für die Massenspektroskopie. Hierüber geriet ein Teil der Vorteile der Infrarotspektroskopie in Vergessenheit.

2.1.1 Vorteile der Infrarotspektroskopie

Die Infrarotspektroskopie vereint eine Reihe von Vorteilen der anderen spektroskopischen Verfahren, wie NMR-Spektroskopie, Massenspektroskopie und UV/Vis-Spektroskopie in sich. Mit Hilfe der Infrarotspektroskopie kann in *allen Phasen*, fest, flüssig, gasförmig und gelöst gemessen werden: fest als KBr-Preßling, flüssig und gasförmig in der Küvette, gelöst z.B. in Halogenalkanen und als Verreibung in Kohlenwasserstoffen.

Weitere Vorteile der Infrarotspektroskopie zeigen sich an der GC-IR-Kopplung, die erst durch die *kurzen Meßzeiten* und den *geringen Substanzbedarf* der IR-Spektroskopie möglich wurde und zudem *zerstörungsfrei* ist. Die GC-IR-Kopplung ist nicht die einzige Kopplung geblieben. So gibt es Verfahren zur LC-Kopplung mit einer Durchflußzelle und zur HPLC-Kopplung mit dem Auftrag des Eluats auf eine Trägerscheibe und anschließende Messung mit einem IR-Mikroskop. Überhaupt die Messung kleinster Probemengen auf oder in einem Trägermaterial ist eine weitere Domäne der Infrarotspektroskopie, man denke nur an die IR-Mikroskopie an Polymerbeads in der kombinatorischen Chemie oder an Matrixisolationsmessungen instabiler Moleküle.

Im Gegensatz zu IR-Matrixisolationsmessungen, geht es bei IR *in-situ* Messungen um Konzentrationsbestimmungen im Mol- und Tonnenmaßstab. Ein interessantes Beispiel ist die quantitative Bestimmung von Toluoldiisocyanat (TDI) im Polymerisationsgemisch vor der Polymerisation.[2] Die Konzentration von TDI beeinflußt die Qualität des Polymerisates. Hierzu wird vom Polarisationsgemisch ein Reflexionsspektrum aufgenommen und die beiden Isocyanatbanden bei 2250 und 1700 cm^{-1} ausgewertet. Vor der Auswertung ist jedoch eine Kalibrierung der Methode notwendig. Gleiches gilt für die Bestimmung sauerstoffhaltiger Additive in Benzin durch Auswertung der CO-Bande.[2]

Die IR-Spektroskopie wirkt nach NMR/ESR und Mikrowellenspektroskopie mit der geringsten Energiemenge auf das zu untersuchende Molekül ein. Damit ist es unwahrscheinlich, daß die Meßbedingungen eine chemische Veränderung der zu untersuchenden Probe bewirken. Zudem wird ein Infrarotspektrum nach der UV/VIS-Spektroskopie den geringsten Energieaufwand für die Messung benötigen, da weder Hochvakuum noch starke Magnetfelder benötigt werden. Damit könnte der Ersatz von MS oder NMR-Messungen durch IR-Messungen auch einen Beitrag zum Umweltschutz liefern.

Bleibt festzuhalten: Die Infrarotspektroskopie sollte eine der interessantesten Methoden für die Routineanalytik sein. Die Kombination von hohem Informationsgehalt der Spektren, hoher Empfindlichkeit und kurzen Meßzeiten verbunden mit einer breiten Auswahl an Meßbedingungen ist relativ einzigartig in der Analytik, obwohl die Infrarotspektroskopie in einzelnen Disziplinen durch die anderen spektroskopischen Techniken übertroffen wird. Die Nachweisgrenze liegt beispielsweise in der Massenspektroskopie niedriger als in der Infrarotspektroskopie und der Informationsgehalt multidimensionaler NMR-Spektren wird vermutlich von keiner anderen Spektroskopieart erreicht, allerdings ist ein Infrarotspektrum in Sekunden gemessen, ein multidimensionales NMR-Spektrum kann hingegen Stunden benötigen.

2.1.2 Die Nutzung der gewonnenen Spektren

Infrarotspektren werden heute üblicherweise in der Analytik auf dreierlei Art genutzt. In der quantitativen Analyse werden Peakflächen zu Konzentrationsbestimmungen genutzt. In der qualitativen Analyse dienen IR-Spektren z.B. in der Qualitätskontrolle zur Substanzidentifikation durch Spektrenvergleich mit Referenzsubstanzen. Desweiteren werden IR-Spektren zur Bestimmung physikochemischer Parameter wie Bindungsstärken und (3D-) Strukturen eingesetzt, indem entweder direkt aus einer Bandenlage etwas abgeleitet oder das gemessene Spektrum mit einem quantenmechanisch berechneten Spektrum verglichen wird. Die Strukturbestimmung durch den Vergleich zwischen experimentellem Spektrum und quantenmechanisch berechnetem Spektrum hat sich jedoch nicht als Standardmethode durchsetzen können. Die hohen Anforderungen quantenmechanischer Berechnungen bzw. die Tatsache, daß die Dauer quantenmechanischer Berechnungen einen sehr wahrscheinlichen Strukturvorschlag für den wirtschaftlichen Einsatz erfordert, waren die Hindernisse.

2.1.2.1 *Quantitative Analyse mit IR-Spektren*

Eine Anwendung von IR-Spektren in der quantitativen Analyse ist die Bestimmung des Gehalts funktioneller Gruppen in einem Makromolekül wie die folgende Vorschrift zur Bestimmung von Methoxygruppen in Lignin zeigt. Lignin besteht aus Phenylpropaneinheiten (C9) die, je nach Herkunft, unterschiedlich viele Methoxygruppen am Benzolring tragen. Zur Klassifikation des Lignins wird nach Sarkanen, Chang und Allan[3] das Verhältnis zwischen den Absorptionen bei 1460 cm^{-1}, 1422 cm^{-1}, 1272 cm^{-1}, 1124 cm^{-1} und dem konstanten Absorptionsmaximum bei 1500 cm^{-1} bestimmt. Die beiden Absorptionsmaxima bei 1460 und 1422 cm^{-1} repräsentieren die Schwingungen der Methylgruppen. Die Absorptionen bei 1272 und 1124 cm^{-1} resultieren aus Schwingungen der C-O Bindungen. Aus den Absorptionen kann dann auf den Gehalt an Methoxygruppen und damit auf das Verhältnis von 4-Hydroxy-3-methoxypropan- und 3,5-Dimethoxy-4-hydroxypropaneinheiten im zu charakterisierenden Lignin geschlossen werden.

Natürlich können IR-Spektren auch zu Gehaltsbestimmungen bei Mischungen eingesetzt werden, wie die oben genannten Beispiele der Additivbestimmung in Benzin oder der Konzentrationsbestimmung von TDI im Polymerisationsgemisch zeigen. Ein weiteres Beispiel ist die Bestimmung von Kohlenwasserstoffen in Böden nach DIN 38409 H18. Nach dieser Vorschrift werden zunächst die vermuteten Kohlenwasserstoffe aus der Bodenprobe mit 1,1,2-Trichlortrifluorethan (TTE) kalt extrahiert und der Extrakt dann in der Küvette zwischen 4000 und 2000 cm^{-1} IR-spektroskopisch vermessen. Da TTE keine Absorptionen oberhalb von 2000 cm^{-1} aufweist, alle CH-Bindungen aber in diesem Bereich aktiv sind, kann aus ihren Absorptionen auf die Konzentration von Kohlenwasserstoffen in der Bodenprobe geschlossen werden.

2.1.2.2 *Identifizierung bekannter Substanzen mit IR-Spektren*

Ist das Referenzspektrum einer Substanz bekannt, kann diese durch die Übereinstimmung von experimentellem Spektrum und Referenzspektrum identifiziert werden. Insbesondere der Vergleich der IR-Spektren im *Fingerprintbereich*, wie er in jedem Buch über IR-Spektren erwähnt[4,5] und in den Praktikumsbüchern für Studenten angewandt wird[6,7], ist hierfür sehr aussagekräftig. Auch in der Industrie wird diese Methode häufig eingesetzt, wobei der Einsatzschwerpunkt in der Qualitätskontrolle liegt. Ein Beispiel für den Einsatz sind die Qualitätskontrollvorschriften „Identität von Festsubstanzen - Infrarotspektroskopie“ und „Identität von wasserhaltigen, flüssigen Substanzen auf Siliciumscheiben - Infrarotspektroskopie“ aus dem Werk Uetersen der Knoll AG (BASF Pharma).[8] In beiden Vorschriften wird die Identität der

Untersuchungssubstanz durch visuellen Vergleich der Infrarotspektren der Probe und des hinterlegten Referenzspektrums festgestellt.

2.1.2.3 Struktur- und Strukturparameterbestimmung mit IR-Spektren

Immer wenn eine IR-Absorption, wie z.B. diejenige von Salzsäure, H-Cl, nur von einer einzelnen Bindung abhängt, kann die Lage dieser Bande zur präzisen Messung von Bindungsstärken, aber auch Atommassen herangezogen werden. So kann beispielsweise die Bindungsstärke von ^{1}H-^{35}Cl nach (1) berechnet werden, wobei die Atommassen von ^{1}H und ^{35}Cl als bekannt vorausgesetzt werden.

$$\omega = (k/\mu)^{1/2} \tag{1}$$

ω	Winkelgeschwindigkeit $\tilde{\nu} = \omega/2\pi c$
k	Bindungsstärke oder auch Kraftkonstante der Bindung
μ	reduzierte Masse

Ist die Bindungsstärke *k* bekannt, kann sie, da sie unabhängig von den jeweiligen Isotopen ist, bei bekannten Atommassen zur Bestimmung von Bandenlagen benutzt werden z.B. für D-^{35}Cl oder, zumindest theoretisch, zur Bestimmung der Masse von Isotopen z.B. bei H-^{37}Cl. Praktisch würde dies wohl eher mit einem Massenspektrometer geschehen.

Strukturparameter zur Beschreibung der Struktur (Konformation) von Polymeren konnten James A. de. Haseth und V. E. Turula aus dem IR-Spektrum erhalten.[9] Sie entwickelten eine LC/FT-IR Kopplung bei der Polymere desolvatisiert werden und auf einen IR transparentes Trägermaterial aufgetragen werden. Die Substanz tragenden Bereiche auf dem Trägermaterial können dann Off-Line mittels IR-Mikroskopie analysiert werden. Mit dieser Technik konnten sie z.B. den Effekt der Chromatographiebedingungen auf die Sekundärstruktur von Proteinen studieren.

2.1.2.4 Identifikation funktioneller Gruppen anhand von IR-Spektren

Bei Molekülen mit mehr als zwei Atomen ist die Beziehung zwischen der Absorptionsbande im IR-Spektrum und einer Bindung im Molekül selten eindeutig, da eigentlich immer, wie quantenmechanische Rechnungen zeigen, mehr als ein Atomabstand sich im Laufe einer Molekülschwingung ändert. Allerdings wurden für viele funktionelle Gruppen typische Absorptionsbereiche im IR-Spektrum definiert und für charakteristische Banden, wie der Carbonylbande, wurden viele, teilweise erfolgreiche, Versuche unternommen, die Bandenlage mit Bindungseigenschaften zu korrelieren. So kann man an der Lage der Carbonylbande im IR-Spektrum erkennen, ob die Veresterung von 4-Methoxybenzoesäure zum Methylester erfolgt ist, da die Carbonylbande des 4-Methoxybenzoesäuremethylesters bei 1715 cm^{-1} liegt während 4-Methoxybenzoesäure bei 1755 cm^{-1} absorbiert.[10]

Dieses Beispiel zeigt, wie sich die Stärke der Carbonylbindung auf die Lage der Bande auswirkt, wenn man davon ausgeht, daß die typische Esterbande aufgrund des induktiven Effekts des Alkylsubstituenten bei 1740 cm^{-1} liegt und die Bande der Carbonsäure üblicherweise bei 1760 cm^{-1}, wenn die Carbonsäure nicht im Dimer vorliegt und die Carboxylgruppe auch sonst keine weiteren Wasserstoffbrückenbindungen aufbaut. Wasserstoffbrückenbindungen können CO-Doppelbindungen erheblich schwächen, so daß Carbonsäuredimere typischerweise im Bereich von 1725 bis 1690 cm^{-1} absorbieren. Dies gilt auch für die 4-Methoxybenzoesäure, denn die 1755 cm^{-1} für die Carbonylbande[11] wurden mittels GC/IR-Kopplung gemessen, während das zweite, offensichtlich in kondensierter Phase gemessene Spektrum der 4-Methoxybenzoesäure, die Carbonylbande bei 1690 cm^{-1} zeigt[12].

Verändert eine Reaktion größere Teile eines Moleküls oder entstehen Konstitutionsisomere, kann eine einzelne Bande keinen Aufschluß über die Produkte der Reaktion geben. In solchen Fällen muß das gesamte IR-Spektrum des oder der Produkte mit den für die Produkte zu erwartenden IR-Spektren verglichen werden.

2.1.2.5 Identifikation isomerer Reaktionsprodukte mittels berechneter IR-Spektren

Sind die IR-Spektren der erwarteten Produkte nicht verfügbar, d.h. weder in einer Datenbank gespeichert noch in einer anderen Form archiviert, was angesichts von gerade mal 100 000 IR-Spektren in der größten IR-Datenbank bei ca. 15 Millionen bekannten Verbindungen nicht unwahrscheinlich ist, müssen die notwendigen Vergleichsspektren für die mutmaßlichen Reaktionsprodukte berechnet werden. Empfohlen wurden hierzu von Scott und Radom[13] DFT-Berechnungen mit dem Becke-3-Lee-Yang-Parr-Funktional (B3-LYP) und einem 6-31 G(d) Basissatz. Das folgende Beispiel anhand der Reaktion von *trans*-1-Chlor-1-ethyl-2-methylcyclopentan **cp1** mit Base soll die Vorgehensweise einmal illustrieren und die Vor- und Nachteile aufzeigen. Die Eliminierung des Chlors nach dem zu erwartenden E2-Mechanismus kann zu den folgenden vier Verbindungen im Reaktionsgemisch führen: dem Edukt, dem *exo*-Eliminationsprodukt 1-Ethyliden-2-methylcyclopentan **cp2** sowie den beiden *endo*-Eliminationsprodukten 1-Ethyl-5-methylcyclopenten **cp3** und 1-Ethyl-2-methylcyclopenten

cp4. Dies ist ein hypothetisches Beispiel, das gewählt wurde, weil es das Problem besser verdeutlicht als eine Dehydratisierung, wo das Edukt durch die OH-Bande deutlich von den Produkten zu unterscheiden wäre. Wer denkt, Information zu diesen recht einfachen Verbindungen müßten doch in Datenbanken zu finden sein, wird enttäuscht werden. Die SpecInfo® Datenbank enthält keine der vier Verbindungen und in der Beilstein Datenbank CrossFire plus Reactions®[14] fehlt das Edukt völlig. Von den drei Produkten kennt diese Datenbank lediglich bei (E)-1-Ethylidene-2-methylcyclopentane eine Referenz für IR-Banden. Die Referenz enthält dann ganze drei Absorptionsbanden.[15]

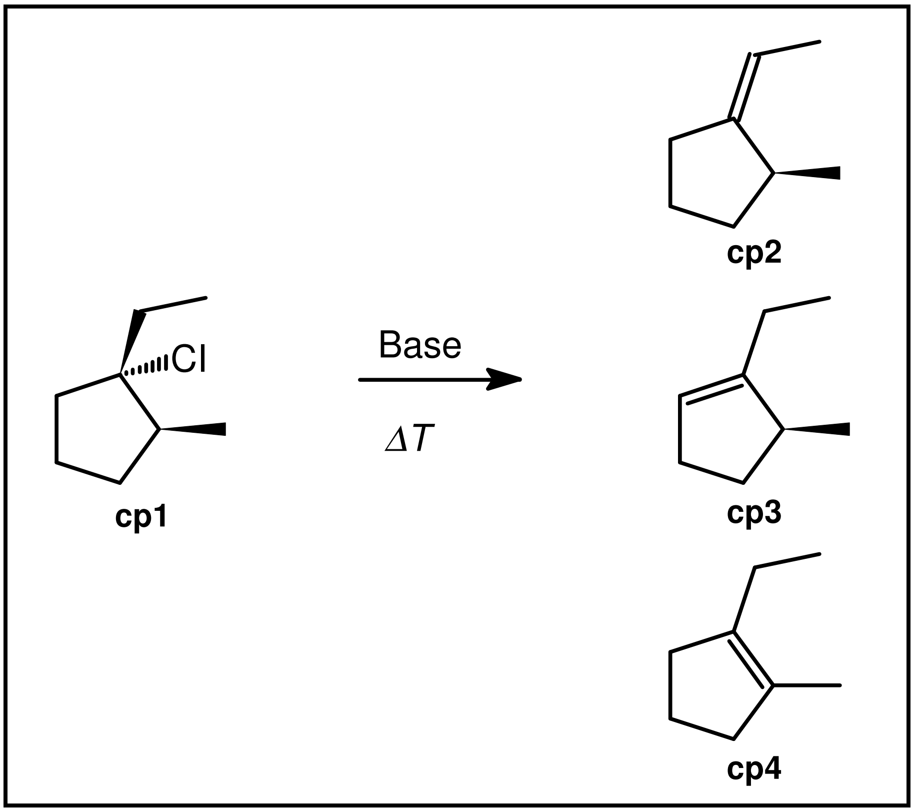
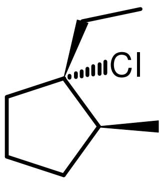


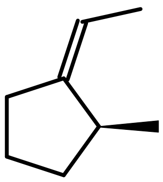
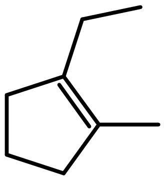
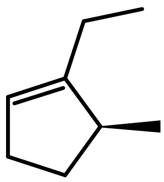
Abbildung 5: Die möglichen Reaktionsprodukte bei der Eliminierung von HCl aus *trans*-1-Chlor-1-ethyl-2-methylcyclopentan

Berechnet man nun die Spektren für die vier Verbindungen und vergleicht sie mit den Spektren der Produkte, die auf jeden Fall mittels GC/IR leicht und rein zu erhalten sein sollten, so wird aus der folgenden Tabelle die Unterscheidbarkeit der Moleküle deutlich. Dabei ist zu beachten, daß berechnete Frequenzen vergleichbarer Intensität mit wenigen cm^{-1} Unterschied in der Frequenzlage im experimentellen Spektrum schwer zu unterscheiden sein werden.

Das Edukt ist klar durch eine relativ intensive Schwingung bei 408 cm^{-1} von den drei Alkenen zu unterscheiden. Bei allen Alkenen fällt zunächst das Fehlen der Schwingung der CC-Doppelbindung auf, deren Valenzschwingung ganz eindeutig hier sehr intensitätsschwach ausfällt (die berechneten Werte liegen bei ca. 1720 cm^{-1} und die Intensität um 1.0). Dennoch werden die Cyclopentenderivate anhand des Infrarotspektrums zu unterscheiden sein. So fällt das Exo-Produkt 1-Methylvinyl-2-methylcyclopentan durch zwei sonst nicht vorhandene Fingerprintbanden bei 990 und 1332 cm^{-1} auf. 1-Ethyl-2-methylcyclopenten zeichnet sich vor allem durch das Fehlen intensiver Absorptionen unterhalb von 1229 cm^{-1} aus und das Charakteristische am Spektrum von 1-Ethyl-5-methylcyclopenten ist die Breite der CH-Bande. Die einzelnen Schwingungen liegen hier in einem Bereich von 2958 bis 3178 cm^{-1}, der damit deutlich größer ist, als bei allen anderen Verbindungen, die im Reaktionsgemisch möglich sind.

Tabelle 1: Die 20 intensivsten Schwingungen, die für die vier Cyclopentanderivate berechnet wurden (DFT, B3-LYP/ 6-311 G(d) (5D, 7F)).



Frequenz	Intensität	Frequenz	Intensität	Frequenz	Intensität	Frequenz	Intensität
408.2600	10.0801	832.9733	8.7878	1229.5549	4.4558	831.6593	8.0801
828.1166	14.9560	990.9694	6.8745	1358.8816	4.1183	859.8968	10.4809
864.4004	31.1882	1332.1884	5.6920	1420.8886	3.9711	1369.1263	7.3385
1430.6819	8.8603	1506.3033	7.5700	1501.2845	6.8719	1426.4342	5.3120
1508.8902	8.4613	1508.8694	10.5695	1515.5333	8.1343	1518.2891	7.9650
1520.5866	8.0626	1523.4182	7.2372	1522.8512	6.9865	1526.0186	5.2530
1524.7975	11.2472	2963.9006	22.1785	2983.6557	46.1016	2958.4164	34.8408
3024.1770	15.3246	3011.7761	52.9331	2948.9334	60.1224	2991.5321	37.1062
3032.8067	14.1047	3013.9683	23.4264	3004.4568	54.6779	2993.1527	58.8381
3040.3307	48.7292	3016.1769	26.6796	3009.8367	45.1401	3012.8345	11.9051
3042.8315	24.5874	3026.4611	31.8720	3026.2792	17.6788	3021.2208	21.9494
3055.2037	53.2211	3029.1116	38.5757	3029.2801	22.1504	3027.7198	43.4294
3074.5299	9.6528	3046.4501	35.5398	3036.7205	61.8679	3032.5346	31.2034
3075.8110	11.1922	3058.6130	33.2925	3044.1076	30.5787	3045.4176	64.4268
3090.8550	19.1930	3073.0958	46.7255	3051.5776	74.8247	3078.6923	13.7070
3095.5035	24.8140	3083.0958	41.9238	3058.5268	8.4118	3083.2392	81.5971
3098.8375	62.6003	3084.9621	59.2305	3085.9938	62.6264	3087.2085	39.8032
3101.9950	55.3870	3091.6739	50.8606	3090.4811	47.1375	3093.6599	44.6530
3113.9402	22.4041	3101.6500	8.2255	3096.8542	51.0482	3094.3273	45.4776
3119.1408	43.3207	3127.2834	46.1657	3103.9673	43.4542	3178.8339	22.0342

Die hochwertigen und damit teuren *ab initio*-Berechnungen, die hier zur Ermittlung der Spektren notwendig waren, die vom Zeit- und Rechenaufwand (im Mittel 36h auf einem Compute-Server Sun Sparc Server mit 8 UltraSPARC 250 MHz Prozessoren, 2GB RAM, 104 GB

HD[16]) für den Routineeinsatz gerade am GC/IR nicht tragbar sind, verhinderten bisher den häufigeren Einsatz der IR-Spektroskopie in der Analytik, trotz der von vielen namhaften Autoren beschriebenen Stärke der Infrarotspektroskopie bei der Identifikation von Molekülen.[17]

2.1.3 Ziele im Bereich der Infrarotspektroskopie

Durch die Entwicklung schneller und zuverlässige Methoden zur Vorhersage von Infrarotspektren die Nutzung von Infrarotspektren zur Strukturaufklärung zu erleichtern, ist ein Ziel dieser Arbeit. Gelingt es einigermaßen zuverlässig das Infrarotspektrum für einen beliebigen Strukturvorschlag in einer Zeit vorherzusagen, die ein interaktives Arbeiten noch zuläßt, wird der analytisch arbeitende Chemiker in die Lage versetzt durch iterative Verbesserungen seines Strukturvorschlages ein Infrarotspektrum zuverlässig aufzuklären.

Aus Entwicklung von Methoden zur Korrelation von Infrarotspektrum und 3D-Struktur einer Verbindung, wie sie für eine schnelle Vorhersage von Infrarotspektren notwendig sind, sind zudem zahlreiche Informationen über den Zusammenhang von Struktur und Infrarotspektrum zu erwarten. Diese herauszuarbeiten und darzustellen ist ein weiteres Ziel dieser Arbeit.

2.2 QSAR/QSPR - die Vorhersage der biologischen Aktivität und anderer Eigenschaften

Die pharmazeutische Industrie, durch ihre Bedeutung für die Menschheit bzw. Gesundheit zu einem bedeutenden Industriezweig geworden, steht vor dem Problem, daß die Entwicklung neuer Medikamente immer teurer wird. Dieses Problem hat zwei Ursachen, zum einen sind die einfach zu findenden Medikamente wahrscheinlich schon gefunden, was sich im Anstieg der neu zu synthetisierenden Verbindungen je neuem Medikament von 150 im Jahre 1950 auf 35.000 Verbindungen 1995 bemerkbar macht, und zum anderen wird die Anzahl der Erkrankten, die mit den neuen Medikamenten behandelt werden können oder sollen, oft kleiner, denn viele Krankheiten können heute bereits therapiert werden.

Um wieviel einfacher täte man sich, wenn direkt von der erdachten oder gar elektronisch generierten Struktur die benötigte biologische Aktivität und andere wichtige Struktureigenschaften, wie der Oktanol/Wasser-Verteilungskoeffizienten, vorhergesagt werden könnten. Bietet doch der Oktanol/Wasser-Verteilungskoeffizient einen Anhaltspunkt für die Aufnahme und die Verteilung eines Stoffes im Körper und erlaubt damit eine Aussage, ob das potentielle Medikament überhaupt an seinen Wirkort kommen könnte. Ressourcen für die Synthese neuer Medikamente könnten so viel zielgerichteter eingesetzt werden.

Moleküle sind aber dreidimensional, deshalb ist die Berücksichtigung der 3D-Struktur eines Moleküls für die Vorhersage aller von der Struktur abhängigen Moleküleigenschaften wichtig. Aufgrund des Schlüssel-Schloß-Prinzips der Enzymwirkung ist insbesondere bei der Entwicklung von Pharmaka die Berücksichtigung der 3D-Struktur eines potentiellen Wirkstoffs wichtig. So ist es wenig verwunderlich, daß die ersten Versuche zur Codierung der 3D-Struktur aus dieser Richtung stammen, man denke nur an die Arbeiten von Willett[18], Soltzberg, Wilkins[19],[20] sowie Gasteiger, Wagener und Sadowski[21]. Übrigens bildet die Arbeit von Soltzberg und Wilkins die Grundlage für die weitere Arbeit, doch dazu später mehr.

2.3 Zusammenfassung

Eine Codierung der 3D-Struktur von Molekülen hat ein enormes Potential für die Vorhersage molekularer Eigenschaften. Im Rahmen dieser Arbeit soll die zu entwickelnde 3D-Strukturcodierung insbesondere zur Vorhersage biologischer Aktivitäten sowie zur Vorhersage von IR-Spektren eingesetzt werden. Gelingt es den Zusammenhang zwischen 3D-Struktur und IR-Spektrum mittels einer 3D-Strukturcodierung zu beschreiben wäre ein wesentlicher Schritt zur vollständigen Interpretation von Infrarotspektren getan. Dies ist von besonderem Interesse, da die IR-Spektroskopie als Meßmethode, dank ihrer Universalität und Unabhängigkeit vom Aggregatzustand der Probe, ein bei weitem nicht ausgeschöpftes Potential birgt.

3 Bekannte 3D-Strukturcodierungen - eine kurze Übersicht

Speicherung und Suche chemischer Information war in der Vergangenheit vielfach der Ausgangspunkt für die Nutzung von Computern in der Chemie im allgemeinen und für die Entwicklung von Strukturcodierungen im besonderen. So wurden die ersten Grundlagen für die Handhabung chemischer Strukturinformation in Computern bereits in den fünfziger Jahren entwickelt.[22] Am Anfang standen Fragment Codes, systematische Namen und lineare Notationen wie der Smiles-String[23]. Die Nutzung von Bindungslisten (Connection Tables) zur Speicherung chemischer Informationen begann zwischen 1965 und 1970 und hat heute die anderen Formen weitgehend verdrängt. Nur der alphanumerische und lineare Smiles-String ist heute noch vor allem als sehr kompaktes und mit der ASCII-Zeichensatz kompatibles Austauschformat für Molekülstrukturen häufiger im Gebrauch.

Die digitale Speicherung und der Umgang mit der 3D-Struktur von Verbindungen am Computer in einem größeren Umfang dürfte seinen Ursprung im Jahr 1965 haben, als mit der Gründung des Cambridge Crystallographic Data Centre (CCDC) der Aufbau der wohl bekanntesten 3D-Strukturdatenbank, der Cambridge Structural Database (CSD), begann.

Die Speicherung der 3D-Struktur von Verbindungen erfolgt heute im allgemeinen in zwei Formen. Zum einen in Form eines vierspaltigen Vektors pro Atom, der die 3D-Koordinaten sowie den Atomtyp enthält, zum anderen entsprechend den Bindungslisten in Form einer Distanzmatrix, die für jedes Atom den Abstand zu allen anderen Atomen enthält.

Weder der vierspaltige Vektor noch die Distanzmatrix eignen sich für schnelle Suchen bei Datenbankabfragen. Beide Codierungen enthalten zwar die vollständige 3D-Information eines Moleküls, um aber z.B. nach einer bestimmten Distanz zwischen zwei Atomen oder gar funktionellen Gruppen in der 3D-Struktur zu suchen, sind NP vollständige Suchalgorithmen erforderlich, was der Anforderung einer schnellen Suche widerspricht.

Um eine schnelle Suche nach bestimmten Atomdistanzen innerhalb der dreidimensionalen Molekülstruktur zu ermöglichen, wurden 3D-Struktur Hashcodes entwickelt, die zusätzlich für jedes Molekül in der Datenbank berechnet und abgelegt wurden. Die 3D-Struktur Hashcodes als eine der ersten Formen einer 3D-Strukturcodierung sollen im folgenden näher betrachtet werden.

3.1 3D-Hashcodes zu Screening-Zwecken

Ein 3D-Haschcode besteht aus der Definition seines Typs und einer geometrischen Angabe. Am weitesten verbreitet sind Atom-Atom-Distanz Hashcodes, die aus den Typen der beteiligten Atome sowie einem Distanzintervall bestehen. Hierbei muß der Atomtyp nicht automatisch der Ordnungszahl des Atoms entsprechen. So schlug Willett den folgenden Aufbau für einen Hashcode vor: A(1)C(1)A(2)C(2)D. Hierbei steht A für die Ordnungszahl, C für die Connectivität der Atome und D für den Distanzintervall. Mit D = 0.01 Å erhielt Willett für 2337 Moleküle insgesamt 743749 Hashcodes. Klar, daß hier eine Reduktion notwendig ist, um eine schnelle Suche zu gewährleisten. So reduzierte Willett die Anzahl der Hashcodes auf 1500 durch Zusammenfassen von Atomtypen und Vergrößerung der Distanzintervalle.

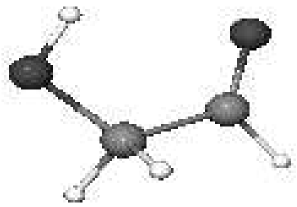
Im Anschluß an die Definition der Hashcodes, wird das Vorkommen der den Hashcodes entsprechenden Distanzen für jedes Molekül der Datenbank beispielsweise in Form eines Bitvektors abgespeichert. Damit reicht es aus, für eine schnelle erste Suche das Vorkommen der Hashcodes in der Anfragestruktur zu ermitteln und anschließend die Suche durch den Vergleich des Bitvektors der Anfragestruktur mit den Bitvektoren der Datenbankmoleküle durchzuführen. Als Suchergebnis erhält man eine Anzahl ähnlicher Strukturen.[18]

Ein Problem ergibt sich hierbei, wenn eine vollständige Übereinstimmung der Bitvektoren gefordert wird. Nicht alle Distanzen sind gleich wichtig, insbesondere größere Distanzen, die länger als drei bis vier Bindungslängen sind, können sich bei konformativ flexiblen Molekülen leicht ändern. Die Bitvektoren enthalten jedoch keinen Mechanismus dies zu berücksichtigen.

Nichtsdestotrotz sind 3D-Hashcodes eine sinnvolle Möglichkeit um eine prinzipielle Ähnlichkeit zweier Moleküle festzustellen. Ist diese festgestellt, was in der Regel nur bei wenigen Molekülen der Datenbank der Fall ist, wird ein direkter Vergleich der 3D-Strukturen notwendig, wofür sich der Vergleich der Distanzmatrizen anbietet.

3.2 Distanzmatrizen - direkte Strukturvergleiche

Eine Distanzmatrix ist eine symmetrische, quadratische Matrix, die für jedes Atom die Distanz zu allen Atomen des Moleküls enthält. Die Abbildung 6 zeigt die 3D-Struktur von Ethanal-2-ol zusammen mit den Atomkoordinaten und der Distanzmatrix.



≅

O1	-2.3517	1.4646	0.0855
C1	-1.6970	0.2671	-0.2162
C2	-0.1984	0.2524	0.0215
O2	0.4665	1.2540	0.1504
H1	-1.8428	2.1782	-0.2805
H2	0.2774	-0.7405	0.0557
H3	-2.1994	-0.4688	0.4365
H4	-1.8926	-0.0248	-1.2677

Distanz	O1	C1	C2	O2	H1	H2	H3	H4
O1	0.00	1.40	2.47	2.83	0.95	3.43	1.97	2.06
C1	1.40	0.00	1.52	2.41	1.92	2.23	1.10	1.11
C2	2.47	1.52	0.00	1.21	2.55	1.10	2.17	2.15
O2	2.83	2.41	1.21	0.00	2.52	2.01	3.19	3.04
H1	0.95	1.92	2.55	2.52	0.00	3.62	2.77	2.41
H2	3.43	2.23	1.10	2.01	3.62	0.00	2.52	2.64
H3	1.97	1.10	2.17	3.19	2.77	2.52	0.00	1.79
H4	2.06	1.11	2.15	3.04	2.41	2.64	1.79	0.00

Abbildung 6: 3D-Struktur und Distanzmatrix von Ethanal-2-ol.

Vorausgesetzt, daß die Anfragestruktur aus n_A Atomen besteht, denen $n_A(n_A\text{-}1)/2$ interatomare Distanzen assoziiert sind, die in einer Suchstruktur aus der Datenbank bestehend aus n_S Atomen mit $n_S(n_S\text{-}1)/2$ interatomaren Distanzen gefunden werden sollen, dann besteht ein einfacher Algorithmus, um eine molekulare 3D-Struktur in einer anderen zu suchen, aus den folgenden drei Schritten.

1. Wähle eine Kombination von n_A Atomen aus den n_S Atomen der Suchstruktur aus.

2. Prüfe die Kombination auf Identität mit der Anfragestruktur

3. Gehe zu Schritt eins bis alle $n_S!/(n_A!(n_S\text{-}n_A)!)$ Kombinationen getestet wurden.

Dieser Algorithmus ist äußerst rechenintensiv, da zum einen die Zahl der zu testenden Kombinationen sehr hoch ist [$nS!/(nA!(nS\text{-}nA)!)$] und zum zweiten die Identitätsprüfung im zweiten Schritt einen vollständigen Vergleich der 3D-Strukturen notwendig macht. Der erste Algorithmus, der dieses Problem NP vollständige Problem zumindest zum Teil umging und relativ schnell arbeitete, war der Lesk-Algorithmus aus dem Jahre 1979.[24] Wesentlich für seine Effizienz ist dabei der Einsatz von Bitoperationen für den Strukturvergleich. Die nachfolgenden Entwicklungen basierten vor allem auf Methoden aus der 2D-Substruktursuche, wie der „set reduction algorithm“[18,25], oder auf der Graphentheorie. Die Distanzmatrix kann dabei als vollständig verbundener Graph aufgefaßt werden mit den Atomen als Knoten und den Distanzen zwischen den Atomen als Kanten. Der Strukturvergleich kann dann als Subgraph-Isomorphismus-Problem aufgefaßt werden. Der Ullmann-Algorithmus ist ein Beispiel für einen schnellen Algorithmus, der auf dieser Basis arbeitet.[26] Der Ullmann-Algorithmus besteht im wesentlichen aus einer systematischen, gerichteten Suche (*backtracking search*) und einer Verfeinerungsprozedur für die Auswahl des nächsten Suchschritts (Ecke im Graph). Die systematisch, gerichtete Suche beginnt bei einem zufällig ausgewählten Atom der Anfragestruktur, *A(x)*, für dieses wird zunächst ein vom Typ her äquivalentes Atom in der Suchstruktur gesucht, *S(z)*. In der Verfeinerungsprozedur wird nun überprüft, ob für alle Atome der Anfragestruktur *A* deren Distanz zu *A(x)* bekannt ist, ein Atom gleichen Typs in derselben Entfernung, wie in der Anfragestruktur zu *A(x)*, auch in der Suchstruktur vorkommt. Erfüllt *S(z)* die Anforderungen der Verfeinerungsprozedur nicht, wird die Suche bei einem anderen Atom fortgesetzt. Erfüllt *S(z)* hingegen die Anforderungen, wird die Suchprozedur mit einem Nachbaratom von *A(x)* fortgesetzt. Ist in einer Ebene keine Fortsetzung mehr möglich, wird Ebene für Ebene im Suchbaum zurückgegangen. Da aber die Verfeinerungsprozedur schon viele Möglichkeiten bei der Auswahl des nächsten Atoms in der Suchstruktur ausschließt, die aufgrund fehlender Nachbaratome in der benötigten Entfernung unmöglich sind, ist der Algorithmus sehr effizient.

3.3 Distanzhistogramme und der Radialcode zwei vektorielle 3D-Strukturcodierungen

Dienten Hashcodes und die Vergleichsalgorithmen der Distanzmatrizen bisher ausschließlich der Suche nach 3D-Strukturen bzw. Strukturelementen in Datenbanken, eröffnen Distanzhistogramme und der Radialcode als vektorielle 3D-Strukturcodierungen erstmals die Möglichkeit zu einem sinnvollen und schnellen Vergleich von 3D-Strukturen. Anders ausgedrückt, dienten 3D-Hashcodes und die 3D-Substruktursuchen der Suche nach identischen Substrukturen in Molekülen, dienen Strukturcodierungen der Suche nach ähnlichen Strukturen. Identische Substrukturen werden beispielsweise bei der Suche nach Verbindungen benötigt, die ein bestimmtes Pharmakophor (für die Wirkung notwendige Substruktur) enthalten. Die Suche nach ähnlichen Verbindungen ist dann notwendig, wenn die Verbindungen ähnliche Eigenschaften haben sollen, wie beispielsweise Infrarotspektren.

Distanzhistogramme sind spätestens seit Willett[18] bekannt, dienten aber anfangs eher der statistischen Auswertung von 3D-Strukturdatenbanken. Für die Erstellung eines Distanzhistogramms wird ein Distanzbereich, *dmin* bis *dmax*, definiert und in Abschnitte aufgeteilt, dann wird die Zahl der Moleküldistanzen für jeden Abschnitt ermittelt. Distanzhistogramme zur Codierung von 3D-Strukturen wurden für die Untersuchung von Ähnlichkeitsbeziehungen u.a. von Bienfait genutzt, der sie als Interatomic Distance Histogramm (IDH) bezeichnet.[27]

Ähnlich, oder unter bestimmten Bedingungen identisch,[28] ist der von Steinhauer eingeführte Radialcode.[29] Dessen wesentlichste Eigenschaft ist, daß er wieder in die 3D-Struktur rücktransformiert werden kann. Der Radialcode basiert auf der folgenden Transformation der Distanzmatrix eines Moleküls:

$$R(r) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} Z_i Z_j e^{-B(r-r_{i,j})} \qquad (2)$$

R	Radialcode
r	Distanz, die in äquidistanten Schritten von *dmin* bis *dmax* läuft.
n	Anzahl der Atome im Molekül
i,j	Atomnummern
Z	Ordnungszahl
B	Temperatur- oder Glättungsfaktor, je größer B desto schärfer die Peaks. Ab B = 1000 entspricht der Radialcode einem Distanzhistogramm.[28]
$r_{i,j}$	Distanz der Atomen i und j

Den Radialcode konnten Steinhauer et al. bereits erfolgreich nutzen. So war mit Hilfe des Radialcodes die Korrelation von Infrarotspektren und der molekularen 3D-Struktur möglich, die dank der Decodierbarkeit des Radialcodes in einigen Fällen die Vorhersage der 3D-Struktur aus dem Infrarotspektrum ermöglichte.[30] Sowohl der Radialcode als auch Distanzhistogramme haben aber, so wie sie beschrieben werden, zwei Schwächen:

1. Molekulare Distanzen, die außerhalb des Codierungsbereiches von *dmin* bis *dmax* liegen, werden von beiden Codierungen nicht berücksichtigt.

2. Die konformative Flexibilität von Molekülen wird, bei den von den Autoren verwendeten kurzen Intervallen <= 0.1 Å für Distanzen über 10 Å, für manche Anwendungen unzureichend berücksichtigt.

Für diese beiden Punkte werden Verbesserungsmöglichkeiten im Rahmen des Ausblicks vorgestellt.

Die mögliche Verwendung elektronischer Atomeigenschaften, wie der partiellen Atomladung anstelle der Ordnungszahl in Gleichung (2), erschwert die Interpretation des Radialcodes, denn die Distanz zweier partieller Atomladungen hat eine ganz andere Aussage als die Distanz zweier Atome. Zudem wird die Rücktransformation des Radialcodes einer Verbindung in die 3D-Struktur durch Verwendung von Atomeigenschaften, die von der Atomposition im Molekül abhängen sehr erschwert. Die Rücktransformation des Radialcodes besteht in einem wesentlichen Schritt aus einem Monte-Carlo-Verfahren bei dem Atompositionen und Arten getestet werden. Hängt nun die Atomeigenschaft, welche zur Codierung verwendet wurde, von der Atomposition ab, so müßte bei jedem Iterationsschritt des Monte-Carlo-Verfahrens, diese Atomeigenschaft neu berechnet werden, was den Zeitbedarf des Verfahrens erheblich steigern würde.

Die IDH Codierung und der Radialcode wurden kurz nach dem 3D-MoRSE Code entwickelt, der im Kapitel 4 vorgestellt wird. Im experimentellen Teil der Arbeit wird ausschließlich der 3D-MoRSE Code genutzt. Dies hatte folgende Gründe:

Ziel der Arbeit ist es die Möglichkeiten des 3D-MoRSE Codes als Strukturcodierung zu beschreiben und seine Handhabung vorzustellen.

Die Verwendung der partiellen Atomladung q_{tot} erwies sich bei der Codierung der 3D-Struktur für die Simulation von Infrarotspektren als wesentlicher Schritt, damit ist der Vorteil der einfachen Interpretation des Radialcodes zumindest zum Teil verloren.

Sowohl der Radialcode als auch Distanzhistogramme berücksichtigen nach der Meinung des Autors die konformative Flexibilität von Molekülen nicht ausreichend, gerade wenn es um die Simulation von Infrarotspektren von Verbindungen geht, die einen Schwerpunkt dieser Arbeit bilden. Die Infrarotspektren von Verbindungen werden nahezu immer die Summe der Infrarotspektren vieler Konformere sein, weshalb eine Einteilung von Distanzen in Schritten von 0.1 Å oberhalb von 3 Å für nicht sinnvoll gehalten wird, zumal bereits im Ethan der Abstand der gegenüberliegenden Wasserstoffatome zwischen der verdeckten und der gestaffelten Konformation zwischen 2.29 Å und 2.55 Å schwankt.

Der Radialcode und der 3D-MoRSE Code können durch eine Fourier-Transformation ineinander überführt werden.

3.4 Die Codierung von Molekülоberflächen

Die molekulare Erkennung nach dem Schlüssel-Schloß-Prinzip von Enzym und aktiver Verbindung beruht auf der molekularen Oberfläche. Damit hängt die Wirkung potentieller Pharmaka, mindestens im ersten Schritt der Erkennung, von der Gestalt der molekularen Oberfläche ab (Findet im Rahmen der therapeutischen Wirkung eine Reaktion zwischen Pharmakon und Enzym statt, spielt neben der molekularen Oberfläche auch die chemische Struktur für die Wirkung eine wesentliche Rolle.).

Aus dem vorstehenden ergibt sich die Motivation zur Entwicklung von Codierungen für Moleküloberflächen. Ein Ansatz ist die „Comparative Molecular Field Analysis“ (CoMFA-Methode), die aus dem Vergleich bekannter Wirkstoffe versucht, strukturelle Voraussetzungen für die Wirkung festzulegen, indem geometrische Elemente im Raum mit bestimmten Eigenschaften (Wasserstoffbrücken-Akzeptor, Lipophilie etc.) definiert werden, die für eine Wirkung notwendig sind. Neben dieser Methode wurden von Gasteiger et al. die folgenden zwei Methoden entwickelt.

3.4.1 Autokorrelationsvektoren von Moleküloberflächen

Autokorrelationsvektoren molekularer Oberflächen beschreiben die Wahrscheinlichkeit, daß eine Oberflächeneigenschaft in einem bestimmten Abstand vorkommt. Anwendungen der Autokorrelationsvektoren waren unter anderem die Vorhersage der biologischen Aktivität von Steroiden mit dem elektrostatischen Potential als Oberflächeneigenschaft und die Vorhersage der Bindungskonstanten von polyhalogenierten aromatischen Kohlenwasserstoffen am cytosolischen *Ah* Rezeptor mit dem Hydrophobizitätspotential als Oberflächeneigenschaft.[31]

$$A(d_u, d_o) = \frac{1}{L}\sum_{ij} p_i \cdot p_j \qquad (d_u \leq d_{ij} \leq d_o)$$

A	Autocorrelationsvektor
d	Distanz
L	Anzahl der Oberflächenpunkte mit einer Distanz zwischen d_u und d_o
p_i, p_j	Eigenschaftswerte der Oberflächenpunkte i, j

Auf die gleiche Art und Weise ist es ebenfalls möglich Autokorrelationsvektoren von 3D-Molekülstrukturen zu berechnen.

3.4.2 Oberflächenkarten von Molekülen

Kohonenkarten oder auch selbstorganisierende Karten (self organizing maps) bieten die Möglichkeit ein multidimensionales Gebilde unter Erhalt der Nachbarschaftsbeziehungen in eine zweidimensionale Karte zu projizieren. Auf diese Weise ist es möglich, von einer Moleküloberfläche eine zweidimensionale Karte der Oberflächeneigenschaften eines Moleküls zu erhalten.[32] Diese Karten bieten sich für einen direkten optischen Vergleich an, können aber auch für einen numerischen Vergleich der Moleküloberflächen herangezogen werden, allerdings nur nach einer Überlagerung der 3D-Strukturen.[33,34]

3.5 Fazit

Neben der Möglichkeit, über geometrische Suchen dreidimensionale Substrukturen in Datenbankmolekülen zu finden, gibt es einige Ansätze zur Codierung der 3D-Struktur von Molekülen. Im selben Zeitraum wie diese Arbeit wurden im Arbeitskreis die 3D-Strukturcodierungen Radialcode und Distanzhistogramm (IDH) eingeführt. Letztere sind, aufgrund ihrer linearen Einteilung der Distanzen zwischen *dmin* und *dmax* problematisch in bezug auf konformativ flexible Moleküle. Daher wurde der nachstehend vorgestellte 3D-MoRSE Code zunächst weiter entwickelt und ausschließlich für die Experimente in dieser Arbeit genutzt. Ob sich in Zukunft der 3D-MoRSE Code oder der Radialcode bzw. Distanzhistogramme von Verbindung mit den Verbesserungen, wie sie im Ausblick dieser Arbeit vorgeschlagen werden, als 3D-Strukturcodierungsverfahren durchsetzten werden, kann im Moment nicht gesagt werden. Aussagekräftige Vergleiche zwischen den 3D-Strukturcodierungen fehlen, da bisher weder Radialcode noch Distanzhistogramme im Hinblick auf eine Aufgabenstellung voll optimiert wurden, noch der Vergleich der Codierungen für eine spezielle Aufgabenstellung eine allgemeingültige Aussage zuläßt. Daher würde auch ein allgemeiner Vergleich der drei Codierungen den Rahmen der Arbeit sprengen, wenn er überhaupt möglich ist. Die Möglichkeit den Radialcode in den 3D-MoRSE Code bzw. umgekehrt durch eine Fourier-Transformation ineinander überführen zu können, bedeutet nicht unbedingt gleiche Ergebnisse bei der Nutzung. Denn der Radialcode bewertet alle atomaren Distanzen (zwischen *dmin* und *dmax*) gleich, während der 3D-MoRSE Code eine Bewertung der atomaren Distanzen r_{ij} mit $\sin(r_{ij})/r_{ij}$ vornimmt.

4 Der 3D-MoRSE Code ein neuer 3D-Strukturcode

Der *3D-Molecule Representation of Structures based on Electron diffraction Code* ist ein neues Codierungsverfahren für die 3D-Struktur von Molekülen. Es basiert auf einer mathematischen Transformation der Distanzmatrix von Molekülen und ist damit translations- und rotationsunabhängig, somit ist eine kanonische Ausrichtung und Positionierung der Moleküle im Raum unnötig. Der 3D-MoRSE Code ist ferner von der Numerierung der Atome im Molekül unabhängig, da im Zuge der Transformation über alle Abstände im Molekül summiert wird. Im folgenden wird der 3D-MoRSE Code ausführlich abgeleitet und seine Eigenschaften detailliert diskutiert.

4.1 Die Ableitung des 3D-MoRSE Codes

Das Elektronenbeugungsbild eines Moleküls bzw. eines Elektronenstrahls an einem Molekularstrahl ist radialsymmetrisch und damit vektoriell darstellbar, in Form eines Beugungswinkels (-bereiches) und der Intensität. Ferner enthält es genug Informationen, daß die Bestimmung der 3D-Struktur eines Moleküls aus diesem Beugungsbild möglich ist. Dies ist die Idee hinter dem 3D-MoRSE Code. Das Elektronenbeugungsbild eines Moleküls enthält die 3D-Information eines Moleküls und ist vektoriell darstellbar! Damit ist der Weg der Ableitung des 3D-MoRSE Codes klar. Ausgegangen wird von der Gleichung zur Beschreibung des Elektronenbeugungsbildes eines Moleküls (3).

$$I_t(s) = K\left\{\sum_{i=1}^{N}\frac{S_i(s)+f_i(s)^2}{s^4} + 2\sum_{i=2}^{N}\sum_{j=1}^{i-1}\frac{f_i(s)f_i(s)}{s^4}F_{ij}\right\} \tag{3}$$

$I_t(s)$ Gesamtintensität der Beugungsstrahlung

s Beugungswinkel nach $s = 4\pi\frac{\sin(\theta/2)}{\lambda}$ wobei

λ Wellenlänge des Elektronenstrahls ist und

θ den Beugungswinkel angibt.

K Apparaturkonstante

f_i/s^2 Formfaktor der Elektronen des Atoms i

$S_i(s)$ nicht elastischer Streufaktor des Atoms i

F_{ij} Interferenzfunktion der Atome i und j

N Anzahl der Atome im Molekül

Der erste Term in Gleichung (3) beschreibt die Hintergrundstreuung der Probe, die auch beim Vorhandensein eines atomaren Gases derselben Zusammensetzung auftreten würde, d. h. dieser Term ist unabhängig von der 3D-Struktur der Probe. Es gilt für die Hintergrundstreuung Gleichung (4):

$$I_b(s) = K\sum_{i=1}^{N}\frac{S_i(s)+f_i(s)^2}{s^4} \tag{4}$$

Normiert man Gleichung (3) auf die Hintergrundstrahlung, erhält man den von der 3D-Struktur abhängigen Teil der Streustrahlung, der in Termen von F_{ij} für alle Atompaare definiert werden kann (5):

$$I_{3D}(s) = \sum_{i=2}^{N}\sum_{j=1}^{i-1} d_{ij}F_{ij} \tag{5}$$

F_{ij} hat hierbei nach Karle die folgende generelle Form:[35]

$$F_{ij} = \int_0^\infty P_{ij}(r)(\sin sr / sr)dr \tag{6}$$

Hierbei gibt $P_{ij}(r)$ die Wahrscheinlichkeit dafür an, daß der Abstand der Atome i und j im Intervall r bis $r+dr$ liegt. Mit der für Moleküle im Grundzustand durchaus üblichen Näherung, daß alle Molekülschwingungen harmonisch sind, kann die Interferenzfunktion F_{ij} in der folgenden Form ausgedrückt werden:

$$F_{ij} = \exp(-\langle l_{ij}^2\rangle s^2 / 2)\sin sr_{ij} / sr_{ij} \tag{7}$$

In Gleichung (7) steht der Wert $\langle l_{ij}^2\rangle$ für die mittlere quadratische Amplitude der harmonischen Schwingung zwischen den Atomen i und j. r_{ij} ist die mittlere Distanz zwischen den zwei Atomen. Gehen wir ferner davon aus, daß das Molekül starr ist, was angesichts der Tatsache, daß aus Kristallstrukturdaten und von 3D-Strukturgeneratoren auch nur starre Molekülmodelle erhältlich sind, nicht abwegig ist, können wir Gleichung (5) weiter zu Gleichung (8) vereinfachen. Mathematisch bedeutet das nichts anderes als $\langle l_{ij}^2\rangle = 0$.

$$I_{3d}(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} d_{ij} \sin sr_{ij} / sr_{ij} \tag{8}$$

Der zweite Faktor d_{ij} in der Gleichung (8) definiert Karle wie folgt (Gleichung 9):

$$d_{ij} = \frac{f_j(s) f_j(s)}{\sum_{i=1}^{N} [S_i(s) + f_i(s)^2]} \tag{9}$$

Der Faktor d_{ij} ist bis auf Spezialfälle unabhängig von s und kann durch die Näherung $f_i = fxZ_i$ ersetzt werden, wobei Z_i. die Ordnungszahl des Atoms ist und fx eine Funktion ist, die den Beugungsquerschnitt des Atoms beschreibt und mit der Ordnungszahl ansteigt. Vereinfacht man nun die Atome zu Punkten, erhält man die folgende, seit den ersten Arbeiten von Wierl[36] über Elektronenbeugung gebräuchliche Gleichung (10), für den durch die 3D-Struktur verursachten Anteil der Streustrahlung.

$$I_{3d}(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} Z_i Z_j \sin sr_{ij} / sr_{ij} \tag{10}$$

Da für Eigenschaften eines Moleküls neben der 3D-Struktur nicht nur die Ordnungszahl der Atome wichtig ist, sondern gerade wenn man die Atome zu Punkten vereinfacht, auch die Ladung, Elektronegativität, der van der Waals Radius oder die Gruppenzugehörigkeit im Periodensystem (um nur einige Beispiele zu nennen) wurde für den 3D-MoRSE Code die Ordnungszahl, Z_i, in (10) zur Atomeigenschaft A_i verallgemeinert, so daß mit (11) die allgemeine Formel für den 3D-MoRSE Code erhalten wird.

$$I_M(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin sr_{ij}}{sr_{ij}} \tag{11}$$

Diese Verallgemeinerung zu einer vom Benutzer auszuwählenden Atomeigenschaft ermöglicht die flexible Anpassung des 3D-MoRSE Code an die jeweilige Problemstellung. Wie in den folgenden Kapiteln beschrieben, erwies es sich beispielsweise als günstig für die Korrelation von 3D-Struktur und IR-Spektrum als Atomeigenschaft im 3D-MoRSE Code die partielle Gesamtladung des Atoms zu verwenden. Verständlich, wenn man bedenkt, das die Stärke einer IR-Bande vom Übergangsdipolmoment abhängt.

Ein allgemeiner Atomparameter A_i muß nicht, kann aber durchaus auch aus einer Kombination von Atomeigenschaften bestehen. Sollen Atomeigenschaften kombiniert werden, ist jedoch ihre Vergleichbarkeit Voraussetzung. Damit ist es erforderlich, die Atomeigenschaften für die Verwendung im 3D-MoRSE Code zu skalieren. Die im Rahmen dieser Arbeit verwendeten Skalierungsfaktoren, die größtenteils aus den Minimal- und Maximalwerten von Atomeigenschaften der in der SpecInfo Datenbank enthaltenen Verbindungen errechnet wurden, finden sich in Anhang 2: Atomeigenschaften und Skalierungsfaktoren.

4.1.1 Der 3D-MoRSE Code und innermolekulare Distanzen

Eine flexible Anpassung des 3D-MoRSE Codes an das Problem über die verwendete Atomeigenschaft ist gut, aber bis jetzt ist weder ein Wertebereich für s festgelegt noch die Bedeutung definiert. Dies soll im folgenden geschehen.

Als Soltzberg und Wilkins 1976/77 Gleichung (10) in einer stark vereinfachten Form erstmals zur Beschreibung von Molekülen verwendeten[37], benutzten sie die Null-Durchgänge von $I(s)$ in 100 äquidistanten Abschnitten zwischen 1 und 31 $Å^{-1}$ zur Codierung von Sedativa. Arbeiten von Náray-Szabó[38,39] definieren eine molekulare Distanz auf Basis der Gleichung (8), zu deren Berechnung über s von 0 bis unendlich integriert wird und somit kein Wertebereich für s festgelegt wird.

Die Bedeutung des Wertebereichs von s erschließt sich über die Definition von s. s ist ein Beugungswinkel und als solcher von der Spaltbreite, bzw. im Fall der Elektronenbeugung vom Abstand zweier Atome abhängig. Die Beugung eines Strahls an einem Spalt ergibt eine radialsymmetrische, sinusförmige Intensitätsverteilung. Der 3D-MoRSE Code liefert für $N = 2$, sprich für zwei Atome, eine vergleichbare Intensitätsverteilung in Abhängigkeit von s. Abbildung 7 zeigt dies für zwei Wasserstoffatome mit $A_i = 1$ mit einem Abstand von 0.74 Å. Zur Beschreibung einer Sinusfunktion muß mindestens ein Maximum beschrieben werden, was hier durch das Maximum von $I(0) = 1$ bis zum Null-Durchgang bei 4.24 $Å^{-1}$ gewährleistet wird.[i] Würde der Ausschnitt weniger als ein Maximum enthalten, wären auch andere Sinusfunktionen denkbar, die dem gezeigten Kurvenverlauf folgen.

---

i Die Funktion sin(x)/x zeigt vom Grenzwert (0,1) zum ersten Nulldurchgang nur ein halbes Maximum aufgrund des Effekts von x->0.

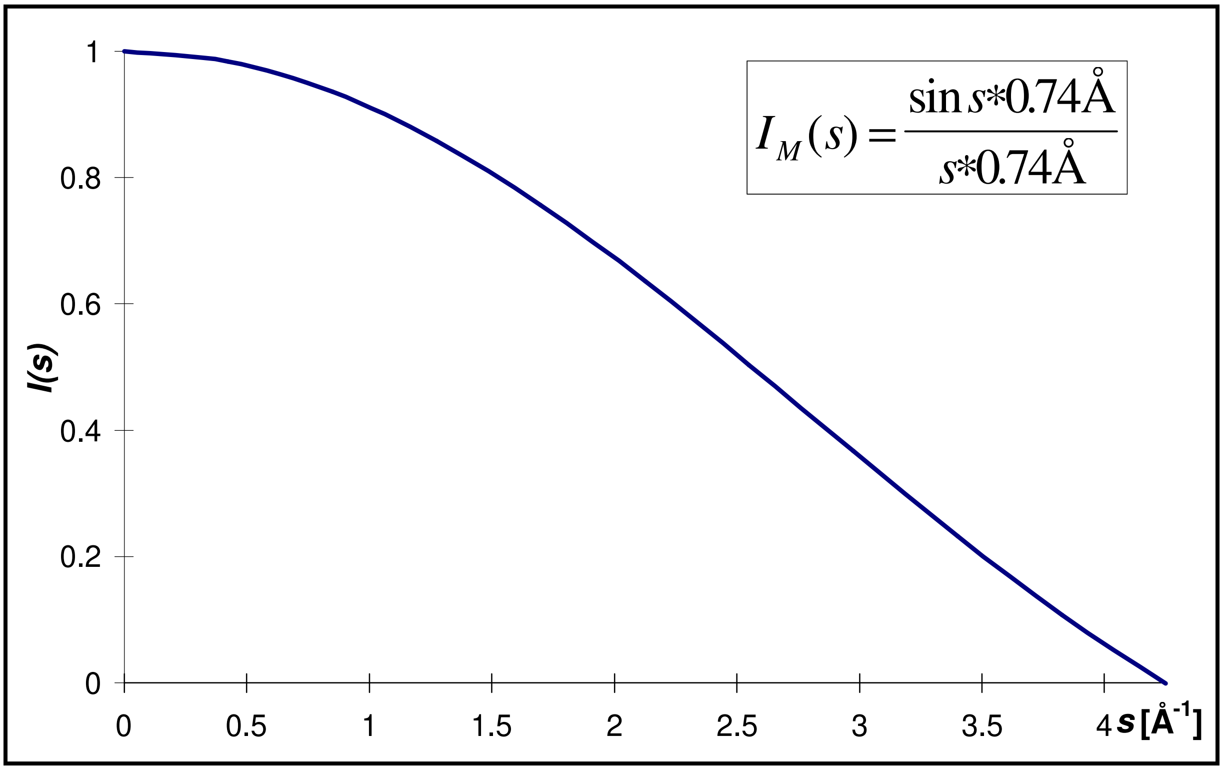


Abbildung 7: 3D-MoRSE Code von Wasserstoff am Auflösungslimit für Bindungslänge von Wasserstoff.

Dies bedeutet nichts anderes, als daß die minimale durch den 3D-MoRSE Code beschreibbare Distanz, r_{min}, durch Gleichung (12) gegeben ist, wenn $s_{min} = 0.0$ $Å^{-1}$ ist.

$$r_{min} = \frac{\pi}{s_{max}} \tag{12}$$

Dabei kann $I(0)$ mittels des Grenzwertes $\lim_{x \to 0} \sin x / x = 1$ berechnet werden. Interessant ist, daß r_{min} auch gleich dem kleinsten beschreibbaren Distanzunterschied ist. Abbildung 8 zeigt dies anhand von zwei Codepaaren mit Atomabständen von 1.39 bzw. 1.54 Å und 1.99 bzw. 2.14 Å. Es galt $A_i = 1$.

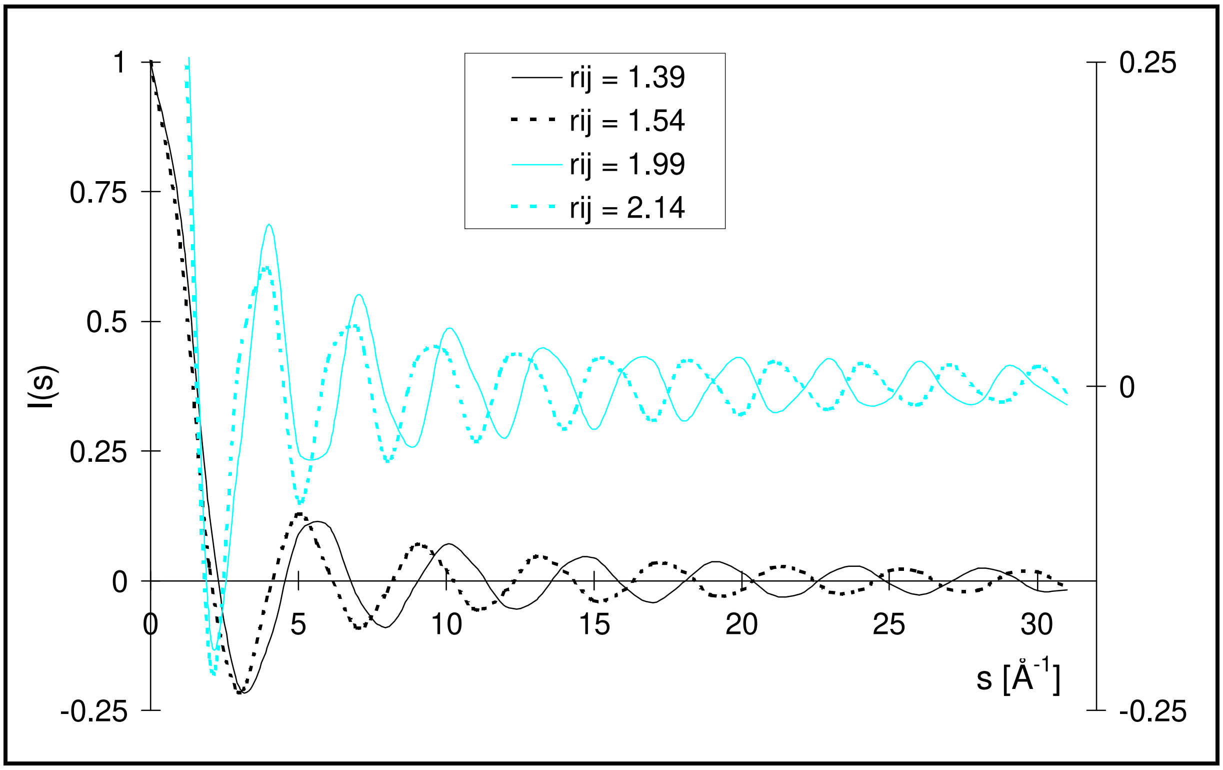


Abbildung 8: 3D-MoRSE Codes mit von zwei Atompaaren mit ein Distanzdifferenz von 0.15 Å, bei s_{max} = 31 Å^{-1} und daraus resultierend einem r_{min} = 0.1 Å. Die Unterschiede der Codes sind deutlich zu sehen, so unterscheidet sich die Zahl der Maxima jeweils um 1.

Aus einer ähnlichen Überlegung heraus läßt sich auch die größte beschreibbare Distanz r_{max} festlegen. Auch die von r_{max} stammende Beugungsfunktion sollte durch mindestens zwei Punkte beschrieben werden, deren Distanz die halbe Breite eines Maximums nicht unterschreitet. Der kleinste Abstand zweier Punkte im 3D-MoRSE Code ist Δs. Damit ergibt sich für r_{max}:

$$r_{max} = \frac{\pi}{\Delta s} = \frac{n\pi}{s_{max}} \tag{13}$$

n Anzahl der Codewerte

Die folgende Abbildung mit den 3D-MoRSE Codes von zwei Atompaaren am Rande des Auflösungslimits verdeutlicht das Gesagte. So beträgt r_{max} bei Δs = 0.3 Å^{-1} gut zehn Angström (10.5 Å) und wie Abbildung 9 zeigt, führt schon eine Überschreitung von r_{max} um 0.5 Å zum Verschwinden von Maxima.

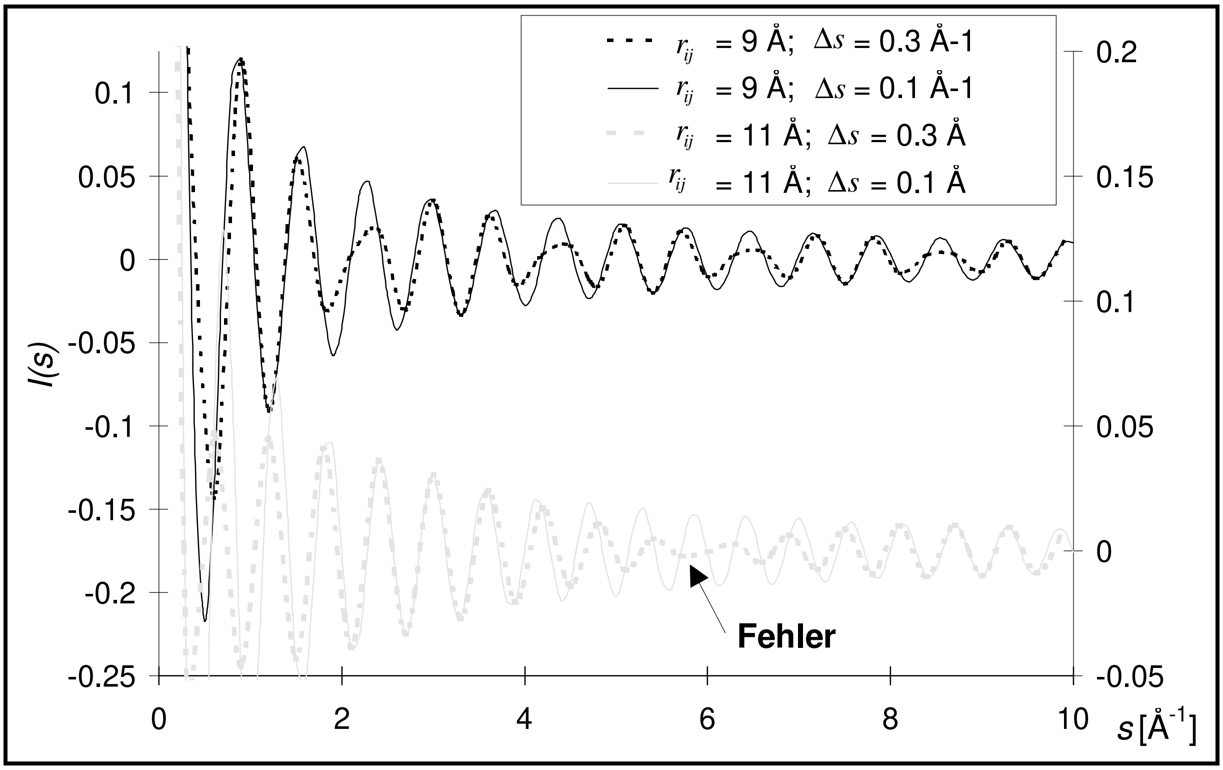


Abbildung 9: Das Verhalten des 3D-MoRSE Codes am Auflösungslimit (gestrichelte Linien). Während knapp unterhalb des Auflösungslimits noch alle Maxima zu sehen sind, fehlt nur 0.5 Å oberhalb des Auflösungslimits bereits ein Maximum (siehe Marke).

Mit den Gleichungen (12) und (13), die den Zusammenhang zwischen dem Bereich von s bzw. dem Wert von s_{max}, der Anzahl der Codewerte und den Auflösungslimits des Codes r_{min} und r_{max}, definieren, ist die Bedeutung dieser Parameter des 3D-MoRSE Codes geklärt. Festzuhalten ist, das r_{min} und r_{max}, über s_{max} miteinander in Beziehung stehen, so daß es nicht möglich ist, unabhängig voneinander, die Auflösung für kleine oder große Distanzen zu verbessern.

4.1.2 Skalierung des 3D-MoRSE Codes

Abbildung 9 mit dem unskalierten 3D-MoRSE Codes eines Atompaares zeigt die Abnahme der Werte des 3D-MoRSE Codes mit $1/s$. Es ist aber für die Verwendung eines Korrelationsverfahrens notwendig, daß alle Eingabewerte dieselbe Größenordnung haben. Daher ist eine Skalierung/Normierung der Werte des 3D-MoRSE Codes notwendig. Da Normierungsverfahren, wie die Normierung der Vektorlänge auf eins,[i] immer einen Informationsverlust bedeuten und

i Bei der Normierung der Vektorlänge auf eins bleibt zwar die Information über die Richtung des Vektors erhalten aber die Information über den Betrag (Länge) des Vektors geht verloren.

die bekannten Skalierungsverfahren nur in bezug auf einen Datensatz arbeiten, was für jeden erzeugten Datensatz andere Skalierungsfaktoren bedeuten würde und einen Vergleich von 3D-MoRSE Codes zwischen verschiedenen Datensätzen ausschließt, wurde entschieden, bezüglich der Skalierung der Werte des 3D-MoRSE Codes, einen neuen Weg zu beschreiten.

Da die Skalierungsfaktoren von der Wahl der Codierungsparameter A_i, s_{max} und n abhängen werden, die sich bei einer Optimierung für eine Fragestellung immer wieder ändern, ist ein relativ schnelles Skalierungsverfahren notwendig, das aber unabhängig von der Fragestellung, immer dieselben Skalierungsfaktoren liefern soll (Vergleichbarkeit der Codes).

Skalierungsfaktoren können aber nur anhand eines Datensatzes festgelegt werden. Daraus wurde die Idee geboren, die Skalierungsfaktoren anhand eines kleinen, standardisierten Datensatzes, des Standarddatensatzes, zu berechnen. Die Verwendung eines Standarddatensatzes, wenn er in etwa die Vielfalt des Einsatzgebietes der Codierung beschreibt, genügt zum einen den vorher gestellten Anforderungen, zum anderen erlaubt er zwischen wichtigen und unwichtigen Abweichungen im Code zu unterscheiden. Normalerweise wird bei der Skalierung eines Codes jeder Codewert auf denselben Wertebereich skaliert. Ist nun aber ein Codewert in einem Datensatz beispielsweise bis auf die numerische Ungenauigkeit konstant, da er eine konstante Eigenschaft des Datensatzes repräsentiert, beispielsweise den Benzolring in einem Datensatz der nur aus Benzolderivaten besteht, würde bei einer normalen Skalierung diese numerische Ungenauigkeit auf den im Rahmen des Skalierungsverfahrens festgelegten Wertebereich skaliert werden. Nicht so bei der Verwendung eines Standarddatensatzes der neben Benzolderivaten auch noch andere Verbindungen enthält, hier bleibt der Codewert für den Benzolring konstant.

Natürlich kann die Verwendung des Standarddatensatzes (Anhang 1 Standarddatensatz) nicht garantieren, daß alle Codewerte im Skalierungsbereich des Standarddatensatzes liegen. So zeigte sich im Rahmen der Arbeit, daß relativ häufig die Grenzen von -1 und 1, die für die Skalierung nach der Min-Max-Methode gewählt wurden, überschritten werden. Dennoch fielen bisher noch keine Codewerte auf, die den Wertebereich um mehrere Zehnerpotenzen verfehlten. Insgesamt kann man die mit der Skalierungsmethode des Standarddatensatzes gemachten Erfahrungen als durchweg positiv beurteilen. Was nicht heißen soll, daß der Standarddatensatz, dessen Moleküle frei erfunden sind, deren Moleküleigenschaften aber im Rahmen der verwendeten Methoden berechenbar sind, nicht verändert oder verbessert werden könnte.

Alle nachstehend verwendeten 3D-MoRSE Codes sind, falls nicht gesondert erwähnt, nach dieser Methode skaliert. Die folgende Tabelle zeigt die nach der vorstehenden Methode erhaltenen Skalierungsfaktoren für den 3D-MoRSE Code mit folgenden Codierungsparametern: A_i =1, s_{max} =31 Å^{-1} und n = 32. Tabellen mit den Skalierungsfaktoren für alle im Rahmen dieser Arbeit verwendeten Codierungsparameter finden sich in Anhang (Kapitel 10.3).

Tabelle 2: Skalierungsfaktoren für den 3D-MoRSE Code mit den Codierungsparametern: A_i =1, s_{max} =31 Å^{-1} und n = 32. Die Codewerte sind erst mit zugehörigen Multiplikationsfaktor zu multiplizieren und dann ist der Summand zu addieren.

Wert	Multiplikationsfaktor	Summand
0	0.001899	-1.053181
1	0.059632	-1.528429
2	0.273005	1.144434
3	0.380064	0.020969
4	0.442235	1.262159
5	0.234176	1.166447
6	0.368331	-1.029467
7	0.662819	-0.963404
8	0.355073	-0.541738
9	0.482703	-0.409363
10	0.574144	0.419264
11	0.567202	0.773683
12	1.080876	-0.067486
13	0.60701	-0.374062
14	0.755747	-1.032286
15	1.334685	0.465189
16	1.122697	-0.552876
17	0.681654	0.965193
18	0.832126	0.718003
19	0.394128	-0.956607
20	0.673101	0.22396
21	0.691088	0.567546
22	1.919132	-0.472361
23	1.081566	0.912867
24	3.24794	-0.346346
25	0.876406	-0.956823
26	2.135839	0.237133
27	1.858686	-0.115191
28	1.089327	0.15254
29	0.967769	0.89079
30	2.090965	-0.044209
31	1.154323	-0.882405

4.2 Der 3D-MoRSE Code und molekulare Bewegungen

Die Auswirkungen molekularer Bewegungen, wie Translation, Rotation, Schwingung einer Bindung und Konformationsänderungen, auf den 3D-MoRSE Code soll im folgenden anhand von Beispielen diskutiert werden.

4.2.1.1 Translation und Rotation des Moleküls

Da Translation und Rotation des gesamten Moleküls, die zur Codierung verwendeten zwischenatomaren Abstände nicht beeinflussen, ist der 3D-MoRSE Code rotations- und translationsinvariant. Abbildung 10 zeigt dies an einem Beispiel.

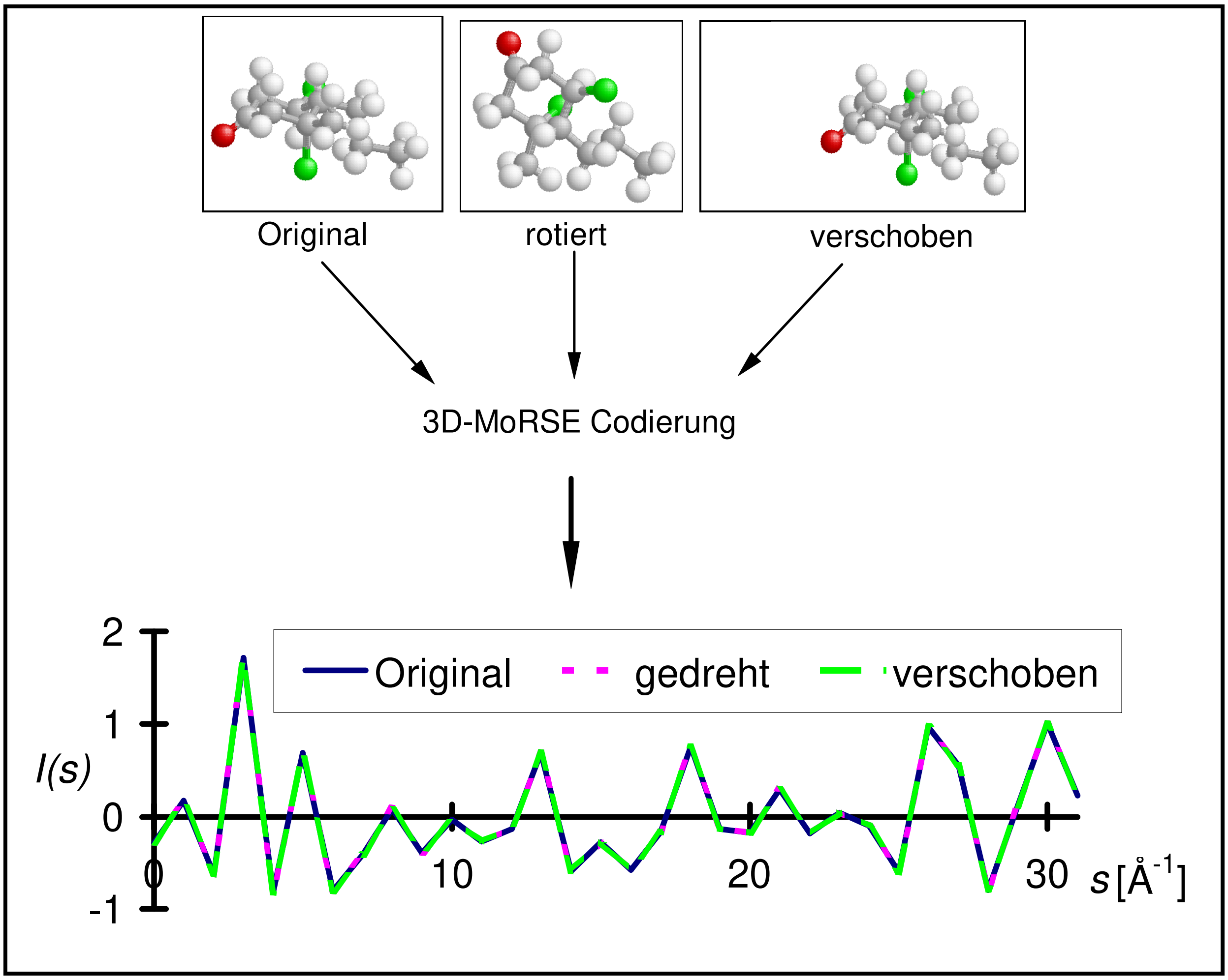


Abbildung 10: Invarianz des 3D-MoRSE Codes gegenüber Rotation und Translation.

4.2.1.2 Verlängerung einer Bindung

3D-Koordinaten enthalten die Information über eine Bindung nur implizit in Form der beteiligten Atome und des Atomabstandes. Bindungslängendifferenzen unter einem Pikometer sind in der Regel nicht signifikant und kaum experimentell meßbar (Röntgenstrukturen haben eine

Auflösung von maximal rund 1 Å = 100 pm bei Proteinen, bei kleineren Molekülen kann die Auflösung theoretisch bis in den Bereich der Wellenlängen der benutzten Röntgenstrahlung gesteigert werden).

Variationen einer Bindungslänge im Bereich von etwa 1 pm bis 5 pm kennzeichnen meist Einflüsse von Substituenten auf eine Bindung, während sich im Bereich darüber zumeist die Bindungsordnung ändert. Längendifferenzen von über 100 pm bzw. die Zunahme einer Bindungslänge um 1 Å wird in der Regel als Zeichen für eine brechende Bindung gelten können und bei einer Distanz über 4 - 5 Å zwischen einen Atom und seinem nächsten Nachbarn werden in der Regel die Atomorbitale nicht mehr überlappen.

Der 3D-MoRSE Code sollte in der Lage sein, die chemischen Auswirkungen einer Bindungslängendifferenz wiederzugeben. Dabei ist zu beachten, daß die Übergänge zwischen den einzelnen Bindungskategorien sicherlich fließend sind.

Anhand der mit AM1/HyperChem[40] optimierten Struktur von 6-Chlor-6-methylheptanon-2 soll die Auswirkung einer Änderung der C-Cl Bindungslänge auf den 3D-MoRSE Code untersucht werden. Das Beispiel wurde gewählt, da sich hier der C······Cl Atomabstand im Rahmen einer S_{N1}-Reaktion auch tatsächlich ändern könnte und zudem die flexible Alkanonkette das Studium von konformativen Einflüssen auf den 3D-MoRSE Code zuläßt, die im nächsten Kapitel diskutiert werden sollen.

Um die Auswirkung von Bindungslängendifferenzen auf den 3D-MoRSE Code zu studieren, wurde ausgehend von der optimierten Struktur die C-Cl Bindungslänge editiert, während die übrige Struktur nicht geändert wurde. Es ergibt sich bei der Verwendung folgender Parameter für den 3D-MoRSE Code: n = 32, s_{max} = 31 Å^{-1} und A_i = 1 folgendes Bild: der 3D-MoRSE Code für eine Bindungslänge von 10 Å weicht an mehreren Stellen signifikant vom Code der mit AM1 optimierten Struktur ab, die eine Bindungslänge von 1.806 Å aufweist. Eine vielleicht erwartete deutlichere Änderung des Codes gibt es nicht und sollte es auch nicht geben, denn mit dieser Bindungslänge wurden nur 24 der 300 Distanzen im Molekül geändert, da es sich um eine terminale Bindung handelt.

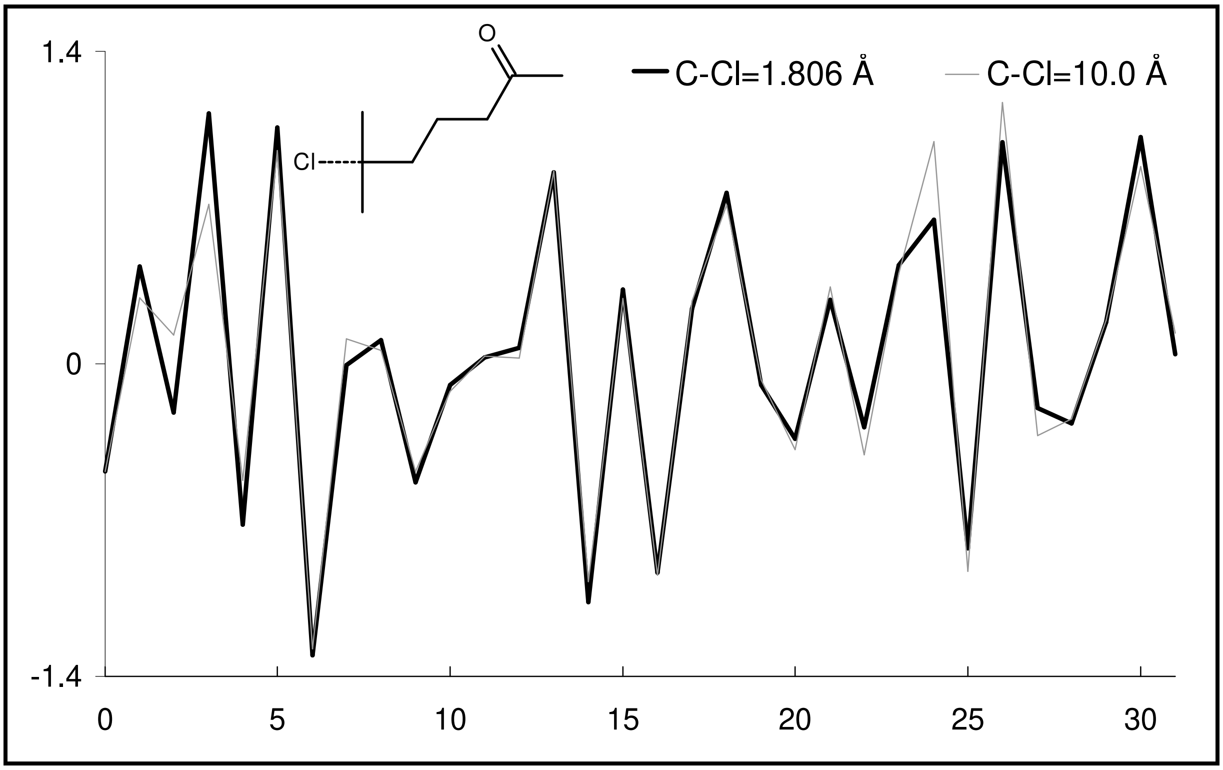


Abbildung 11: Verhalten des 3D-MoRSE Codes beim Bruch der terminalen C-Cl Bindung in 6-Chlor-6-methyheptanon-2. Deutlich sind die kleinen aber signifikanten Abweichungen des mit n = 32, s_{max} = 31 Å^{-1} und A_i = 1 berechneten Codes bei einer Atomabstand C·····Cl von 10 Å zu erkennen.

Wie Abbildung 11 zeigt, schlägt sich der Bruch der terminalen C-Cl Bindung im 6-Chlor-6-methyheptanon-2 im 3D-MoRSE Code nieder, damit ist zunächst die Grundforderung erfüllt, daß eine solche Änderung im Code sichtbar sein muß. Im vorstehenden waren aber noch weitere Anforderungen an den Code, bezüglich der Abbildung von Bindungslängendifferenzen, gestellt worden. Abbildung 12 und Abbildung 13 zeigt, daß diese erfüllt werden. Hierbei zeigt sich, daß der 3D-MoRSE Code bei gebrochener Bindung, wie gefordert, im Prinzip konstant ist und die „weitere Diffusion" der Bindungspartner von 6.0Å auf 10.0 Å keinen Einfluß auf den Code hat.

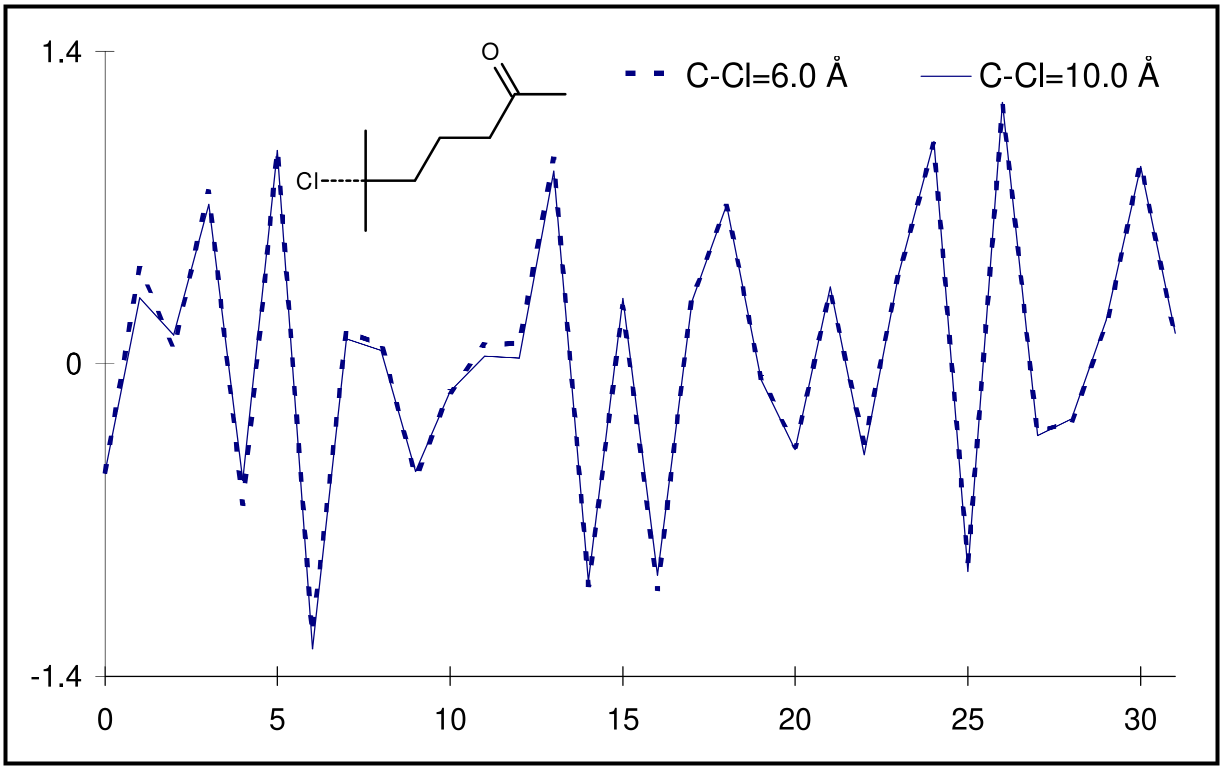


Abbildung 12: Der 3D-MoRSE Code von 6-Chlor-6-methylheptanon-2 bei gebrochener C-Cl Bindung (Atomabstände von 6 bzw. 10 Å).

Die geforderte Konstanz des Codes bei Bindungslängendifferenzen von unter 0.01 Å wird gefunden (C-Cl Bindungslängen 1.806 und 1.800 Å). An den Stellen A und B (Abbildung 13) ist deutlich der Effekt von Bindungslängendifferenzen im Bereich von 0.2 Å zu sehen, die häufig einer Änderung der Bindungsordnung oder einem starken Einfluß von Substituenten auf die Bindung entsprechen. Damit kann folgendes gezeigt werden:

- Der 3D-MoRSE Code reagiert nicht auf insignifikanten Abstandsänderungen (1.806 und 1.800 Å).
- Der 3D-MoRSE Code zeigt eine deutliche Reaktion auf Bindungslängedifferenzen, die einer Änderung der Bindungsordnung entsprechen.
- Der Einfluß der Abstandsänderung C·······Cl von 6 auf 10 Å auf den 3D-MoRSE Code ist trotz der hohen prozentualen Abstandsänderung, wie gefordert, gering.

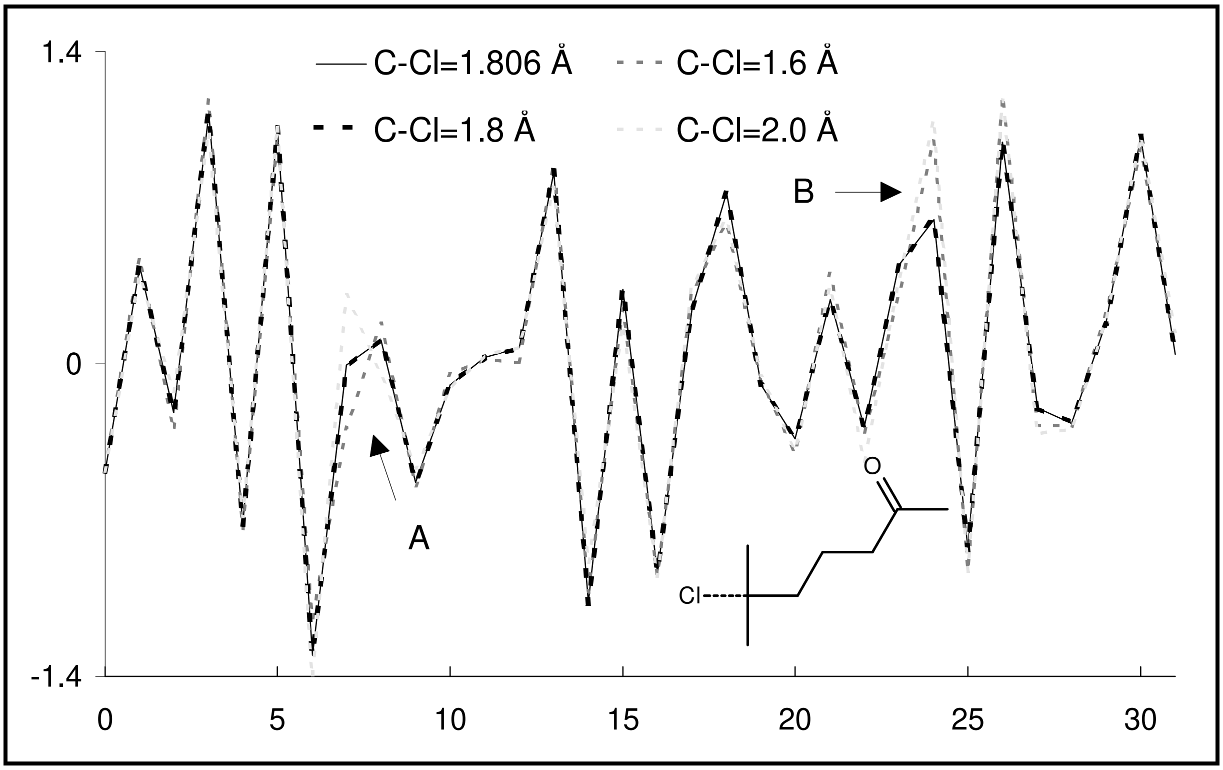


Abbildung 13: Die Konstanz des 3D-MoRSE Codes insignifikanten Änderungen des Bindungsabstandes (die Werte für 1.800 und 1.806 Å) und die deutliche Reaktion des 3D-MoRSE Codes auf Bindungslängendifferenzen im Bereich von Bindungsordnungsänderungen (Werte für 1.600, 1.806 und 2.000 Å für die gestrichelte Bindung)

Ein Bindungsbruch in der Kohlenstoffkette, sollte eine weit größere Auswirkung auf den Code haben, da davon alle interatomaren Abstände zwischen den beiden Fragmenten betroffen wären und die bisherige 3D-Struktur in zwei gleichgroße Fragmente zerbrechen würde, statt wie im Fall des Bruchs der C-Cl Bindung im wesentlichen erhalten zu bleiben. Abbildung 14 zeigt das Verhalten des 3D-MoRSE Codes in diesem Fall.

Die Änderungen im 3D-MoRSE Code fallen hier stärker als im Fall der C-Cl Bindung aus. Dennoch ändert sich der Code nicht vollständig, da bei den gewählten Randbedingungen der maximale Abstand auf 3.1 Å begrenzt ist, bei dem sich noch Auswirkungen auf den Code ergeben. Da aber die maximale Distanz im Molekül im Bereich von 10 Å liegt, fallen die sich ändernden Distanzen zwischen den Molekülenden aus der Betrachtung raus. Auch die minimale Distanz r_{min} = 0.1 Å wirkt hier entsprechend. So ergibt sich für die Verringerung des berechneten Abstandes von 1.52614 Å auf 1.50 Å (0.026 Å) noch keine wesentliche Änderung im Code während diese bei einer Verlängerung auf 1.60 Å (0.074 Å) schon deutlicher Ausfallen.

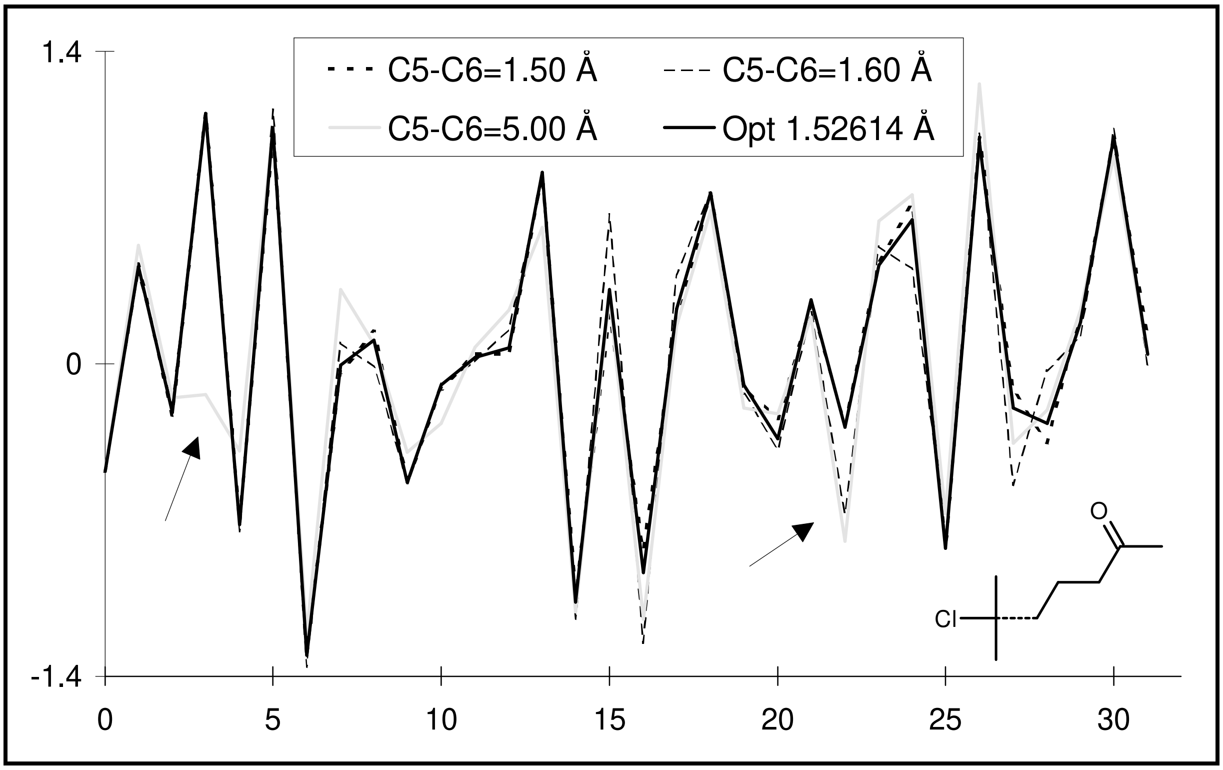


Abbildung 14: Das Verhalten des 3D-MoRSE Codes bezüglich einer veränderten Bindungslänge zwischen C5 und C6. Berechnet mit denselben Parametern wie sie für Abbildung 12 genutzt wurden.

Senkt man nun s_{max} auf 15.5 $Å^{-1}$ und verlängert damit r_{max} auf 6.2 Å und benutzt zusätzlich die Ordnungszahl als Atomparameter, womit sich die Bedeutung der vielen Distanzen zu den Wasserstoffatomen verringert und die Bedeutung auch gerade der gebrochenen CC-Bindung steigt, findet man im Anfangsbereich weit deutlichere Änderungen des 3D-MoRSE Codes als in den vorstehenden Beispielen (vgl. Abbildung 15).

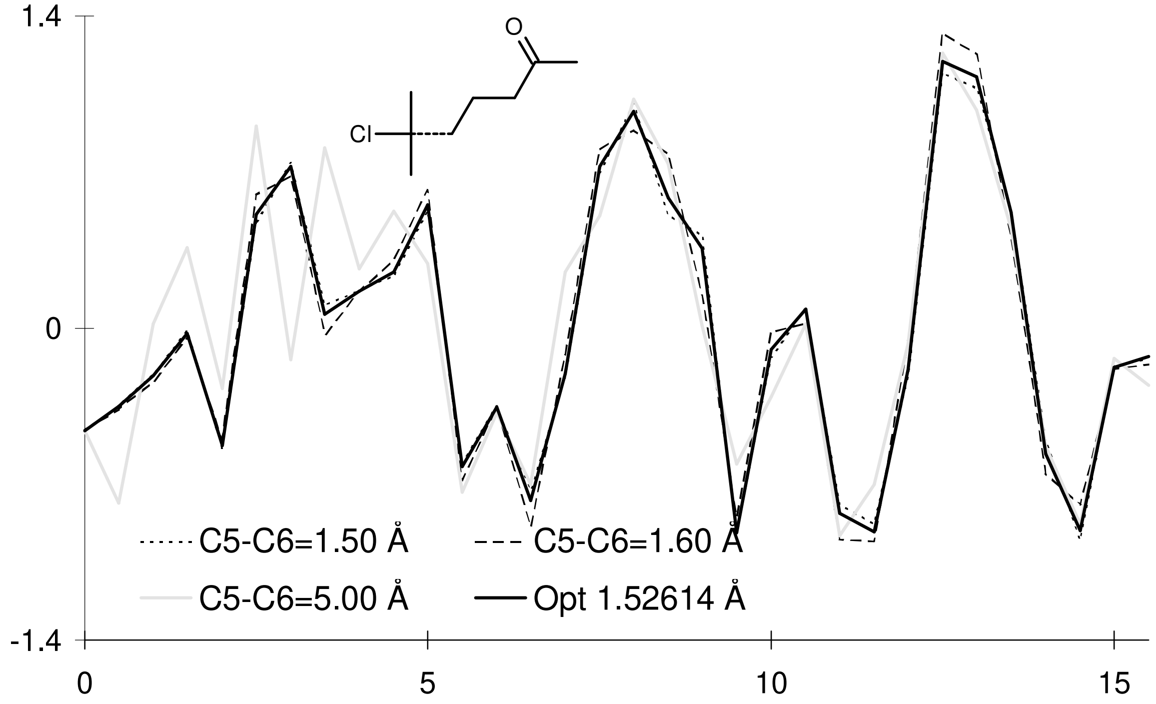


Abbildung 15: Das Verhalten des 3D-MoRSE Codes beim Bruch der Bindung zwischen C5 und C6 von 6-Chlor-6-methylheptanon-2. Der Code wurde hier mit n = 32, s_{max} = 15.5 $Å^{-1}$ und A_i = Ordnungszahl berechnet, um längere Distanzen einzubeziehen und zudem Distanzen zwischen nicht Wasserstoffatomen zu betonen. Die gebrochene Bindung ist gestrichelt gezeichnet.

4.2.1.3 Konformationsänderungen

Ändert sich ein Torsionswinkel im Molekül und damit die Konformation, ändern sich die Distanzen zwischen den Molekülteilen, deren Atome weiter als eine Bindung von der Bindung entfernt sind, deren Torsionswinkel sich geändert hat. Vereinfacht kann man in der organischen Chemie davon ausgehen, daß eine Rotation von Molekülteilen um Einfachbindungen, sofern sie nicht Teil von Ringsystemen sind, frei erfolgen kann. Aus diesem Grund werden Konformationsisomere zumindest im flüssigen und gasförmigen Zustand in einer Mischung entsprechend der Boltzmannverteilung nebeneinander vorliegen und das einzelne Molekül die Möglichkeit haben, zwischen den Konformationen überzugehen. Wenn die Eigenschaften des Gemisches an Konformationsisomeren Gegenstand einer Untersuchung sind, wie es beispielsweise bei der Vorhersage von Infrarotspektren in kondensierte Phase der Fall ist, sollte eine Codierung der 3D-Struktur eines Moleküls nur im begrenzten Umfang von der Konformation abhängen. Um dies zu überprüfen, soll im folgenden das Verhalten des 3D-MoRSE Codes gegenüber einer Rotation um die Bindung zwischen C5 und C6 in dem schon bekannten Beispielmolekül 6-

Chlor-6-methylheptanon-2 untersucht werden. Im optimierten Zustand beträgt der Torsionswinkel hier exakt -177.6°. Um hieraus ein geringfügig unterschiedliches Rotamer zu erhalten wurde dieser Winkel, und zwar nur dieser Winkel, auf -150° geändert.

Abbildung 16 zeigt die beiden Rotamere im Vergleich.

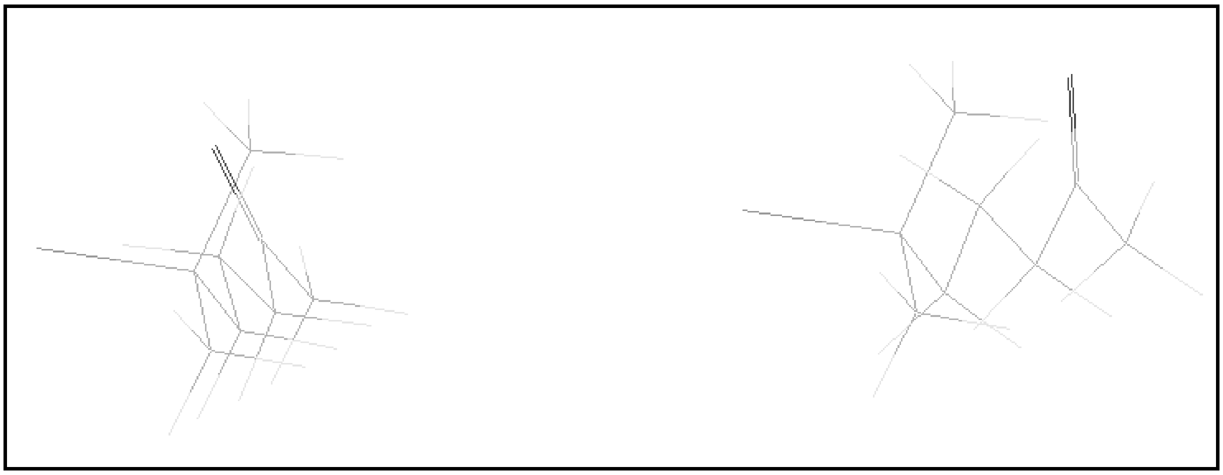


Abbildung 16: Die zwei Rotamere des 6-Chlor-6-methylheptanons-2, die für die folgenden Untersuchungen verwendete wurden. Links die mit AM1 optimierte Struktur mit dem Torsionswinkel von -177.6° rechts das Rotamer in dem dieser Torsionswinkel auf -150° geändert wurde.

Der 3D-MoRSE Code kann nicht unbeeinflußt von einer Konformationsänderung bleiben, da diese die intramolekularen Abstände beeinflußt (Es sei denn r_{max} wird auf den Bereich einer Bindungslänge beschränkt). Wie Abbildung 17 aber zeigt, kann der Grad des Einflusses der Konformationsänderung auf den 3D-MoRSE Code variiert werden.

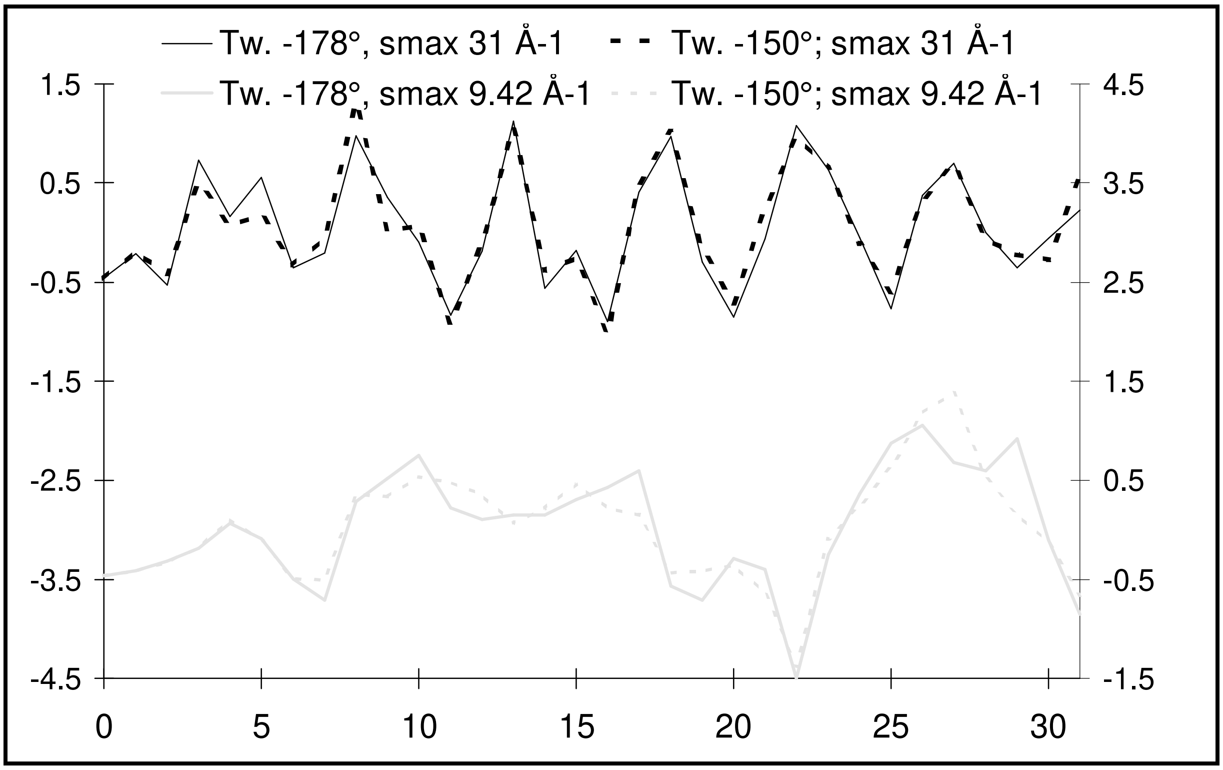


Abbildung 17: Der Einfluß des geänderten Torsionswinkels aus Abbildung 16 auf den 3D-MORSE Code bei unterschiedlichen Werten von s_{max} bzw. r_{max} (gleiche Anzahl der Werte, n = 32, A_i = Ordnungszahl). Die Werte für den 3D-MoRSE Code mit s_{max} = 31 $Å^{-1}$ wurde auf linke y-Achse Aufgetragen, der *rms*-Wert der Codes für die beiden Konformationen beträgt 0.17. Die Werte für *smax* = 9.42 $Å^{-1}$ sind hellgrau und auf der rechten y-Achse dargestellt, der *rms*-Wert der Codes beträgt hier 0.25.

Abbildung 17 zeigt, daß der 3D-MoRSE Codes durch eine Änderung der Konformation beeinflußt wird, die Stärke des Einflusses der Konformation auf den Code jedoch steuerbar ist.

Als letzte Abbildung zu diesem Thema soll Abbildung 18 zeigen, daß die Konformation zwar den 3D-MoRSE Code beeinflußt, dieser aber sehr wohl in der Lage ist, zwischen Konformeren und Stellungsisomeren zu unterscheiden. So zeigt Abbildung 18 die 3D-MoRSE Codes der Konformere von 6-Chlor-6-methylheptanon-2, bei denen eine Rotation um die C5-C6 Bindung in Schritten von 30° durchgeführt wurden, sowie den 3D-MoRSE Code des Isomers 5-Chlor-2,6-Dimethyl-hexanal.

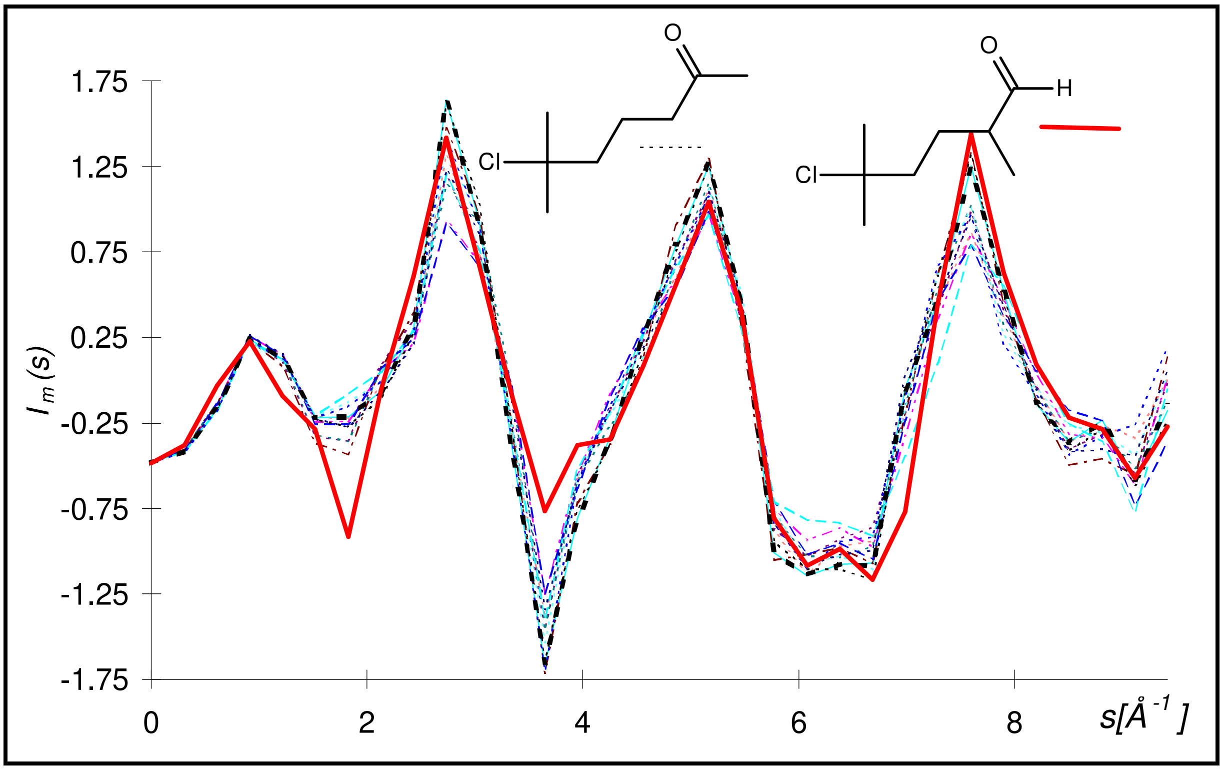


Abbildung 18: 3D-MoRSE Codes (n = 32, s_{max} = 9.42 $Å^{-1}$, A_i = 1) der Rotamere um die C5-C6 Bindung von 6-Chlor-6-methylheptanon-2 (gestrichelt) und seines Isomers 5-Chlor-5,6-dimethylhexanal.

4.3 Fazit

Der 3D-MoRSE Code ist zur Codierung von 3D-Strukturen geeignet. Er ist in der Lage zwischen insignifikanten Bindungslängendifferenzen (> 1 pm) und Änderungen der Bindungsordnung sowie Bindungsbrüchen zu unterscheiden. Ferner kann der Grad des Einflusses von Konformationsänderungen auf den 3D-MoRSE Code mit Hilfe von s_{max} bzw. r_{max} gesteuert werden, insofern die Bedeutung längerer Atomdistanzen mit $1/sr_{ij}$ sinkt, wobei oberhalb von r_{max} der Einfluß auf den Code klein ist und schwankt, da Maxima, die auf Abständen größer r_{max} beruhen, nicht vollständig von den Codewerten wiedergegeben werden. Insofern sind die Parameter für die Anwendung des 3D-MoRSE Codes mit bedacht zu wählen.

Die Möglichkeit den Grad des Einflusses von Konformationsänderungen auf den 3D-MoRSE Code zu bestimmen ist ein Vorteil gegenüber anderen 3D-Strukturcodierungen, sofern das Gemisch an Konformationsisomeren einer Verbindung in Untersuchungen nur durch eine Konformation repräsentiert wird, wie es in den folgenden Untersuchungen der Fall ist. Vergleicht man beispielsweise die Radialcodes derselben Rotamere, die auch Abbildung 17 zugrunde liegen, so zeigen sich hier sehr deutliche Änderungen im Code, wie sie bei einer getrennten Behandlung von Konformeren nützlich sein mögen, für die Repräsentation einer Verbindung durch den 3D-Strukturcode einer Konformation aber nur fraglich geeignet sind.

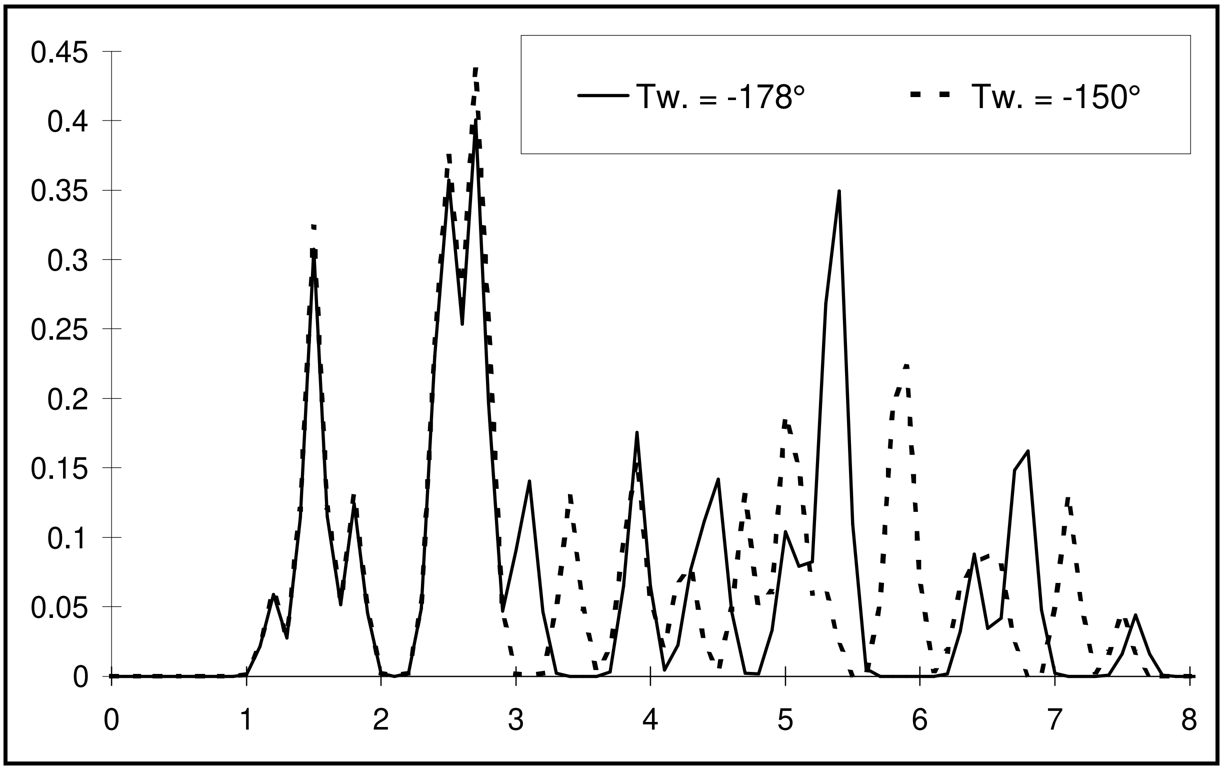


Abbildung 19: Der Einfluß des geänderten Torsionswinkels aus Abbildung 16 auf den Radialcode. Radialcode und Abbildung nach M. Hemmer.[41]

5 Der Zusammenhang zwischen 3D-Struktur und IR-Spektrum

5.1 Der Einfluß des Moleküls auf bekannte Gruppenschwingungen

5.1.1 Die Carbonylschwingung von Ethanal bis Cyclobutanon

Die Lage der Carbonylbande variiert bei den kurzkettigen Aldehyden und Ketonen mit einem bis vier Kohlenstoffatomen erheblich, selbst wenn man darauf achtet, daß die Spektren in flüssiger Phase gemessen wurden. So liegt die Carbonylbande von Ethanal bei 1762 cm^{-1} und sinkt mit zunehmender Länge der Kohlenstoffkette über 1732 cm^{-1} für Propanal auf 1722 cm^{-1} für Butanal. Wer jetzt eine einfache Korrelation zwischen der Zahl der Kohlenstoffatome und der Lage der Carbonylbande aufstellt geht fehl, denn sobald weitere Änderungen im Molekül hinzukommen beeinflussen diese die Lage der Carbonylbande, so liegt die Carbonylbande von Cyclobutanon bei 1784 cm^{-1}, und nimmt in Propenal durch die Konjugation mit einer CC-Doppelbindung von 1732 auf 1688 cm^{-1} ab. Die folgende Tabelle zeigt noch einmal einige Werte.

Tabelle 3: Die Lage der C=O - Bande im Infrarotspektrum. Experimentelle Werte[10,42,43] und mittels DFT-Funktional Becke-3-Lee-Yang-Parr (B3LYP) berechnete und durch Multiplikation mit 0.98 skalierte Werte.[44]

Verbindung	*C=O exp. [1/cm]*	*C=O B3LYP [1/cm]*	*Bemerkungen*
Ethanal	1762		
Propanal	1732		
2-Methyl-propanal	1738		
Aceton	1712		1737 EPA Gasphase
Propenal	1688		1734/1714 Doppelbande EPA Gasphase
Butanal	1722		1744 EPA Gasphase
Butanon	1720		
cis But-2-en-al	1695		GC 1720/1710 des cis/trans Gemisches
Cyclobutanon	1784	1784.592	ca 1810 EPA Gasphase

5.1.2 Die CC-Doppelbindungsschwingung in Hexenolen

Für Alkene findet man in jedem Tabellenwerk zu charakteristischen IR-Absorptionen, die Angabe einer charakteristischen Absorption mit variabler Intensität für die Streckschwingung der CC-Doppelbindung zwischen 1600 und 1700 cm^{-1}, so z.B. bei Hesse et al. von 1680-1620 cm^{-1}.[45] In der Legende steht dann, daß die Intensität sehr schwach sein kann, falls die Doppelbindung "mehr oder weniger symmetrisch substituiert" ist. Mit anderen Worten: ob die "charakteristische Bande" sichtbar ist, hängt von der Substitution der Bindung ab. Um diese Aussagen nachzuvollziehen und Einflüsse auf die Bande zu studieren, wurden für einen Satz von Hex-2-enol-Derivaten die IR-Spektren berechnet.

5.1.2.1 Berechnung der IR-Spektren von Hex-2-enol-Derivaten

Da bei den Hex-2-enol-Derivaten die Effekte der Struktur auf eine spezielle Absorption studiert werden sollen, kann auf exakte absolute Wellenzahlen verzichtet werden. Daher wurde zur Berechnung der IR-Spektren das semiempirische AM1-Verfahren gewählt[46], dessen Schwingungswerte aufgrund der harmonischen Näherung für die Oszillatoren zwar im allgemeinen zu hoch liegen, die aber dennoch einen Trend beschreiben können. Die Vorgehensweise für alle Moleküle war:

- Eingabe der Struktur mittels Editor
- Generierung einer Ausgangskonformation
- Optimierung der Geometrie mittels AM1
- Berechnung der Schwingungsfrequenzen basierend auf der optimierten 3D-Struktur mit AM1.

5.1.2.2 Ergebnisse

Alle Ergebnisse liegen, wie erwartet ca. 250 cm^{-1}, zu hoch. Die Intensität der Absorptionsbande schwankt stark, während die Frequenz der Streckschwingung der CC-Doppelbindung relativ konstant bleibt (Ausnahme: eine deutliche Erniedrigung der Schwingungsfrequenz tritt bei konjugierten Doppelbindungen auf). Das folgende Diagramm (Abbildung 20) zeigt alle Ergebnisse. Interessant ist, daß eine zweite, nicht konjugierte Doppelbindung keine Verschiebung der Bandenlage der CC-Doppelbindung gegenüber der CC-Doppelbindungsbande von Cyclohex-2-enol bewirkt, obwohl die erhöhte Ringspannung einen solchen Effekt vermuten läßt. Sind zwei Doppelbindungen im Cyclohexanring vorhanden, so koppeln diese. Aus der Kopplung resultiert die Aufspaltung der CC-Doppelbindungsbande in eine symmetrische und eine asymmetrische Bande bezüglich der Längenänderung der beiden Doppelbindungen. Beim Cyclohexa-2,5-dienol, dessen Doppelbindungen nicht konjugiert sind, liegt die Frequenz der asymmetrischen Streckschwingung der CC-Doppelbindungen niedriger als die der symmetrischen. Bei den konjugierten Doppelbindungen des Cyclohexa-2,4-dienols erfolgt die Frequenzaufspaltung genau andersherum. Hier liegt die Frequenz der symmetrischen Streckschwingungsbande der CC-Doppelbindung unter der asymmetrischen.

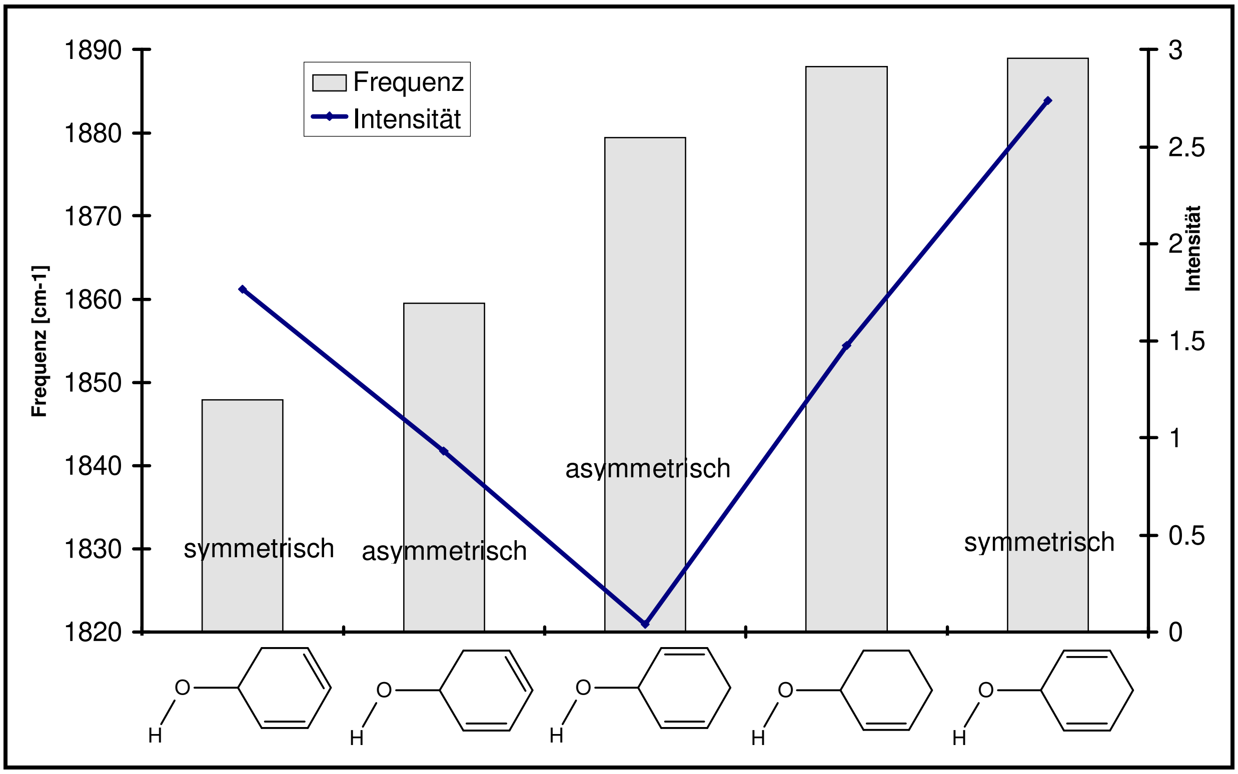


Abbildung 20: Frequenz und Intensität der Streckschwingung der CC-Doppelbindung bei Cyclohexenolderivaten, berechnet mit AM1.[46] Die Frequenzen der Schwingungen sind auf der linken y-Achse aufgetragen die Intensitäten auf der rechten.

Bei der offenkettigen *trans*-Verbindung, *trans*-Hex-2-enol (nicht in Abbildung 20 enthalten), liegt die Bande für die Streckschwingung der CC-Doppelbindung mit 1896 cm^{-1} höher als bei den cyclischen Derivaten, die Bande hat aber nach der asymmetrischen Bande der Streckschwingung der CC-Doppelbindung von Cyclohexa-2,5-dienol die zweitschwächste Intensität. Für das offenkettige *cis*-Hex-2-enol wird mit 1886 cm^{-1} dieselbe Frequenz wie für das cyclische Derivat erhalten, allerdings ergibt sich hier die gleiche schwache Intensität wie für die *trans*-Verbindung.

5.2 Sterische Einflüsse auf das IR-Spektrum

5.2.1 Der Einfluß der cis/trans-*Isomerie am Beispiel von Fumar- und Maleinsäure*

Schon die berechneten Spektren der offenkettigen *cis/trans* Hex-2-enol-Derivate zeigten einen Unterschied in der Bandenlage der Streckschwingung der CC-Doppelbindung von 10 cm^{-1}.

Wesentlich größere Effekte auf das Infrarotspektrum lassen die *cis/trans*-Isomeren der 1,4-Butendisäure, Malein- bzw. Fumarsäure, erwarten. Die Betrachtung der experimentellen Spektren (Abbildung 21) zeigt deutliche Unterschiede über den gesamten Meßbereich. Selbst in Wellenzahlbereichen, für die allgemein akzeptierte Zuordnungen zu bestimmten Valenzschwingungen bestehen, wie der O-H - Schwingung der Carboxylgruppe, der Carbonylbande, sowie der Streckschwingung der CC-Doppelbindung sind erhebliche Unterschiede zu sehen. Beispielsweise ist die Lage der Carbonylbande der Maleinsäure gegenüber der Bande von Fumarsäure um 30 cm^{-1} zu höheren Wellenzahlen verschoben (Fumarsäure 1660 cm^{-1}; Maleisäure 1692 cm^{-1}).

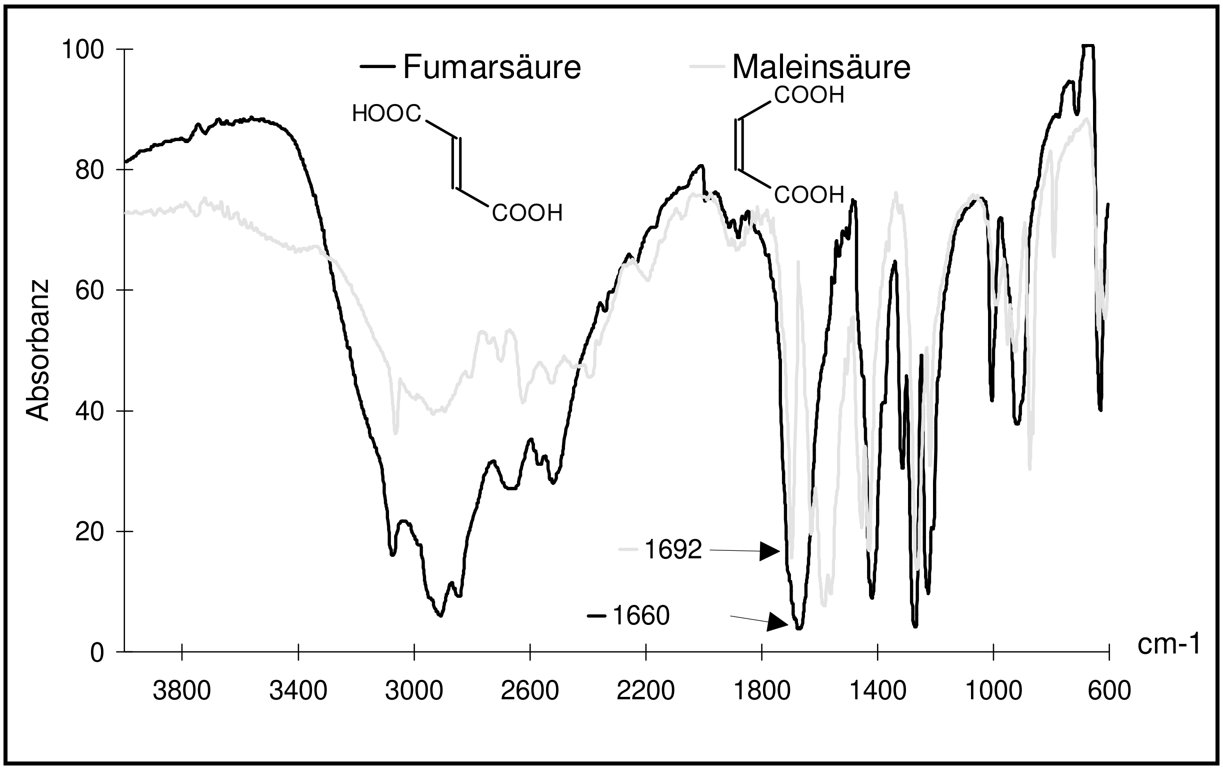


Abbildung 21: Die experimentellen Infrarotspektren[i] des *trans*- und des *cis*-Isomers der But-2-en-1,4-disäure, Fumarsäure bzw. Maleinsäure.[47]

[i] Die Spektren wurden mit Hilfe UN-SCAN-IT[i] digitalisiert und als XY-Werte gespeichert. Die Darstellung erfolgte mit Hilfe von MS-EXCEL.

5.2.2 Unterschiede im IR-Spektrum der Diastereomere von 1,2,3,4,5,6-Hexachlorcyclohexan.

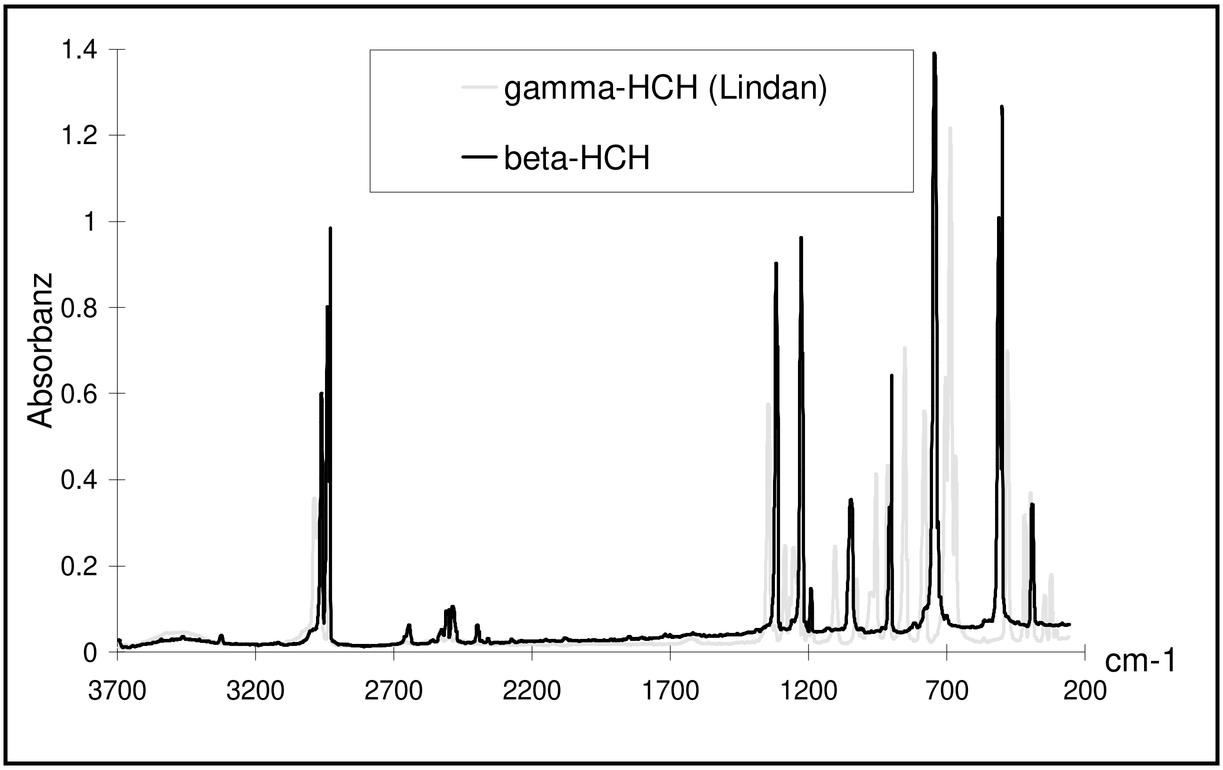


Abbildung 22: Die experimentellen Spektren der diastereomeren 1,2,3,4,5,6-Hexachlorcyclohexanderivate Lindan (γ-HCH) und β-HCH.

Wie die beiden experimentellen Spektren von γ-HCH und β-HCH in Abbildung 22 zeigen, können sich die Infrarotspektren von Diastereomeren ganz erheblich unterscheiden, insbesondere wenn die Symmetrie der Diastereomere von einander abweicht. So hat das all-*trans*-Derivat β-HCH aufgrund der hohen Symmetrie deutlich weniger Banden als das *cis*-1,2-*trans*-3,4,5,6- Hexachlorcyclohexan, γ-HCH, dessen Symmetrie geringer ist.

5.2.3 Stereoisomere

Die chiralen Isomere einer Verbindung sind im Gegensatz zu Diastereomeren einer Verbindung mit Hilfe der Infrarotspektroskopie nicht zu unterscheiden. Eine Aufklärung der absoluten Konfiguration ist deshalb mit Hilfe der Infrarotspektroskopie nicht möglich, es sei denn, man setzt die zu untersuchende Verbindung vorher mit einem chiralen Auxilliar um.

6 Die Simulation von IR-Spektren

6.1 Quantenmechanische Verfahren zur Simulation von IR-Spektren

Die Berechnung von Infrarotspektren mit quantenmechanischen Methoden, ob nun semieempirisch oder *ab initio*, beruht grundsätzlich auf der zweiten Ableitung der Energie nach der (vorher optimierten) 3D-Struktur. Bei der Berechnung der Frequenzen wird grundsätzlich die harmonische Näherung verwendet, was zur Folge hat, das die Werte nicht der anharmonischen Realität chemischer Bindungen entsprechen. So können beispielsweise Obertöne von Schwingungen mit der harmonischen Näherung nicht vorhergesagt werden und auch die Energien für das erste Schwingungsniveau liegen meist etwas zu hoch. Allerdings ist der Einfluß der Anharmonizität auf die Frequenzen der Moleküle im Grundzustand meistens gering und bewegt sich in der Regel in der selben Größenordnung, so daß versucht wird mit Hilfe von Skalierungsfaktoren die Abweichung, die sich aus der Verwendung der harmonischen Näherung und andere Einflüsse der verwendeten Modelle ergibt, auszugleichen.

Welcher Skalierungsfaktor dabei für welche Methode und für welchen Basissatz zu verwenden ist und wie zuverlässig, die sich damit ergebenden Ergebnisse sind, ist damit Gegenstand wissenschaftlicher Diskussionen. Erschwert wird diese Diskussion zum einen dadurch, daß es sehr schwer ist IR-Banden eindeutig einer Schwingung zuzuordnen, wobei auch die Datenbasis nicht gerade üppig ist, und zum anderen, daß die verwendeten Datensätze nur schwer vergleichbar sind. Denn je nach Größe der Moleküle ändern sich die Meßmethoden und auch für die verwendeten *ab initio* Verfahren ist eine Abhängigkeit von der Molekülgröße und von den vorkommenden Elemente zu erkennen.

So lehnen beispielsweise Scott und Radom die beiden semiempirischen Methoden AM1 und PM3 zur Berechnung von Schwingungsfrequenzen ab,[13] da sie für ihren Datensatz aus 122 teilweise kleinen anorganischen Molekülen mit zusammen 1066 einzelnen Schwingungen bei über einem Drittel der Frequenzen Abweichungen fanden, die größer als 10 % waren, während Healy und Holder die AM1 Methode empfehlen, da sie für ihren Datensatz an 42 kleinen organischen Molekülen nach Skalierung nur einen mittleren Fehler von 6 % bekamen.[48] Um die Verwirrung zu vervollständigen werden zudem oft für einzelne Frequenzbereiche verschiedene Skalierungsfaktoren vorgeschlagen, z.B. für Valenzschwingungen oberhalb von 1600 cm^{-1}, für Deformationsschwingungen zwischen 1600 und 400 cm^{-1} und niederfrequente Schwingungen unterhalb von 400 cm^{-1}. Zudem schweigen sich die meisten Publikationen zu Intensitäten aus.

Dennoch können einige Trends festgestellt werden:

- Die Skalierungsfaktoren sind in der Regel größer als 0.85 und meistens kleiner als 1.0. Wobei die Skalierungsfaktoren in der Regel umgekehrt proportional zur Schwingungsfrequenz sind.

- Dichtefunktional-Methoden insbesondere mit Hybridfunktionalen wie B3LYP sind in der Regel besser zur Berechnung von Schwingungsfrequenzen geeignet als ab initio Molekülorbital-Verfahren wie Hartree-Fock- (HF) und Moller-Plesset- (MP2) Methoden. Langhoff erreicht mit Dichtfunktionalmethoden sogar bei der Intensität der Schwingungen polycyclischer aromatischer Kohlenwasserstoffe eine relativ gute Übereinstimmung mit dem Experiment.[49]

- Die Qualität der berechneten Frequenzen steigt nicht unbedingt mit der Größe des Basissatzes. So findet sich bei Scott und Radom für MP2 Rechnungen fast kein Unterschied zwischen fu/6-31G(d) und fc/6-311G(d,p) und für die HF-Methode ist zwar zwischen 3-21G nach 6-31G(d) ein Qualitätssprung zu sehen aber 6-311G(df,p) bringt wieder einen Qualitätsverlust.

Aus dem vorstehenden wurde beschlossen für die Arbeit bei einfachen organischen Molekülen die AM1-Methode zu verwenden, vor allem wenn keine absoluten Frequenzen notwendig waren. Andernfalls wurden DFT-Methoden verwendet, wobei einer Empfehlung nach Erfahrungen innerhalb des Computer-Chemie-Centrums gefolgt wurde, und das Hybridfunktional B3-LYP mit dem 6-31G(5p, 7f) Basissatz und einem Skalierungsfaktor von 0.98 verwendet wurde.[44] Zumal Scott und Radom ohne weitere Basissätze zu testen dasselbe Funktional mit einem etwas kleineren Basissatz (B3-LYP/6-31G(d) mit einem Skalierungsfaktor von 0.9614) vorgeschlagen hatten, weil sie damit die besten Ergebnisse erzielt hatten.

Zusammenfassend kann man sagen, daß *ab initio* Verfahren die Simulation der Infrarotspektren von 3D-Strukturen erlauben, wobei für HF/6-31G(d) und Dichtefunktionalmethoden nach Skalierung die Zahl der berechneten absoluten Frequenzen, die von den experimentellen Werten um mehr als 10 % abweichen, unter 10 % Prozent liegt. Diese Werte werden vielfach an sehr kleinen (zweiatomigen) Molekülen ermittelt. Bei konformativ flexiblen Molekülen wird die Zuordnung von Absorptionsbanden zu molekularen Schwingungen noch dadurch erschwert, daß sowohl das Spektrum als auch seine Berechnung von der Konformation der Moleküle abhängen, was sich im Infrarotspektrum meistens zwar nur durch eine Peakverbreiterung bemerkbar macht, aber auch durchaus zur Peakaufspaltung führen kann (vgl. Tabelle 3). Daher ist bei konformativ flexiblen Molekülen zur IR-Spektrensimulation zum Teil eine Berechnung der Infrarotspektren für alle relevanten Konformationen notwendig, die dann entsprechend der Boltzmannverteilung gewichtet werden müssen.

Ab initio Methoden sind die einzige Möglichkeit zur Simulation von Infrarotspektren für neue Stoffklassen, zu denen noch keine experimentellen Daten existieren, und für kleine oder hoch symmetrische Moleküle, bei denen jede Substitution das Spektrum deutlich verändern würde, was die Nutzung empirischer Methoden erheblich erschwert. Beispiele für solche Moleküle wären Methan, Benzol, Methanol und 3-Aminopropanol. Tabelle 4 faßt noch einmal empfohlene Methoden und Skalierungskoeffizienten zusammen.

Tabelle 4: Empfohlene Methoden und Skalierungskoeffizienten. Literaturstellen Scott und Radom[13], Langhoff[48], Bauschlicher und Partridge[50], CCC-Empfehlung[44].

Methode	Basissatz	Skalierungskoeffizient	Literaturstelle
AM1	-	0.9532	Scott und Radom
HF	6-31G(d)	0.8953	Scott und Radom
B3-LYP	4-31G	0.9580	Langhoff
B3-LYP	6-31G(d)	0.9614	Scott und Radom
B3-LYP	6-311+G(3df,2p)	0.9890	Bauschlicher und Partridge
B3-LYP	6-311G(d)(5d,7f)	0.9800	CCC-Empfehlung

6.2 Korrelation von Struktur und Spektrum - empirische Verfahren

Neben der Berechnung von IR-Spektren, mit Hilfe quantenmechanischer Methoden, hat es immer Versuche zur empirischen Korrelation von IR-Spektrum und Struktur gegeben. Bis vor kurzem ging der überwiegende Teil dieser Versuche von der Korrelation zwischen Strukturfragmenten und Teilspektren aus. Egal wie diese Korrelationen erhalten wurden, ob mit einem Additionsschema[51] oder einem neuronalen Netz, alle Ansätze haben drei prinzipielle Limitationen:

- Es können keine Spektren außerhalb der Datenbasis vorhergesagt werden. Mit anderen Worten, daß zur Simulation notwendige Wissen muß im Trainingsdatensatz, dem Datensatz der zur Aufstellung der Korrelation genutzt wurde, enthalten gewesen sein.

- Ein Satz von Substrukturen kann nie vollständig sein, da deren Anzahl prinzipiell unbegrenzt ist. So gibt es beispielsweise bei Biphenyl alleine 221 verschiedene Möglichkeiten der Substitution bei einem bis zehn identischen Substituenten.

- Abweichungen sind bei einem auf Substrukturen beruhenden Ansatz immer dann zu erwarten, wenn im IR-Spektrum Schwingungen auftreten bzw. Faktoren Valenzschwingungen, wie die Carbonylschwingung, beeinflussen, die nicht auf eine Substruktur beschränkt sind bzw. durch sie beschrieben werden können. Bei der Infrarotspektroskoie ist dies recht häufig der Fall, da beispielsweise fast immer die Molekülsymmetrie beachtet werden muß. Aber selbst bei einer so auf das einzelne Atom zentrierten Spektroskopie, wie der NMR-Spektroskopie, kann es zu Abweichungen bei der Spektrenvorhersage mit einem Substrukturansatz kommen, wenn weit entfernte Substituenten in einem konjugierten System eine Rolle spielen.[52]

Für die Infrarotspektroskopie konnten Pretsch et al.[53] sogar nachweisen, daß die Korrelation von Fragmenten mit Teilspektren nicht sinnvoll ist. So erlaubten literaturbekannte Banden bei einem Datensatz von 20 257 IR-Spektren nur bei rund 50% der Moleküle einen sicheren Rückschluß auf 62 getestete terminale Strukturfragmente mit drei bis acht Atomen.

Trotz der genannten Probleme wurden in den vergangenen Jahrzehnten fragmentbasierte Systeme zur Simulation von IR-Spektren publiziert. Systeme mit 640 und 720 Fragmenten waren in Teilbereichen durchaus erfolgreich[54,55], ein Durchbruch konnte mit ihnen aber nicht gelingen. Weiter führt wohl ein Ansatz von Clerc et al., die eine Serie von Deskriptoren (Hauptkomponenten von „path counts in chemical node-colored molecular graph") benutzten, die die Abfolge der Atome im Molekül und den Grad der Verzweigung im Molekül beschreiben und aus der Bindungsliste abgeleitet werden.[56] Dieser Ansatz hat den Vorteil, ohne vordefinierte Substrukturen auszukommen. Neben den Deskriptoren ist die zweite wesentliche Komponente des Ansatzes die multidimensionale Interpolation des zu simulierenden IR-Spektrums über eine Hauptkomponentenanalyse (PCA). Die Hauptkomponentenanalyse wird auf den Deskriptoren aller Moleküle der zur Simulation benutzten Datenbank durchgeführt. Anschließend werden die Hauptkomponenten für die Anfragestruktur bestimmt. Die Anfragestruktur ist die Struktur, für die das IR-Spektrum simuliert werden soll. Das simulierte Spektrum wird dann aus den IR-Spektren der im Raum, den die ersten acht Hauptkomponenten aufspannen, benachbarten Molekülstrukturen gewichtet interpoliert. Der Faktor für die

Wichtung der Infrarotspektren der Nachbarverbindungen ist dabei die Entfernung zwischen Datenbankstruktur und Anfragestruktur im Raum der Hauptkomponenten.

Der Clerc'sche Ansatz berücksichtigt die 3D-Struktur des Moleküls nicht. Dies soll in Zukunft durch die Einbeziehung geeigneter 3D-Strukturbeschreibungen, wie dem 3D-MoRSE Code, geändert werden.[57]

Alle empirischen Methoden zur Simulation von IR-Spektren beruhen auf folgendem Prinzip:

1. Codierung der Molekülstrukturen

2. Modellierung der Beziehung zwischen Struktur (Strukurcode) und IR-Spektrum einer Substanz mit einem Korrelationsverfahren

3. Codierung der Anfragestruktur und Vorhersage des IR-Spektrums unter Nutzung der in zwei bestimmten Korrelation zwischen IR-Spektrum und Strukturcode.

Abbildung 23 stellt diesen Ablauf noch einmal schematisch dar.

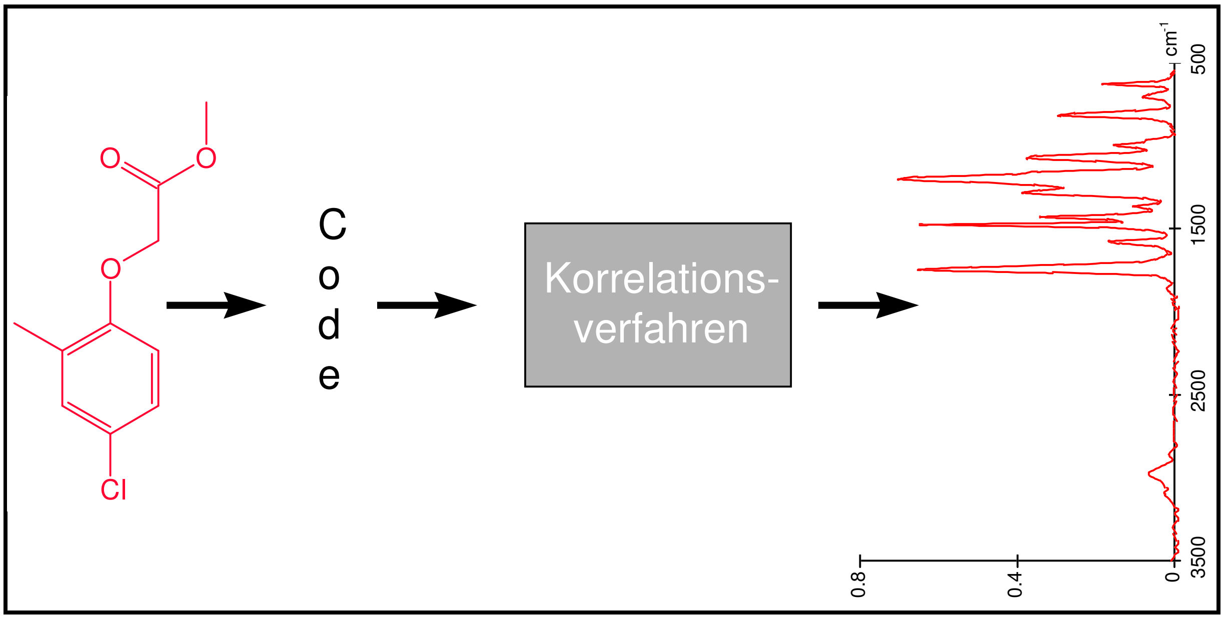


Abbildung 23: Schematische Ablauf der Simulation von IR-Spektren. Ausgehend von der Molekülstruktur wird diese zunächst codiert und aus dem Code mit Hilfe eines Korrelationsverfahrens das IR-Spektrum vorhergesagt.

6.3 Korrelation von 3D-Struktur und Infrarotspektrum - eine neue Methode

Im Gegensatz zu den vorstehenden Methoden, basiert der im Rahmen dieser Arbeit entwickelte Ansatz zur Simulation von Infrarotspektren, auf der Codierung der 3D-Struktur unter Verwendung geeigneter atomarer Eigenschaften. Die hier entwickelte Methodik zur Simulation von Infrarotspektren setzt die folgenden drei Methoden voraus:

- Den 3D-MoRSE Code zu Codierung der 3D-Struktur von Molekülen in einem Vektor fester Länge

- ein Verfahren zur Datenreduktion von IR-Spektren nach Zupan basierend auf der Hadamard Transformation

- Neuronale Counterpropagation-Netze (CPG-Netze) zur Korrelation von IR-Spektrum und codierter 3D-Struktur.

Zwei weitere Programme sind zwar nicht für die Methode an sich essentiell, doch ohne sie wäre die Entwicklung der Methode und die nachfolgenden Experimente nicht denkbar. Der 3D-Strukturgenerator CORINA[58,21] wurde benutzt, um die notwendigen 3D-Strukturen schnell und zuverlässig zu generieren. CORINA benötigt auf einer Workstation (Sun Sparc 10-40 unter Solaris 2.4.1) in der Regel eine halbe Sekunde zur Generierung der 3D-Struktur eines Moleküls mit rund 15 Nicht-Wasserstoff Atomen.[21,59,60,61,62] Das zweite Verfahren war die Partial Equalization of Orbital Electronegativities (PEOE-Methode) nach Gasteiger et al. zur schnellen Abschätzung partieller Atomladungen, die als Atomeigenschaft im Rahmen des 3D-MoRSE Codes benutzt wurden.[63,64,65]

Insgesamt stellt sich der Weg von der Bindungsliste der Anfragestruktur bis zum simulierten IR-Spektrum, wenn ein trainiertes CPG-Netz vorausgesetzt wird, wie folgt dar (Abbildung 24):

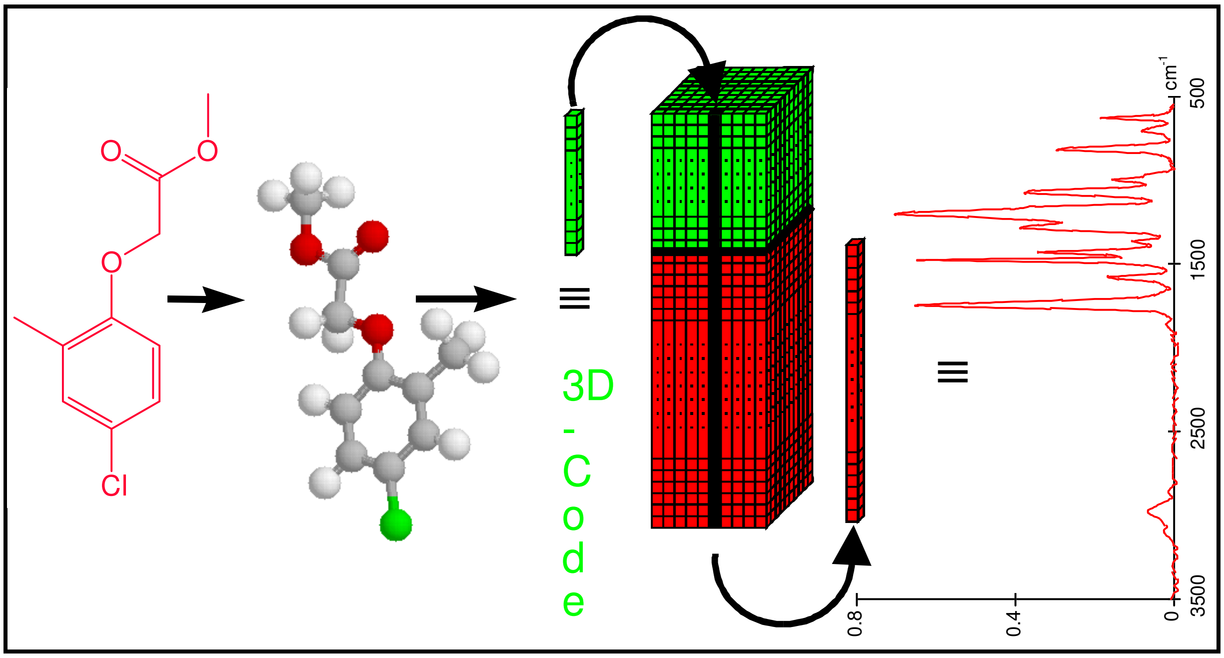


Abbildung 24: Methodik der IR-Spektrensimulation ausgehend von der Molekülstruktur über die 3D-Struktur und den 3D-MoRSE Code und ein trainiertes neuronales Counterpropagation-Netz.

6.3.1 Neuronale Counterpropagation-Netze

Neuronale Netze haben bei der Modellierung komplexer Zusammenhänge den Vorteil, die Abhängigkeiten implizit aus (Trainings-) Daten zu lernen, ohne daß die mathematische Form der Abhängigkeit explizit formuliert werden muß. Der Vorgang, in dem das neuronale Netz den Zusammenhang zwischen Ein- und Ausgabedaten lernt, im Fall der Simulation von IR-Spektren zwischen 3D-MoRSE Code und IR-Spektrum, nennt man Training; den Datensatz aus Paaren von Ein- und Ausgabedaten Trainingsdatensatz. Gespeichert wird die Abhängigkeit in den sogenannten Gewichten des neuronalen Netzes.

Grundsätzlich könnte ein neuronales Netz, wie jedes andere Modellierungsverfahren, beliebige Zusammenhänge auswendig lernen, wenn mehr Variable bzw. Gewichte vorhanden sind als Paare von Ein- und Ausgabedaten. Deshalb ist es notwendig nach jedem Training das trainierte neuronale Netz mit einem *Testdatensatz* von Ein- und Ausgabepaaren, die nicht zum Training verwendet wurden, zu prüfen.

Der Aufbau der verwendeten neuronalen Netze, neuronale Counterpropagation-Netze, ist in Abbildung 25 dargestellt:

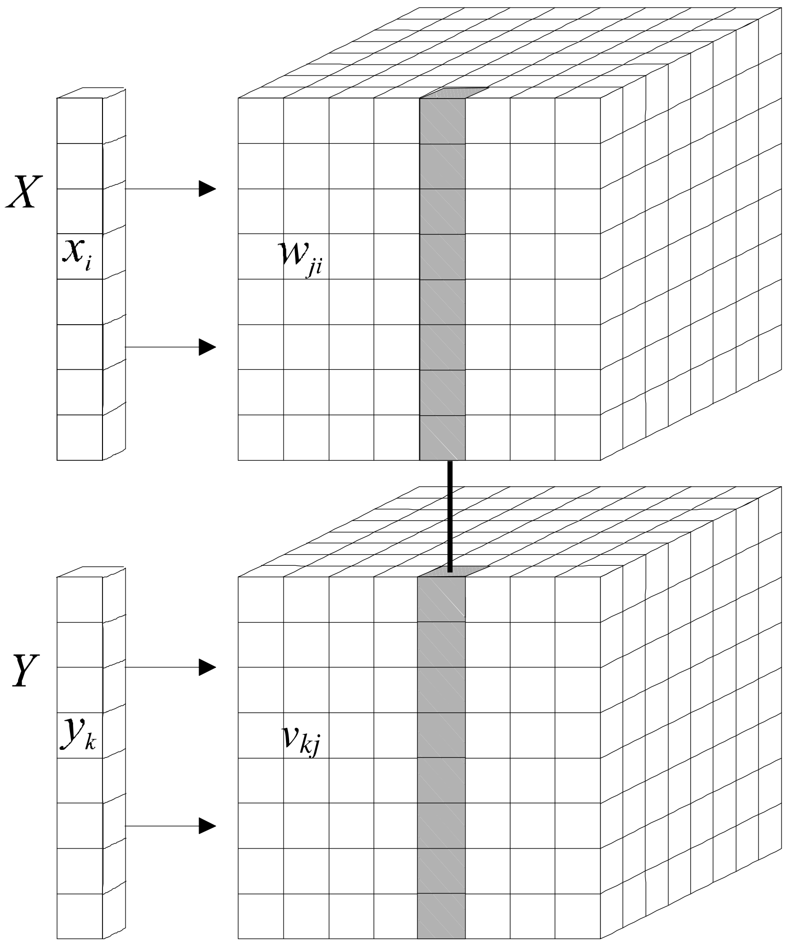


Abbildung 25: Prinzipieller Aufbau eines Counterpropagation-Netzes. Zweidimensionale Anordnung von Neuronen (Säulen), die in einen Eingabe- (oben) und einen Ausgabeteil (unten) aufgespalten sind. Die einzelnen Gewichte sind hier als Würfel dargestellt, diese bilden zweidimensionale Schichten. Erläuterung der Symbole siehe Text.

In einem Counterpropagation-Netz wird der Zusammenhang zwischen einer oder mehreren abhängigen Variablen $Y(y_1, y_2, ..., y_k, ...)$ und den Eingabedaten $X(x_1,x_2, ..., x_i, ...)$, den unabhängigen Variablen, modelliert. Die Beziehung $Y = f(X)$ wird nicht explizit in einer mathematischen Form ausgedrückt, sondern in den Gewichten der Neuronen des künstlichen neuronalen Netzes gespeichert. Ein Counterpropagation-Netz besteht aus zwei Blöcken, einem Eingabeblock für die X-Variablen und einem Ausgabeblock für die oder mehrere Y-Variable. Die Blöcke sind durch Neuronen verbunden, die als Säulen (vgl. graue Markierung in Abbildung 25) vom Eingabeteil in den Ausgabeteil durchgehen und die funktionalen Einheiten des Netzes darstellen. Jedes Neuron j hat ebenso viele Eingabegewichte w_{ji}, wie unabhängige Variable X vorhanden sind. Die Anzahl der Ausgabegewichte v_{jk} entspricht folglich der Menge der abhängigen Variablen Y.

Das Training eines Counterpropagation-Netzes ist ein kompetitiver Lernprozeß, in dem das Netz durch wiederholte Anpassung an die Trainingsdaten den Zusammenhang zwischen Ein- und Ausgabedaten lernt; im Fall der Simulation von IR-Spektren also zwischen Strukturcode und Infrarotspektrum. Erster Schritt jeder Anpassung ist die Suche nach dem nächstliegenden Neuron, Neuron w_c, zum Datenpunkt s. Vgl. Gleichung 14.

$$out_c \Leftarrow \min\left[\sum_{i=1}^{m}\left(x_{si} - w_{ji}\right)^2\right] \qquad j = 1, 2, 3, ..., n \qquad (14)$$

mit:

m	Anzahl der betrachteten Variablen
X	Datenpunkte
s	aktueller Datenpunkt
i	Index der Variablen bzw. des korrespondierenden Gewichtes
w_j	Gewichte des Neurons j
j:	Neuron
n:	Anzahl der Neuronen im Counterpropagation Netzwerk

Für ein Counterpropagation Netzwerk gibt es zwei Trainingsmethoden: beaufsichtigtes und unbeaufsichtigtes Training (*supervised* und *unsupervised*). Die Trainingsmethode bestimmt, ob nur die Eingabegewichte X, (im Fall des unbeaufsichtigten Trainings) oder ob sowohl Eingabe- wie auch Ausgabevariablen (beaufsichtigtes Training) X und Y, bei der Suche nach dem ähnlichsten Neuron (*winning neuron*) berücksichtigt werden.

In beiden Fällen, unabhängig von der Trainingsmethode, werden Eingabe- und Ausgabegewichte w_{ji} und v_{kj} angepaßt und zwar so, daß alle Gewichte der Neuronen des Counterpropagation Netzes dem eingegebenen Datenpunkt ähnlicher werden. Im Anpassungsprozeß werden die Gewichte des ähnlichsten Neurons w_c am stärksten den Eingabedaten angepaßt. Mit der Zunahme des Abstandes des Neurons vom ähnlichsten Neuron, w_c, nimmt die Änderung der Neuronengewichte ab.

Exakt beschrieben wird die Änderung der Neuronengewichte im Training durch die folgenden zwei Gleichungen:

$$w_{ji}^{neu} = w_{ji}^{alt} + \eta(t)\ \mathrm{a}(d_c - d_j)(x_i - w_{ji}^{alt}) \qquad (15)$$

$$v_{jk}^{neu} = v_{jk}^{alt} + \eta(t)\ \mathrm{a}(d_c - d_j)(y_k - v_{jk}^{alt}) \qquad (16)$$

d_c-d_j: topologische Distanz zwischen Neuron *c* und Neuron *j*

$a(d_c$-$d_j)$: Von der Topologie abhängige Skalierungsfunktion für die Lernrate η

t: Nummer der aktuellen Trainingsiteration

η: Lernrate

Ein unüberwachter Lernprozeß, bei dem ausschließlich die Eingabevariablen, *X*, zur Bestimmung des ähnlichsten Neurons benutzt werden, ist äquivalent zum Lernprozeß einer selbstorganisierenden Eigenschaftskarte, wie beispielsweise einem Kohonen-Netzwerk.[66,67]

Nach dem Training eines Counterpropagation Netzes wird üblicherweise ein Erinnerungstest des trainierten Netzes durchgeführt, dabei wird das trainierte Netz mit dem Trainingsdatensatz getestet. Im Rahmen des Erinnerungstests (recall test) wird jedes Muster des Trainingsdatensatzes dem ähnlichsten Neuron assoziiert. Dabei können auch mehrere Datenpunkte demselben Neuron assoziiert werden, wenn die Unterschiede zwischen den Mustern geringer sind als zwischen dem Neuron und seinen Nachbarn. Neuronen, denen ein Muster assoziiert wurde, nennt man auch belegte Neuronen. Die Neuronen, denen kein Muster assoziiert wurde heißen leere (*empty*) Neuronen.

Untersucht man die zweidimensionale Anordnung der Neuronen mit den assoziierten Molekülen, indem man von oben auf das Counterpropagation Netz schaut, kann man die Verteilung des gesamten Datensatzes über das Netzwerk erkennen. Ein trainiertes Counterpropagation Netz wird dabei die Ähnlichkeiten zwischen den Objekten zeigen, insofern die ähnlichen Objekte benachbarten Neuronen assoziiert werden. Dabei kann der Abstand zwischen zwei Neuronen als qualitatives Maß für Ähnlichkeit zweier Moleküle benutzt werden.

Aber ein Counterpropagation Netz zeigt nicht nur die Ähnlichkeitsbeziehungen zwischen den Datenpunkten des Trainingsdatensatzes, es kann auch für Vorhersagen genutzt werden. Testdaten werden in ein trainiertes CPG-Netz in der selben Art und Weise gemappt wie die Trainingsdaten, indem Gleichung (14) zur Bestimmung des ähnlichsten Neurons benutzt wird.

Selbstverständlich wird beim Test keine Korrektur der Gewichte mehr durchgeführt und nur die Eingabeschicht mit den Gewichten w_{ji} zur Bestimmung des ähnlichsten Neurons genutzt.

Betrachtet man bei einem Test die Ausgabeschicht des Netzes, v, so können die Ausgabegewichte des ähnlichsten Neurons, c, als Vorhersage für das Eingabeobjekt betrachtet werden. Wurde zum Beispiel ein Counterpropagation-Netz mit der Molekülstruktur als Eingabe und dem korrespondierenden Infrarotspektrum als Ausgabe trainiert, erlaubt die Eingabe einer Struktur die Vorhersage ihres IR-Spektrums.

6.4 Prinzipielles zur Simulation von IR-Spektren mit einem Counterpropagation Netz

Das Training eines neuronalen Netzes ist ein induktiver Lernprozeß, d.h. hier wird anhand von Beispielen (Trainingsdatensatz) gelernt. Das bedeutet, es können nur Beziehungen zwischen Ein- und Ausgabe gelernt werden, die im Trainingsdatensatz enthalten sind. Bezogen auf die Infrarotspektroskopie bedeutet das, daß Schwingungsmuster, die nicht im Trainingsdatensatz enthalten sind, nicht vorhergesagt bzw. simuliert werden können. Wenn beispielsweise im Trainingsdatensatz eines Netzes nur Benzolderivate enthalten waren, wird man mit diesem Netz die Gerüstschwingungen eines offenkettigen Alkans nicht vorhersagen/simulieren können.

Eine zweite Einschränkung ergibt sich aus der Verwendung Ähnlichkeitsbeziehung zwischen Eingabvektor und Neuron. Das neuronale Counterpropagation-Netz lernt im Endeffekt Analogien. Dies bewahrt das Counterpropagation-Netz im wesentlichen vor den Problemen des Auswendiglernens von Fakten (*Overtraining*), die für Backpropagation Netze typisch sind. Counterpropagation Netze setzen allerdings voraus, daß genügend ähnliche Beispiele, also Struktur-Spektren-Paare vorhanden sind, so daß ein Analogieschluß möglich ist. Dies ist aber bei dem komplexen Zusammenhang zwischen Struktur und IR-Spektrum nur bei enger verwandten Molekülen möglich, wie die nachfolgenden Ergebnisse zeigen werden. So werden die Vorhersagen um so besser, je ähnlicher die Moleküle sind.

Deshalb möchte ich an dieser Stelle dazu aufrufen, das IR-Spektrum jeder neuen Verbindung in computerlesbarer Form zu hinterlegen, um die Aufklärung der Struktur ähnlicher Verbindungen mittels simulierter IR-Spektren zu erleichtern. Denn nur wenn Struktur und IR-Spektrum einer Verbindung in elektronischer Form vorliegen, können die Daten zur Entwicklung bzw. Verbesserung von Systemen zur Simulation von Infrarotspektren genutzt werden.

6.5 Die Repräsentation der Infrarotspektren

Alle in dieser Arbeit verwendeten Infrarotspektren wurden der SpecInfo-Datenbank[10] entnommen. Es wurden nur Absorbanz-Vollspektren verwendet. Vollspektren, die als Transmissionsspektren abgespeichert waren, wurden zunächst in Absorbanzspektren konvertiert. Als nächstes wurden alle Spektren durch Interpolation auf dieselbe Auflösung in den Frequenzbereichen 3500-2000 cm^{-1} und 2000-552 cm^{-1} gebracht. Im ersten Bereich wurden die Spektren auf 150 Punkte mit einem Abstand von 10 cm^{-1} digitalisiert. Im zweiten Bereich zwischen 2000 und 552 cm^{-1} wurde die Auflösung, wegen des informativen Fingerprintbereichs, auf 4 cm^{-1} erhöht, gleich 362 Punkten. Insgesamt ergibt das 512 Absorbanzwerte zwischen 3500 und 552 cm^{-1}, die im nächsten Schritt in 512 Hadamard-Koeffizienten transformiert wurden. Hierzu wird die schnelle Hadamard Transformation (FHT) benutzt, die sich von der schnellen Fourier-Transformation (FFT) (vgl. Gleichung 17) durch die Verwendung einer Rechteckfunktion (Walsh Funktion) anstelle der Sinusfunktion unterscheidet. Die Funktionsgleichung für die Hadamard-Transformtion zeigt Gleichung (18), dabei steht **H** für die Hadamardfunktion und *k* ist der *k*-te Hadamard-Koeffizient.

$$g(\Omega) = \int_{-\infty}^{\infty} f(t) \exp(-i\Omega t) dt \qquad (17)$$

$$\begin{aligned} &Y = \mathrm{H}\,X \qquad \mathit{mit} \\ &\mathrm{H}_k = \begin{Vmatrix} \mathrm{H}_{k-1} & \mathrm{H}_{k-1} \\ \mathrm{H}_{k-1} & -\mathrm{H}_{k-1} \end{Vmatrix} \\ &\mathrm{H}_0 = 1 \end{aligned} \qquad (18)$$

Zur weiteren Reduktion der Daten werden die Koeffizienten 129-512 auf Null gesetzt. Diese werden in 512 Werte aus 128 Gruppen mit je vier gleichen Absorbanzwerten zurücktransformiert. Für die Repräsentation der Infrarotspektren wird je ein Wert aus den 128 Gruppen verwendet, was zu 128 Absorbanzwerten mit einer Auflösung von 40 cm^{-1} zwischen 3500 und 2020 cm^{-1} bzw. von 16 cm^{-1} zwischen 2000 und 560 cm^{-1} führt. Damit entspricht diese Repräsentation weitgehend der von Novic und Zupan vorgeschlagenen, die sich im wesentlichen dadurch unterscheidet, daß Novic und Zupan die Hadamardkoeffizienten noch durch den ersten Hadamard-Koeffizienten dividieren, der dem Integral des Spektrums entspricht.[68] Damit wird eine Normierung der Spektren auf die Gesamtintensität durchgeführt, die aber auch den Verlust der Information über die Gesamtintensität des Spektrums bedeutet. Die resultierenden

Spektren aus beiden Verfahren sind sehr ähnlich, wie das folgenden Beispiel zeigt (Abbildung 26).

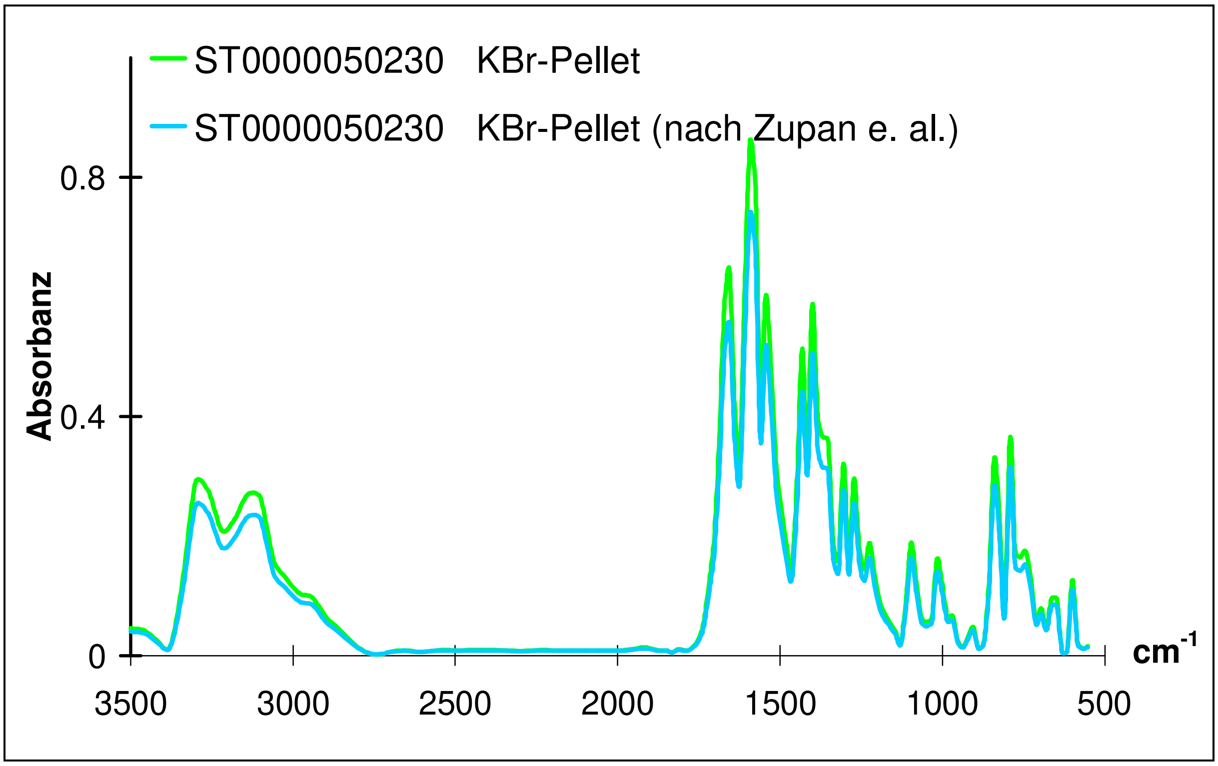


Abbildung 26: Das IR-Spektrum ST0000050230 N-(3,5-Dichlorphenyl)-ethansäureamid repräsentiert nach dem hier verwendeten Verfahren und dem von Zupan et al. vorgeschlagenen.[68]

6.6 Mono-, di- und trisubstituierte Benzolderivate

Benzolderivate sind in der Chemie von erheblicher Bedeutung. Die Aromatizität, 1865 von Kekulé als Denkkonzept vorgestellt, prägte das folgende Jahrhundert der organischen Chemie. Kein wichtiger chemischer Industriezweig indem nicht Benzolderivate eine Rolle spielen oder gespielt hätten. Angefangen mit Anilin und Vanillin in der Aromachemie, dem Anilinschwarz in der Farbenchemie, der Acetylsalicylsäure als Aspirin® oder ASS ratiopharm® bei der Pharmazeutischen Industrie. Die Kunstoffindustrie begann mit Bakelit aus Phenol und Formaldehyd und noch heute ist Styropor oder Polystyrol als Verpackungs- und Dämmaterial allgegenwärtig. DDT aus der Agrarchemie macht uns heute noch Sorgen, dank seiner Stabilität als Benzolderivat. Die Kühl- und Schmiermittelindustrie machte sich die Stabilität des chlorsubstituierten Benzolkerns zunutze und verwendete polychlorierte Biphenyle als Kühl- und Isolierflüssigkeit in ganzen Generationen von Transformatoren. Im Benzin ist es das Benzol, das als Ersatz für $Pb(ET)_4$ für die Klopffestigkeit sorgt und in Sprengstoffen, wie dem TNT oder der Pikrinsäure, sorgt der Benzolkern für die sichere Handhabbarkeit im Gegensatz zum Nitroglyzerin. Kurz: kontrollierbare, vorhersagbare Reaktivität und große Stabilität der Produkte kennzeichenen die Chemie der Benzolderivate und begründen die herausragende Stellung der Benzolderivate in der Chemie.

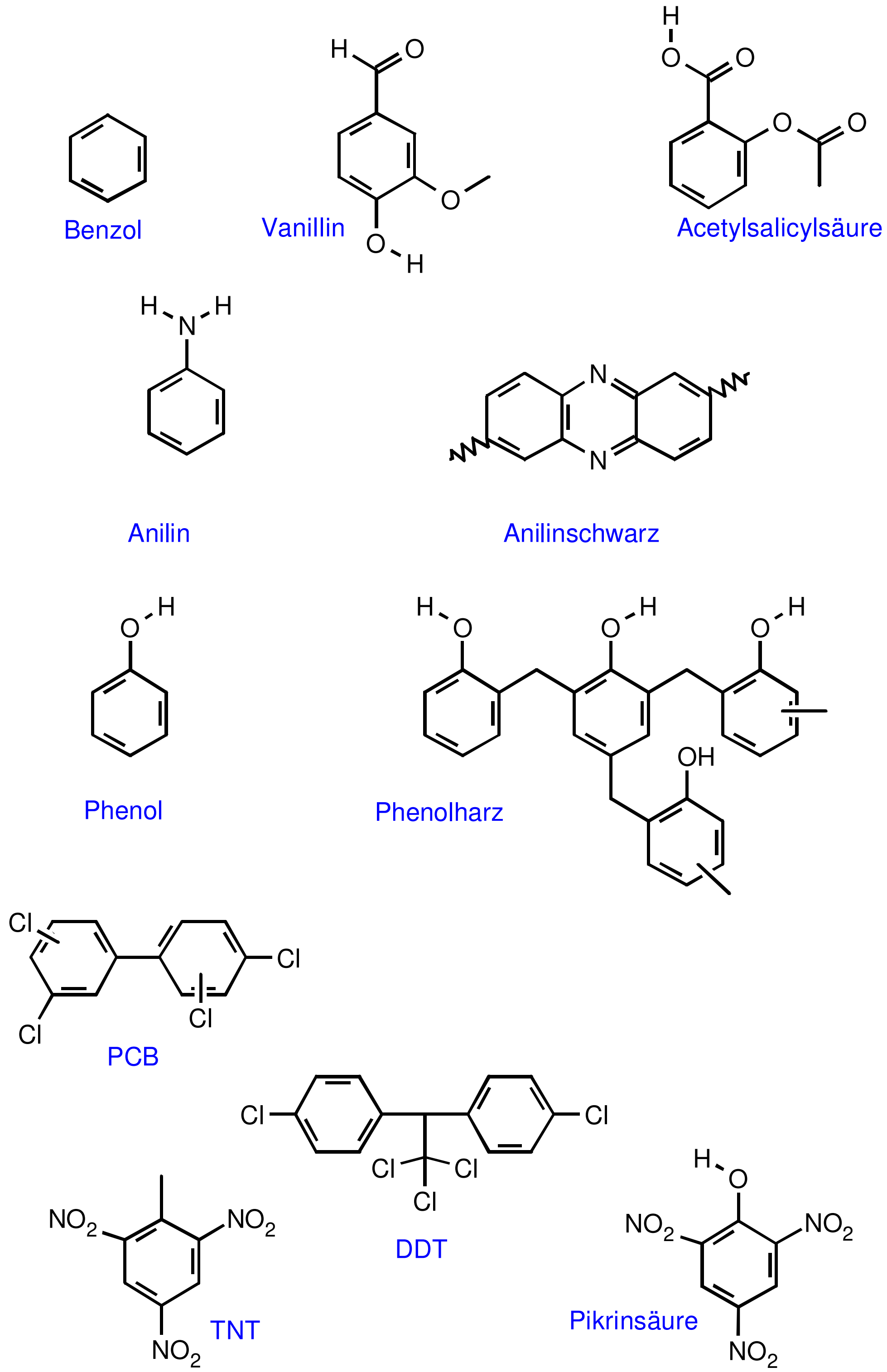


Abbildung 27: Benzol und wichtige Benzolderivate

Als bevorzugte Stoffklasse in der Chemie sind Benzolderivate auch in der SpecInfo-Datenbank vergleichsweise gut repräsentiert. Diese Tatsache und ihre herausragende Bedeutung für die Chemie sind der Grund dafür, daß der erste Versuch zur Simulation von Infrarotspektren mit den Benzolderivaten der SpecInfo-Datenbank unternommen wurde.

6.6.1 Auswahl der Benzolderivate

Für alle Benzolderivate galten die folgenden Auswahlregeln, um sicherzustellen, daß erstens das Benzolgerüst das wesentliche Merkmal im Datensatz ist, und zweitens, daß die Berechnung der physikalischen Eigenschaften der Moleküle mit dem Programmsystem PETRA 6.0[69] möglich ist.

- IR-Vollspektrum in der SpecInfo-Datenbank
- kein Substituent länger als 8-aufeinanderfolgende Bindungen
- keine anderen Elemente als: C, H, N, O, F, Cl, Br.

Dies ergab einen Datensatz von 871 Benzolderivaten, in dem die Substitutionsmuster wie folgt verteilt waren:

Tabelle 5: Verteilung der Substitutionsmuster im Benzoldatensatz

Substitutionsmuster	Anzahl der Derivate	Prozentualer Anteil der Derivate
mono	185	21%
ortho-di	94	11%
meta-di	57	7%
para-di	222	25%
1,2,3-tri	72	8%
1,2,4-tri	190	22%
1,3,5-tri	51	6%
Summe	871	100%

Für einige Benzolderivate enthielt die SpecInfo-Datenbank mehr als ein IR-Spektrum. In einem solchen Fall wurde das Substanzspektrum mit der höheren Reinheit der untersuchten Substanz gewählt. War der Reinheitsgrad der Substanzen gleich, so wurden Flüssigkeitsspektren bevorzugt. Dies führte zu 341 KBr-Spektren und 530 Flüssigkeitsspektren im ausgewählten Datensatz.

Um die Vorhersagequalität der neuronalen Netze beurteilen zu können, wurde der Datensatz in einen Trainings- und einen Testdatensatz aufgeteilt. Zur Aufspaltung des Datensatzes wurde folgende Methode verwendet:

Zuerst wurde ein planares Counterpropagation Netz bestehend aus 30 x 30 Neuronen mit dem gesamten Datensatz trainiert. Das Netz zugleich für die in Kapitel 6.6.2 beschriebene Auswahl der Atomeigenschaft genutzt. Nach dem Erinnerungstest wurde von jedem belegten Neuron dasjenige Molekül für den Trainingsdatensatz selektiert, dessen Code den Gewichten im Eingabeteil des Neurons am ähnlichsten war. Alle anderen Moleküle wurden dem Testdatensatz zugeordnet. Dies ergab bei Verwendung von $A_i = q_{tot,\, i}$ einen Trainingsdatensatz mit 487 Molekülen und einen Testdatensatz von 384 Molekülen.

6.6.2 Optimierung des 3D-MoRSE Codes

Der Benzolring hat einen Durchmesser von etwa 2.6 Å, damit sollte eine maximale codierbare Distanz von 3.0 Å zur Beschreibung ausreichen, wenn man davon ausgeht, daß die Ladungsverhältnisse im und direkt am Benzolring das Infrarotspektrum bestimmen (Eine Annahme die sich später als nur teilweise richtig erwies). Ändert sich die Ordnung einer Bindung, so führt das typischerweise zu einer Veränderung der Bindungslänge um ca. 0.1 Å (sofern es sich um Bindungen zwischen Atomen aus der zweiten Periode handelt). Deshalb sollte das untere Auflösungslimit für den 3D-MoRSE Code bei 0.1 Å liegen. Damit ergeben sich für s_{max} und n die folgenden Werte:

$$s_{max} = \frac{\pi}{r_{min}} = \frac{3.142}{0.1\text{Å}} = 31\text{Å}^{-1} \qquad (19)$$

$$n = \frac{s_{max} \cdot r_{max}}{\pi} = \frac{31\text{Å}^{-1} \cdot 3\text{Å}}{\pi} = 32 \qquad (20)$$

Zur Auswahl der atomaren Eigenschaft A_i wurde der Test auf die Eigenschaften beschränkt, die für molekulare Schwingungen wesentlich sind. In der klassischen Mechanik ist die Schwingungsfrequenz eines Körpers durch die Kraftkonstante und die reduzierte Masse bestimmt. Aus diesem Grund wurde die Masse m_i als Atomeigenschaft in den Test einbezogen. Andererseits besagt das spektroskopische Ausschlußprinzip der Quantenmechanik, daß jeder Schwingungsübergang mit einer Änderung des Dipolmomentes einhergehen muß. Deshalb wurde die partielle Atomladung und die multiplikative Kombination von partieller Atomladung und

Atommasse in die Gruppe der zu testenden Atomeigenschaften aufgenommen. Zum Test wurde der Recalltest des Counterpropagationnetzes mit 30 x 30 Neuronen genutzt, das für jede Atomeigenschaft unbeaufsichtigt trainiert worden war, herangezogen. Die Ergebnisse für die einzelnen Atomeigenschaften (siehe Tabelle 6) weichen nicht sehr deutlich voneinander ab, jedoch ist der Wert für $q_{tot,i}$, der beste unter den drei Varianten.

Tabelle 6: Ergebnisse von Erinnerungstests mit dem gesamten Datensatz an Benzolderivaten für die 3D-MoRSE Codes mit den verschiedenen Atomeigenschaften. Es wurde jeweils der Korrelationskoeffizient zwischen experimentellem und simulierten IR-Spektrum bestimmt und dieser über alle 871 substituierten Benzolderivate gemittelt bzw. dessen Standardabweichung bestimmt.

Atomeigenschaft A_i	mittlerer Korrelationskoeffizient	Standardabweichung des Korrelationskoeffizienten
m_i	0.901	0.104
$q_{tot,i}$	0.915	0.093
m_i * $q_{tot,i}$	0.899	0.106

Damit ergeben sich die folgenden optimierten Randbedingungen für den 3D-MoRSE Code zur Simulation der IR-Spektren von substituierten Benzolderivaten:

- Atomeigenschaft A_i: gesamte partielle Atomladung, $q_{tot,i}$
- Wertebereich von s: 0.0 ... 31.0 $Å^{-1}$
- Anzahl der 3D-MoRSE Codewerte pro Molekül: 32

6.6.3 *Training des CPG-Netzes zur Simulation der IR-Spektren*

Der Datensatz an 871 Benzolderivaten wurde zunächst nach der in Kapitel 6.6.1 vorgestellten Methode in einen Test- und Trainingsdatensatz aufgeteilt. Die Aufteilung ergab einen Trainingsdatensatz aus 487 Benzolderivaten und einen Testdatensatz mit 384 Benzolderivaten.

Zur Darstellung und Training des CPG-Netzes im Computer wurde das Programm *kmap*[70] mit den folgenden Parametern benutzt:

- Netzgröße 25 x 25 Neuronen im Quadrat
- maximale Korrekturentfernung: 8 Neuronen

- Netztopologie: toroidal
- erste Lernrate: 0.95
- automatische Anpassung von Lernrate und Korrekturentfernung

Das Training benötigte 46400 Iterationen und dauerte auf einer Sparc 10-40/Solaris 2.5.1 18 Minuten und 20 Sekunden.

6.6.4 Simulation der IR-Spektren mit $A_i = q_{tot,i}$

Das trainierte und zur Simulation der IR-Spektren verwendete CPG-Netz mit 25 x 25 Neuronen erlaubte die Vorhersage von IR-Spektren für den Testdatensatz mit einem mittleren Korrelationskoeffizienten von 0,71. Der mittlere Korrelationskoeffizient für zwei in der SpecInfo-Datenbank abgespeicherten IR-Spektren derselben (!) Substanz beträgt 0,85 und für zwei zufällig ausgewählte IR-Spektren verschiedener Substanzen liegt der Korrelationskoeffizient im Mittel bei 0,28. Damit ist das Ergebnis positiv zu werten, auch wenn das theoretische Maximum von 0.85 für den mittleren Korrelationskoeffizienten nicht erreicht worden ist. In den folgenden Kapiteln soll das Ergebnis ausführlich untersucht werden und die Ursachen für gute und schlechte Simulationsergebnisse herausgearbeitet werden.

Eine wesentliche Ursache für das Ergebnis ist sicherlich die Zusammensetzung des Datensatzes. Im Datensatz sind die einzelnen Substanzklassen unterschiedlich stark repräsentiert, das bedeutet, es gibt Lücken in der Wissensbasis durch fehlende Beispiele für Verbindungsklassen. Mutmaßliche Fehler in der Datenbank und Ausreißer durch untypische Meßbedingungen erschweren ebenfalls die Simulation, zumal Meßbedingungen einen nicht zu vernachlässigenden Einfluß auf das Spektrum haben. Bei der Auswahl des Datensatzes wurde aber stillschweigend davon ausgegangen, daß ähnliche Verbindungen bei Zimmertemperatur im selben Aggregatzustand (fest, flüssig, gasförmig) vorliegen und deshalb nach der gleichen Methode gemessen wurden, was oft aber nicht immer zutrifft.

Ferner wurde Unmögliches versucht (vgl. Simulation von Anisol). Beispielsweise wird es mit der vorgestellten Methode nicht möglich sein, das IR-Spektrum von Benzol vorherzusagen, weil Benzol als chemischer Grundkörper nur einmal existiert. D.h. es gibt kein Molekül dessen Struktur und Infrarotspektrum dem von Benzol ähnlich ist, da jede Änderung von Symmetrie, Kraftkonstanten, und/oder Massen das Infrarotspektrum stark beeinflussen würde. Leider gibt es keinen Weg, dies bei der Zusammenstellung von Datensätzen bereits im Vorweg zu berücksichtigen.

6.6.4.1 Ergebnis des Erinnerungstests mit dem Trainingsdatensatz

Die 487 Moleküle des Trainingsdatensatzes wurden 341 Neuronen des aus 625 Neuronen bestehenden, trainierten Counterpropagation-Netzes assoziiert. Damit verbleibt rund die Hälfte des Netzes für die Interpolation von IR-Spektren. Die Assoziierung von 487 Molekülen an 341 Neuronen bedeutet aber auch, daß etlichen Neuronen mehrere Moleküle assoziiert wurden, und damit in Teilen des Netzes eine Kompression stattgefunden haben muß, denn bei der Selektion des Trainingsdatensatzes stammten ja alle Moleküle von verschiedenen Neuronen. Bei genauerer Untersuchung des trainierten Netzes findet man, daß von 341 belegten Neuronen 224 einfach belegt sind, 94 zweifach, 17 dreifach und 6 vierfach. Insgesamt sieht die Verteilung der Moleküle im trainierten Netz wie folgt aus:

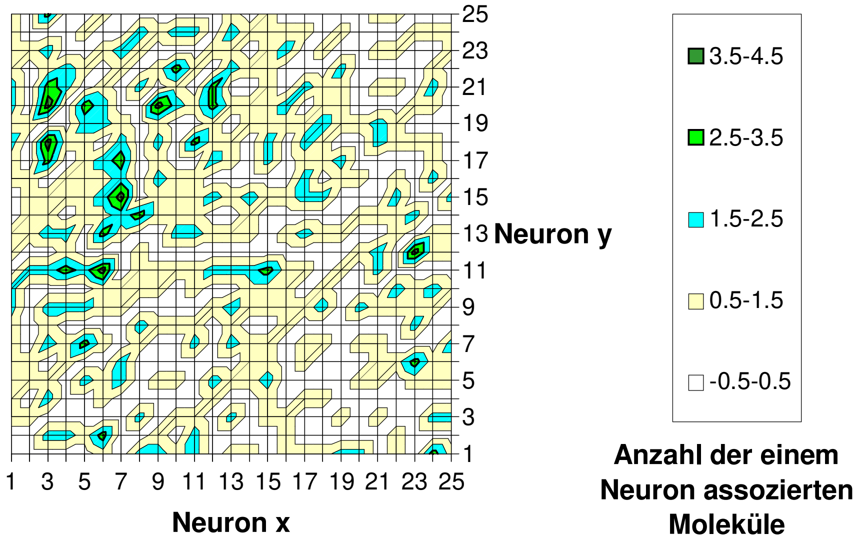


Abbildung 28: Verteilung des Trainingsdatensatzes auf die Neuronen des trainierten CPG-Netzes

Die Inspektion der mehrfach belegten Neuronen zeigt, daß sich die assoziierten Moleküle meistens nur in den aliphatischen Seitenketten unterscheiden, was selten zu einem Verlust an Simulationsqualität führt, da der Unterschied zwischen einer Ethyl- oder Propylseitenkette im IR-Spektrum so gut wie nicht sichtbar ist. Allerdings muß dies nicht immer gelten und wird vor allem bei unpolaren Molekülen problematisch, wo die Unterscheidungsfähigkeit des 3D-

MoRSE Codes aufgrund der kleinen Werte von $q_{tot,i}$ als der verwendeten Atomeigenschaft nachläßt, aber aufgrund der geringen Gesamtintensität des Spektrums Änderungen bei intensitätsschwachen Gerüst- und CH-Schwingungen deutlicher hervortreten.

Abbildung 29 zeigt den Inhalt der sechs vierfach belegten Neuronen aus dem Erinnerungstest. Als Gründe für die Vierfachbelegungen lassen sich aus der Grafik drei Ursachen ableiten, wenn man zusätzlich den Anteil der Stoffklassen am gesamten Datensatz in Betracht zieht. Eine der möglichen Ursachen ist ein großer Unterschied einer kleinen Gruppe von Molekülen zum Rest des Datensatzes, wie es die vier Phenyloxiranderivate von Neuron (3,18) zeigen. In diesem Fall werden alle acht Phenyloxiranderivate des Gesamtdatensatzes zwei benachbarten Neuronen zugeordnet, wobei alle vier Phenyloxiranderivate des Trainingsdatensatzes dem Neuron (3,18) zugeordnet werden.

Nicht weit entfernt von den Phenyloxiranderivaten werden dem Neuron (3,20) vier Ether zugeordnet. Diesmal ist die Ursache für die Vierfachbelegung des Neurons (3,20) der hohe Anteil der Phenolether am gesamten Datensatz. 65 der 871 Moleküle des Datensatzes sind Phenolether. Kommt dann eine Schwäche des Codes bei der Unterscheidung von aliphatischen bzw. Halogensubstituenten hinzu, wird die Kompressionen im Datenraum bei den Phenolethern verständlich, sie führt aber lediglich bei einem Molekül zu Problemen mit der Simulation.

Sinkt die Polarität der Moleküle weiter, werden die Probleme bei der Simulationsqualität größer. Ganz deutlich zeigt sich dies bei den Benzolderivaten ohne Heteroatom, die dem Neuron (7,15) assoziiert wurden. Die geringsten *rms*-Werte zwischen Neuron und 3D-MoRSE Code aller Verbindungen aus Abbildung 29 führen im Mittel zu den zweitschlechtesten Simulationsergebnissen im Vergleich der vierfach belegten Neuronen. Eine geringe Polarität von Verbindungen ist damit wohl die gefährlichste Ursache für Kompressionen im Datenraum.

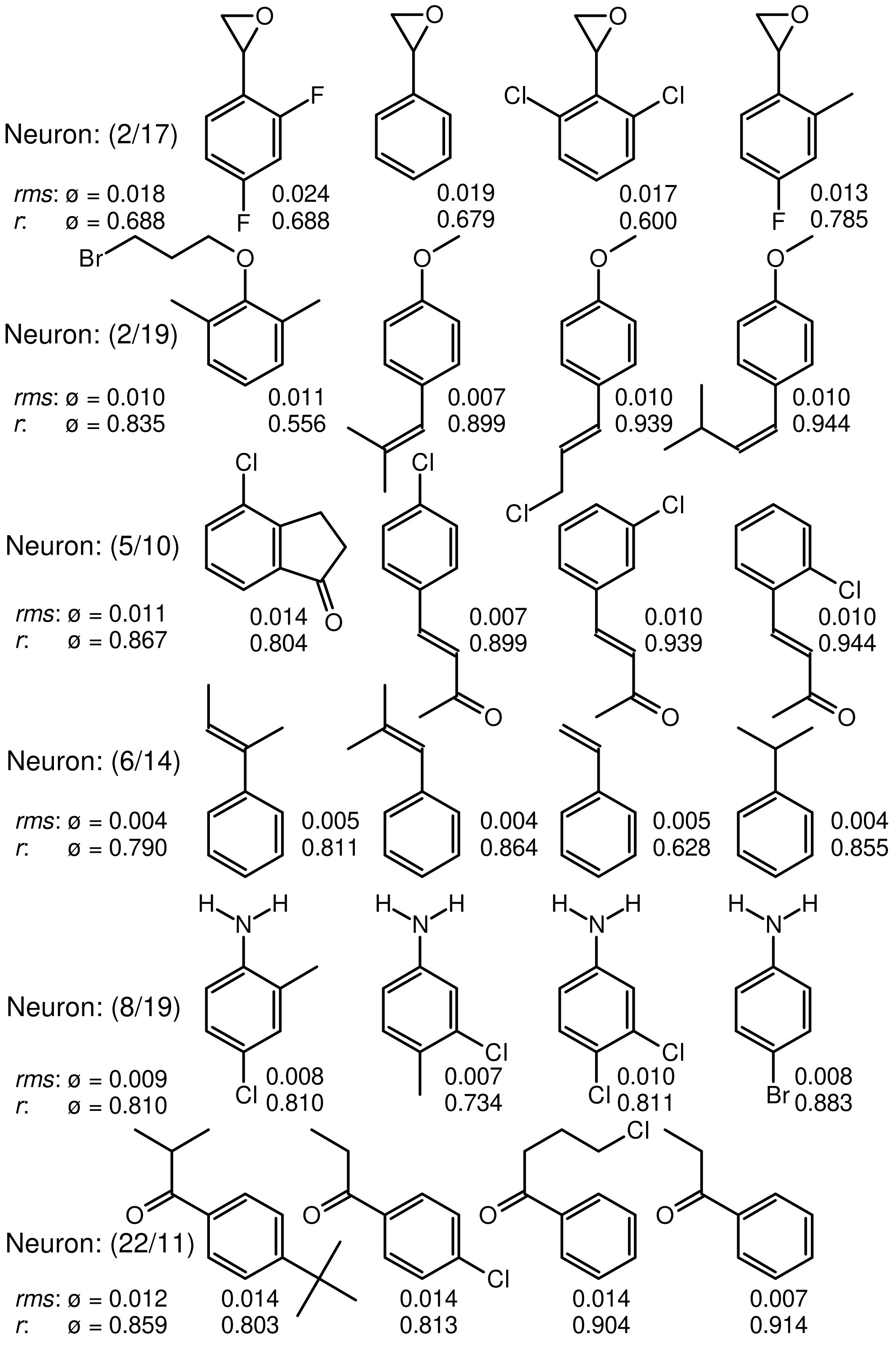


Abbildung 29: Die Verbindungen des Trainingsdatensatzes von den Neuronen, denen im Erinnerungstest vier Moleküle assoziiert wurden. Angeben sind der *rms*-Wert zwischen Neuron und Code sowie der Korrelationskoeffizient zwischen experimentellem und simulierten (vom Neuron gespeicherten) Spektrum.

Den Effekt der abnehmenden Polarität auf den 3D-MoRSE Code zeigt Tabelle 7. Angegeben ist der *rms*-Wert zwischen dem 3D-MoRSE Code von Benzol als Grundkörper, und dem 3D-MoRSE Code des Benzolderivates. Klar zeigt sich: je geringer die Polarität eines Benzolderivates, desto ähnlicher wird der 3D-MoRSE Code des Derivats dem des Grundkörpers Benzol. Aus diesem Grund landeten die vier Benzolderivate ohne Heteroatom trotz größerer Unterschiede in der Seitenkette auch gemeinsam auf dem Neuron (7,15), wie auch sechs weitere aus dem Testdatensatz. Insgesamt sind alle 23 monosubstituierten Benzolderivate ohne Heteroatom, die dieselben Substrukturen aufweisen wie die vier Trainingsmoleküle, im Umkreis von 2 Neuronen um Neuron (7,15) zu finden.

Tabelle 7: *rms*-Werte für den 3D-MoRSE Code zwischen Benzolderivat und Benzol.

Verbindung	Styrol	Phenyoxiran	Phenylethylketon
rms zu Benzol	0.064	0.188	0.265

Im nachhinein bleibt festzuhalten, daß zur besseren Unterscheidung unpolarer Moleküle ein von der Atomladung unabhängiger Codeteil sinnvoll gewesen wäre. Dies bleibt aber späteren Untersuchungen überlassen.

Die Verteilung des Korrelationskoeffizienten für die 487 Moleküle des Trainingsdatensatzes im Erinnerungstest (Abbildung 30) zeigt, daß das CPG-Netz in der Lage war die Information des Trainingsdatensatzes zu lernen.

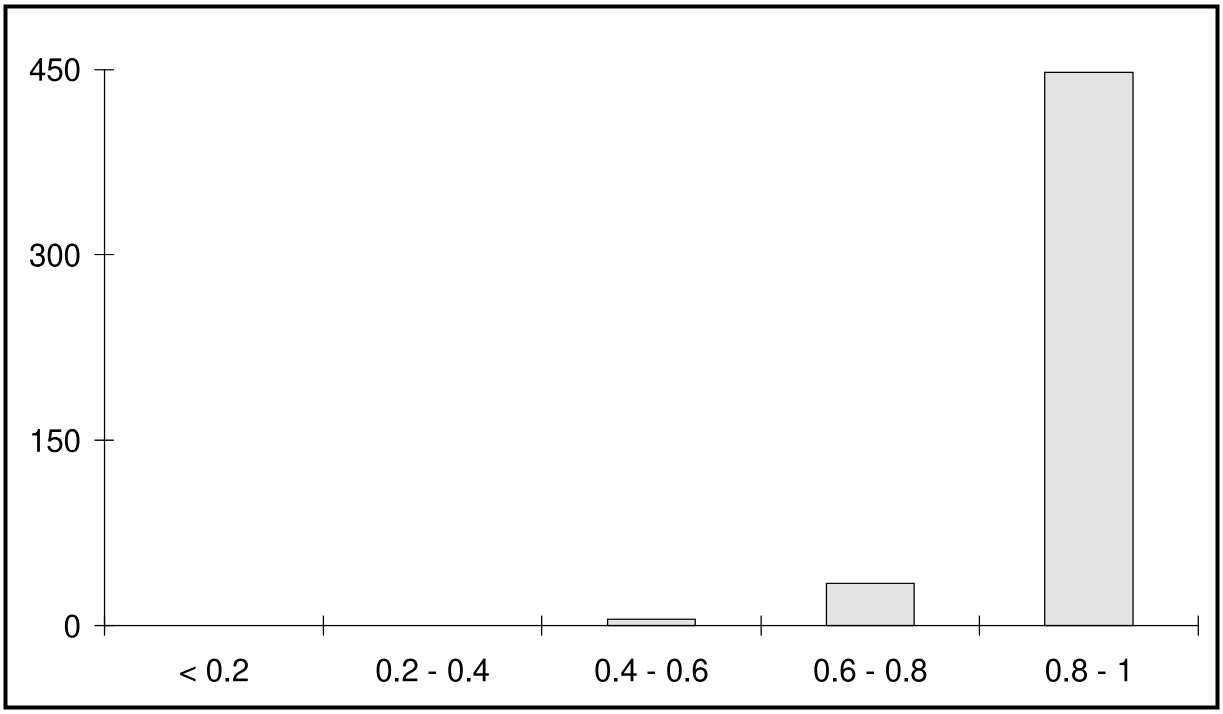


Abbildung 30: Verteilung des Korrelationskoeffizienten beim Erinnerungstest mit dem Trainingsdatensatz. Für 92 % der Moleküle ist der Korrelationskoeffizient größer als 0.8, der kleinste Korrelationskoeffizient beträgt 0.489.

Insgesamt wurde für 92.0 % der Moleküle des Trainingsdatensatzes im Erinnerungstest ein Korrelationskoeffizient von über 0.8 erreicht, bei 217 Molekülen war der Korrelationskoeffizient größer als 0.99, womit für die Hälfte der Moleküle quasi das theoretische Maximum von 1 erreicht wurde. Damit ist klar, daß das Netz in der Lage war die Informationen des Trainingsdatensatzes über den Zusammenhang zwischen Strukturcode und Infrarotspektrum weitestgehend zu speichern.

Abbildung 31 zeigt das Ergebnis einer der sieben besten Simulationen des Trainingsdatensatzes. Bei diesen Simulationen wurde der maximale Korrelationskoeffizient von 1.0 erreicht.

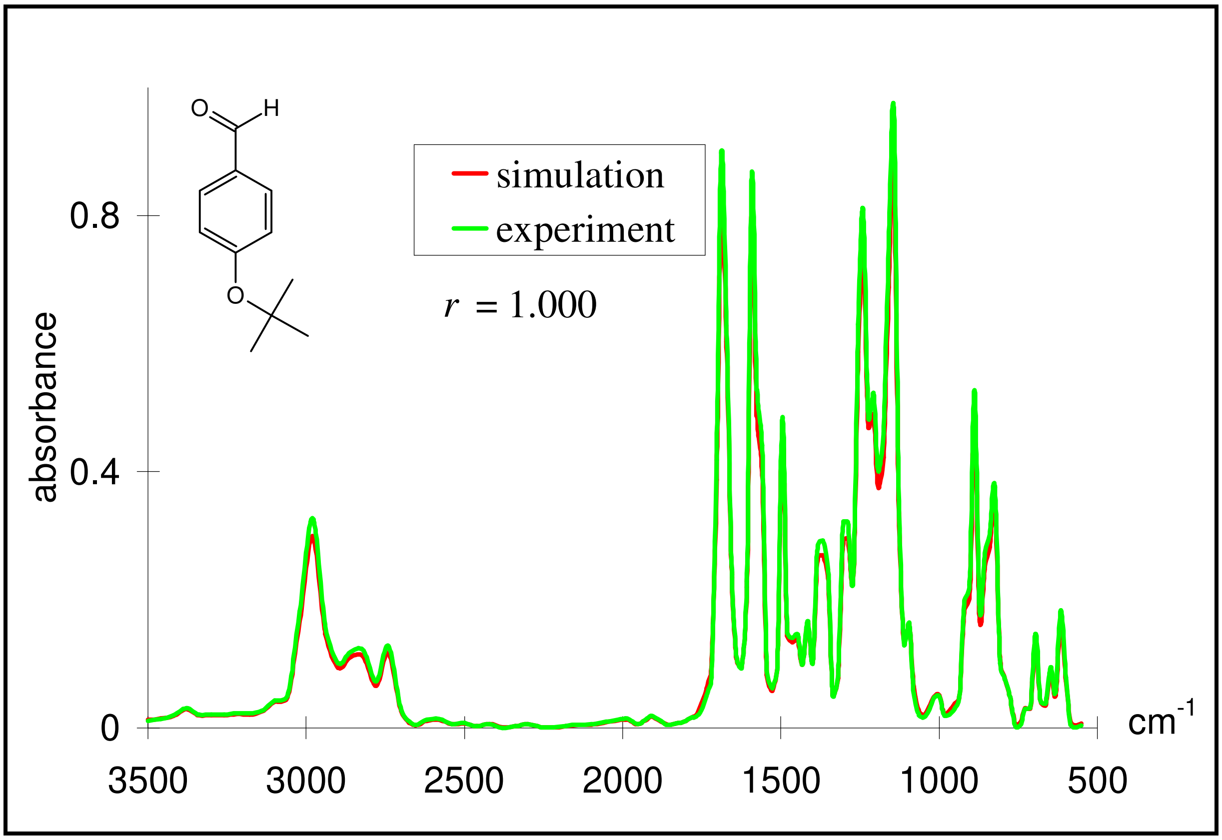


Abbildung 31: Simulation des IR-Spektrums von 4-(1,1-Dimethylethyl)-1-methanoylbenzol aus dem Trainingsdatensatz im Rahmen des Erinnerungstests.

Die folgende Simulation (Abbildung 32) mit einem Korrelationskoeffizienten von 0.803 markiert bereits die untere Grenze des Bereichs in dem 92% der Simulationen des Erinnerungstest liegen. Relativ typisch für das Verfahren ist die Tendenz zu geringeren Gesamtintensitäten in den Teilbereichen der simulierten IR-Spektren gegenüber den experimentellen Spektren, in denen sich die benachbarten Neuronen bzw. die IR-Spektren ihrer Trainingsmoleküle unterscheiden. Diese Tendenz ist auch hier wieder ein Grund für den relativ geringen Korrelationskoeffizienten von 0.803 trotzt der guten Übereinstimmung der Bandenlagen, die alle bis auf einen kleinen Peak bei etwa 700 cm^{-1} übereinstimmen. Der Grund für die Tendenz zu geringeren Intensitäten bei simulierten Infrarotspektren liegt in der gegenseitigen Beeinflussung der Neuronen im Training, die im Durchschnitt zu einer Intensitätsabnahme führt.

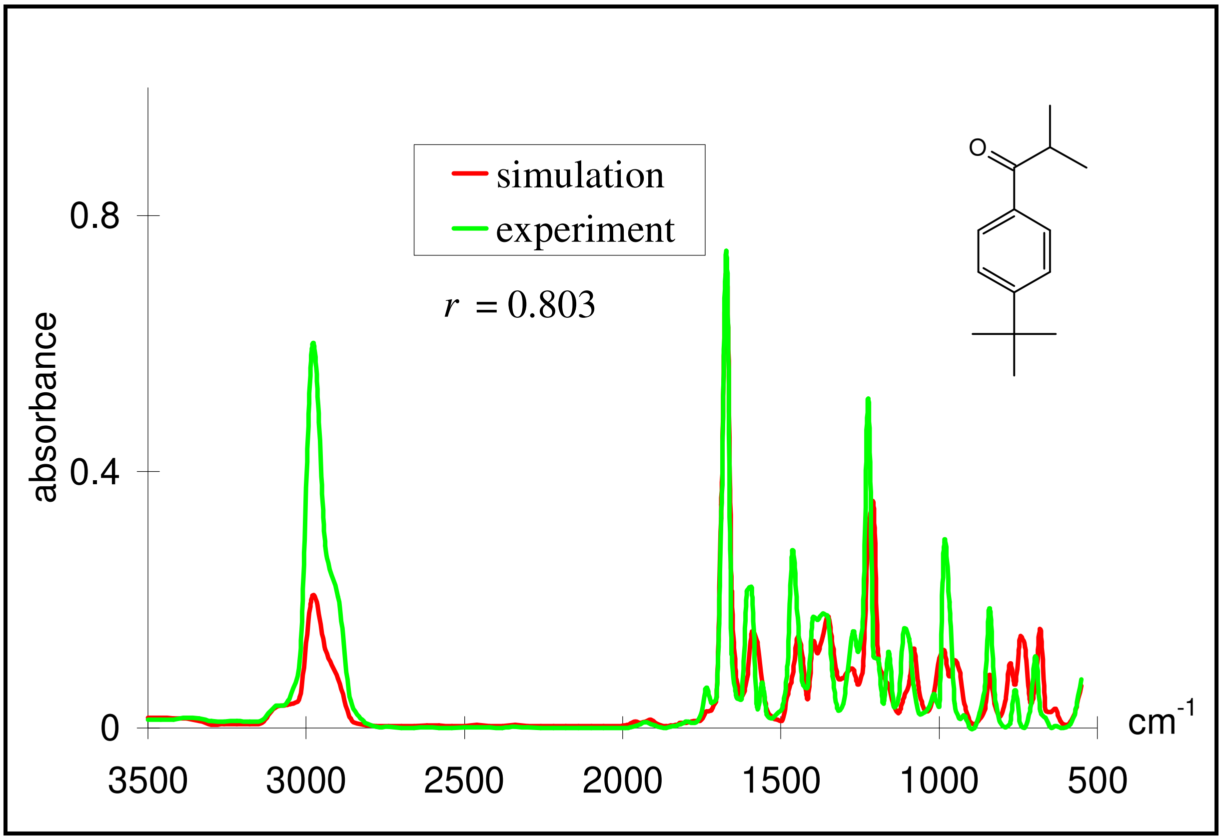
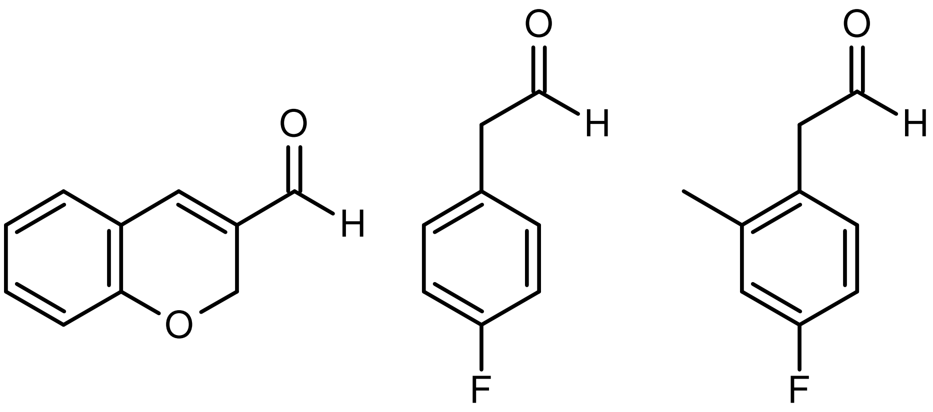


Abbildung 32: Simulation von 1-(2-Methyl-1-oxopropyl)-4-(1,1-dimethylethyl)-benzol mit einem Korrelationskoeffizienten von 0.803 im Rahmen des Erinnerungstests.

Abbildung 33 zeigt die schlechteste Simulation des Erinnerungstests mit einem Korrelationskoeffizienten von 0.489. Der Grund hierfür ist nicht vollständig aufzuklären, zum einem liegt **TA1** am Rande des Datenraums und es ist fraglich ob die Codewerte das Molekül gut repräsentieren, zum anderen bestehen, aufgrund der nachfolgenden quantenmechanischen Berechnungen, Zweifel an der Qualität des experimentellen Spektrums.



TA1 **TA2** **TA3**

Schema 1: TA1 und die anderen Trainingsmoleküle vom Neuron (5,7)

Die Lage von **TA1** am Randes des Datenraums wird daran deutlich, daß es im gesamten Datensatz kein zweites Molekül mit dem Ringsystem von **TA1** gibt. Das neben **TA1** noch (4-Fluorphenyl)-ethanal (**TA2**) und (2,4-Difluorphenyl)-ethanal (**TA3**) dem zur Simulation genutzten Neuron (5,7) assoziiert werden läßt es ferner fraglich erscheinen, ob **TA1** durch die Codewerte gut repräsentiert wurde. Die Ursache für die schlechte Repräsentation von **TA1** muß dabei nicht notwendigerweise im 3D-MoRSE Code liegen, da bei der Ladungsberechnung mit Hilfe der PEOE-Methode eine Aufteilung der Ladung zwischen σ- und π-Ladung im Verhältnis von 1:1 vorgenommen wird, womit die zu erwartenden Ladungsunterschiede zwischen konjugierter (**TA1**) und nicht konjugierter Aldehydfunktion (**TA2** und **TA3**) zumindest teilweise nivelliert werden.

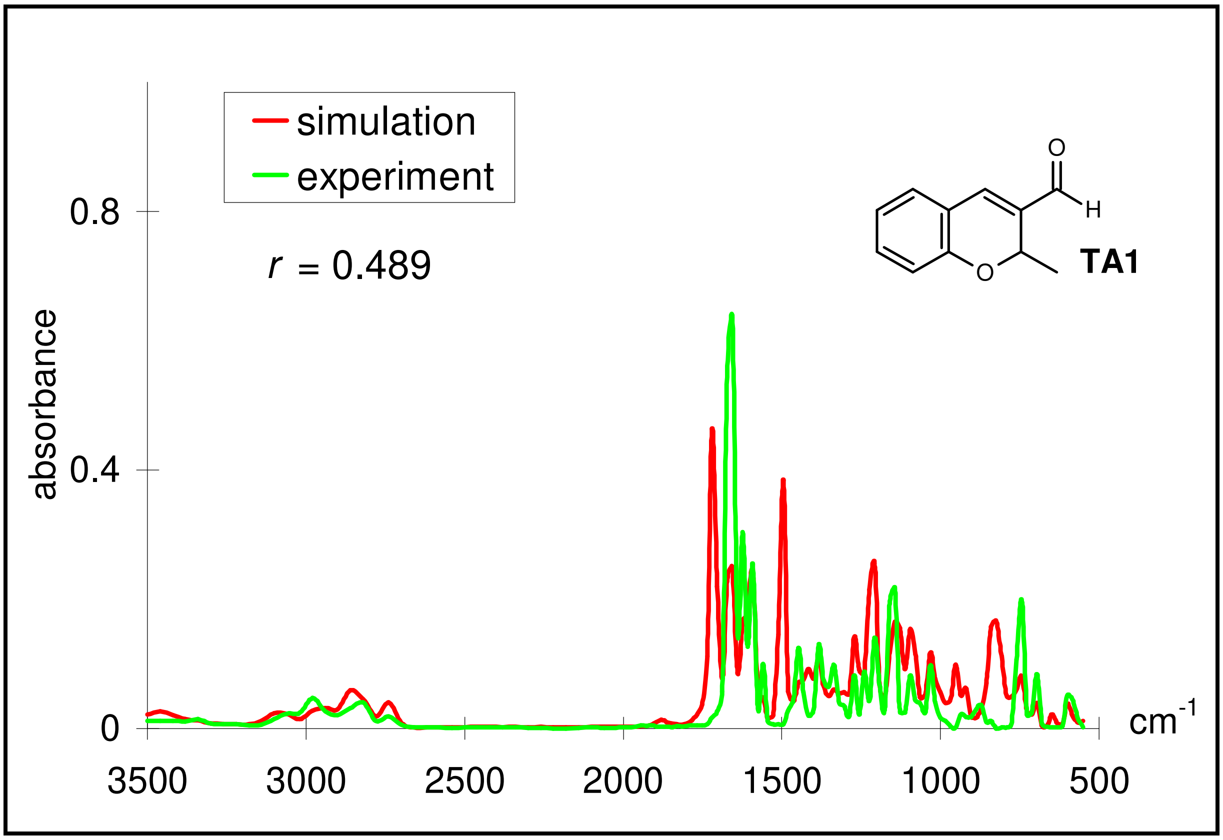


Abbildung 33: Die schlechteste Simulation des Erinnerungstestes für die Benzolderivate.

Betrachtet man das experimentelle Spektrum im Vergleich zur Simulation, fallen einem die Unterschiede bei 1700 cm^{-1} und 1500 cm^{-1} deutlich ins Auge. Ein genauerer Vergleich des Bereiches von 1800 bis 1450 cm^{-1} in den die Ergebnisse einer DFT Rechnung (B3-LYP/6-311G(d)(5d,7f)) einbezogen wurden ist in Abbildung 34 zu sehen.

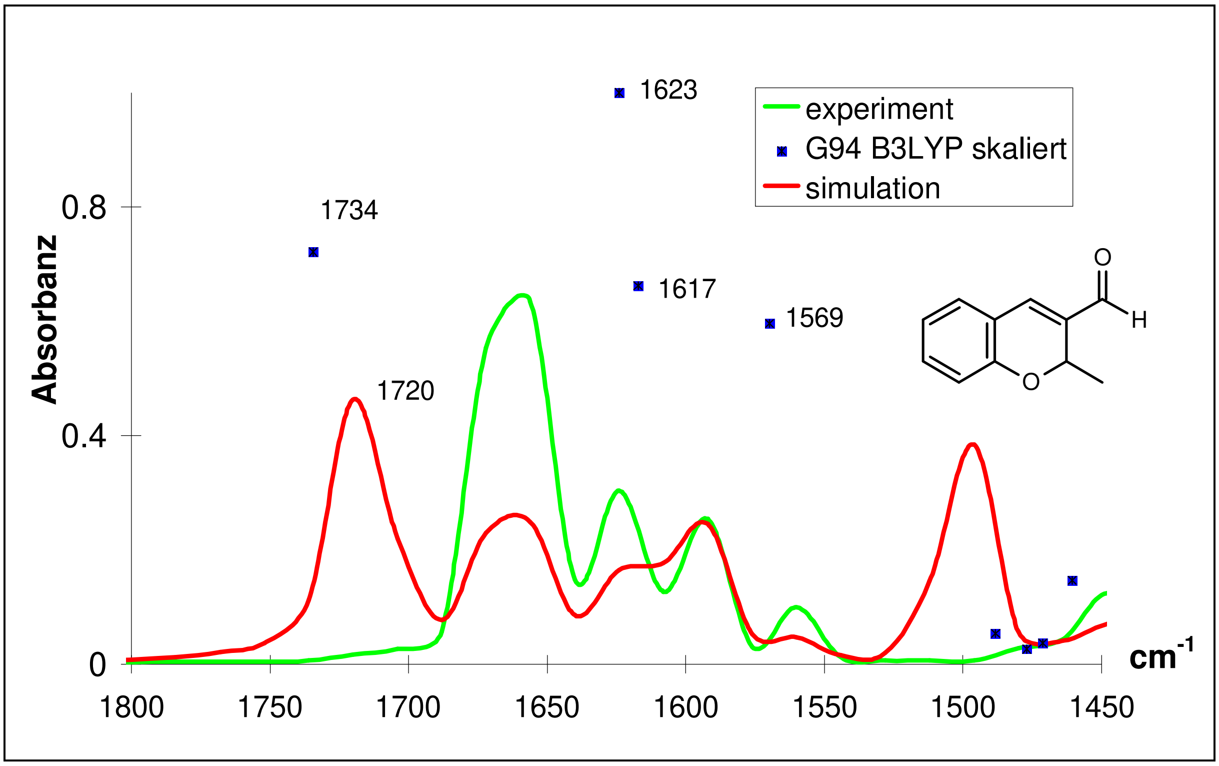


Abbildung 34: Spektrenausschnitt von 1800 und 1450 cm^{-1} des Spektrums von 2-Methyl-*2H*-chromene-3-carbaldehyde. Deutlich zu sehen ist das Fehlen der Carbonylbande im experimentellem Spektrum und die fehlende Absorption bei 1490 cm^{-1}.

Vergleicht man die drei Spektren miteinander, wird das Fehlen der Carbonylbande im experimentellem Spektrum deutlich, die sich im simulierten Spektrum bei 1720 cm^{-1} und im berechneten Spektrum bei 1734 cm^{-1} findet. Bei 1490 cm^{-1} finden sich sowohl im simulierten als auch im berechneten Spektrum CH-Schwingungen der Methylgruppe, die aber im experimentellen Spektrum fehlen. Insgesamt bestehen aus diesem Grund erhebliche Zweifel an der Qualität des experimentellen Spektrums. Das schlechte Simulationsergebnis kann insofern als Ausreißer gewertet werden.

6.6.4.2 Ergebnisse mit dem Testdatensatz

Für mehr als 75% der 384 Moleküle im Testdatensatz konnte ein Korrelationskoeffizient im Bereich von 1.0 und 0.6 zwischen simuliertem und experimentellem IR-Spektrum erreicht werden. Für mehr als 22.5 % der Simulation lag der Korrelationskoeffizient mit 0.85 bis 1.0 gleich oder höher als der durchschnittliche Korrelationskoeffizient für zwei verschiedene Spektren derselben Substanz aus der SpecInfo-Datenbank. Abbildung 35 zeigt die Verteilung des Korrelationskoeffizienten für den Testdatensatz. Wie aus der Verteilung zu erkennen ist, wird für knapp die Hälfte der Verbindungen (160 = 42%) ein Korrelationskoeffizient zwischen 0.6 und 0.8 erreicht. Für 35% der Verbindungen (136) liegt der Korrelationskoeffizient über 0.8. Für 88 Verbindungen, entsprechend 23% des Testdatensatzes, liegt der Korrelationskoeffizient unter 0.6 und damit, wie sich später zeigen wird, zu niedrig für ein verwertbares Simulationsergebnis.

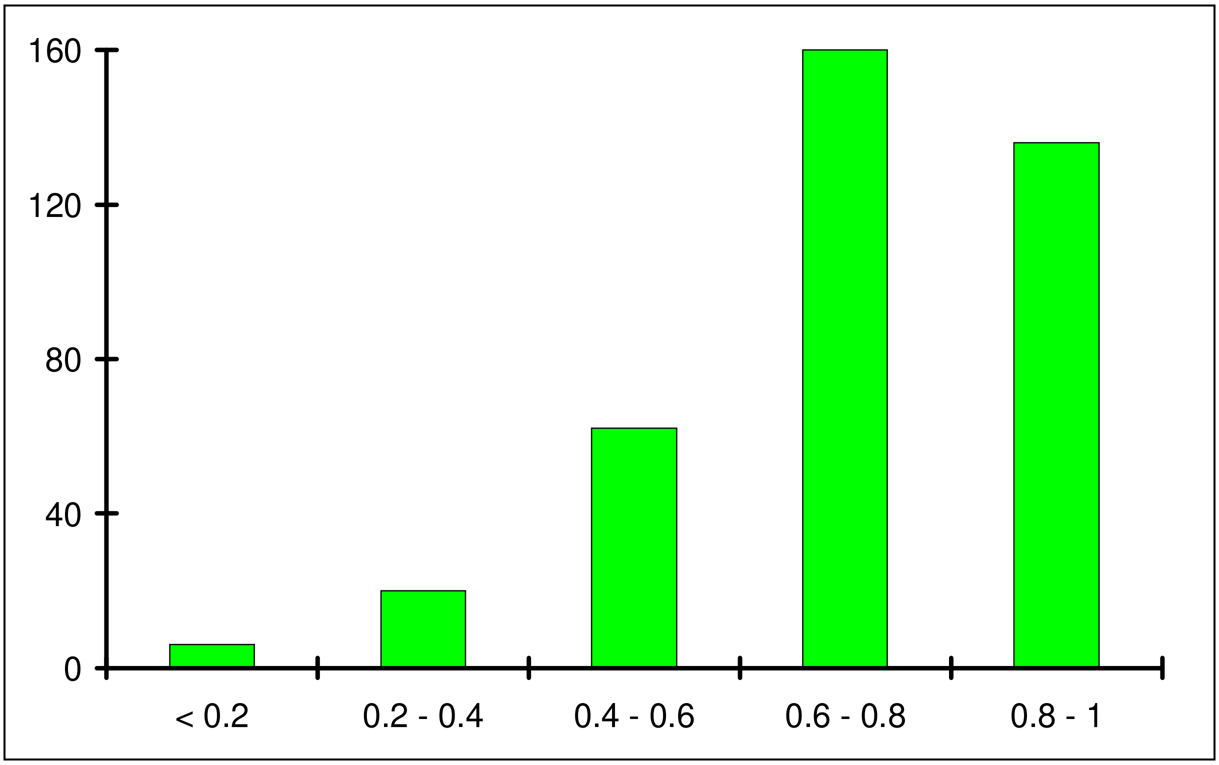


Abbildung 35: Verteilung des Korrelationskoeffizienten zwischen experimentellem und simuliertem Infrarotspektrum für die 384 Verbindungen des Testdatensatzes.

6.6.4.3 10 Beispiele aus den 25 besten Simulationen des Testdatensatzes

Die folgenden zehn Beispiele aus den 25 besten Simulationen des Testdatensatzes (Abbildung 36 - Abbildung 45) sollen die Anwendungsbreite der Methode illustrieren und zeigen, welche Bedingungen für gute Simulationen notwendig waren. Zusammen mit den folgenden Beispielen für geringere Korrelationskoeffizienten soll die Qualität der Simulationen in Verbindung mit dem Wert des Korrelationskoeffizienten diskutiert werden.

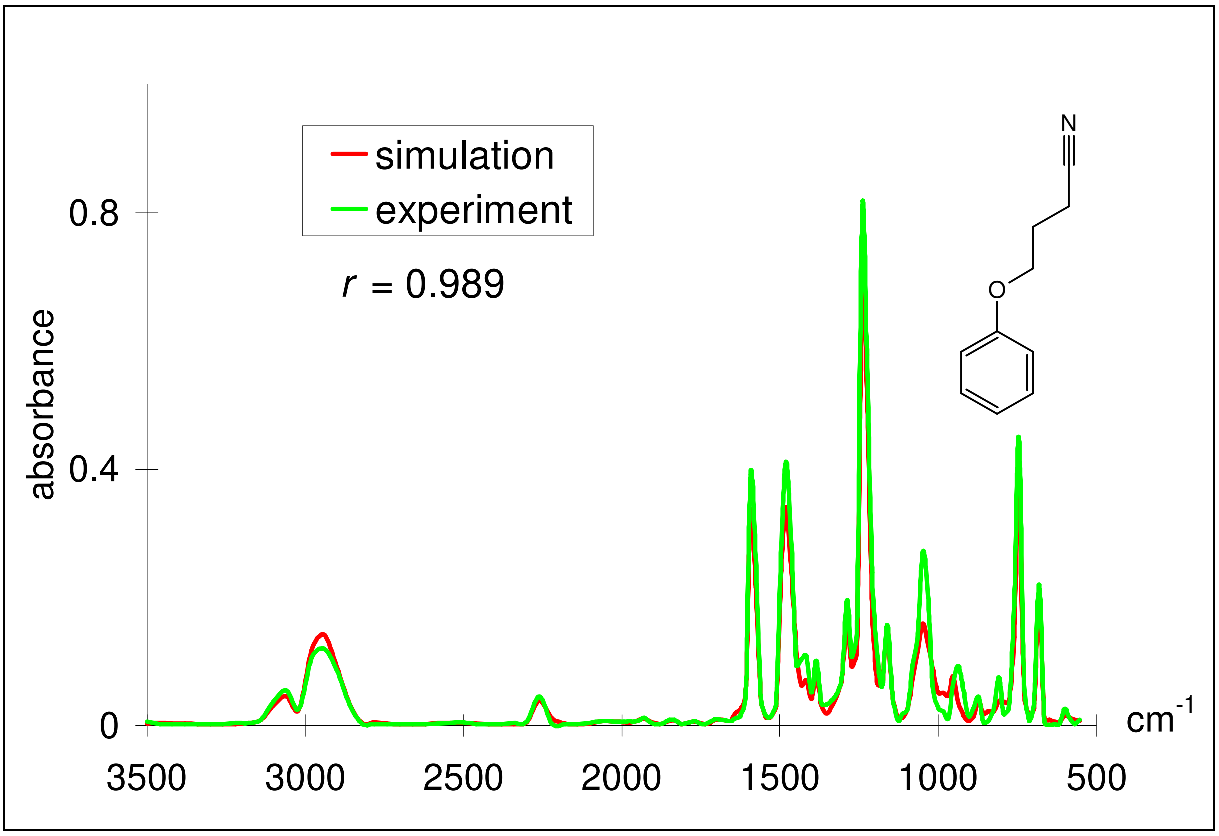


Abbildung 36: Die beste IR-Spektrensimulation des Testdatensatzes mit einem Korrelationskoeffizient von 0.989 zwischen simuliertem und experimentellem IR-Spektrum.

Das simulierte und experimentelle IR-Spektrum von 4-Phenoxybutannitril unterscheidet sich nur geringfügig in der Intensität, alle Banden wurden aber korrekt reproduziert. Ausschlaggebend für die Qualität dieser Simulation war das Vorhandensein von 5-Phenoxypentannitril, das sich von 4-Phenoxybutannitril nur durch den um eine CH_2-Gruppe längeren Substituenten unterscheidet.

Die IR-Spektren Simulation für 4-Phenoxybutannitril ist damit typisch für gute Simulation, die den zwei folgenden Merkmalen genügt:

- Verbindungsklasse = Phenolether
- Ähnliche nur durch eine CH_x-Gruppe unterschiedene Verbindung im Trainingsdatensatz (hier 5-Phenoxy-pentan-nitril)

Die zweitbeste IR-Spektren Simulation für Butoxybenzol bestätigt dies, da diese Simulation exakt dieselben Merkmale aufweist. So wurde für die IR-Spektrensimulation von Butoxybenzol aus dem Testdatensatz ein Korrelationskoeffizient von 0.980 erreicht. Das Molekül im Trainingsdatensatz war hier (4-Brombutoxy)-benzol und zeigt damit eine weitere Quelle IR-spektroskopisch sehr ähnlicher Moleküle: die Halogenderivate. Damit wird zugleich deutlich, wie wichtig es war, die partielle Atomladung $q_{tot,i}$ als Atomeigenschaft zu verwenden und nicht die Ordnungszahl. Denn, die Ordnungszahl von Wasserstoff und Brom ist grundverschieden, $q_{tot,i}$ muß es aber nicht sein.

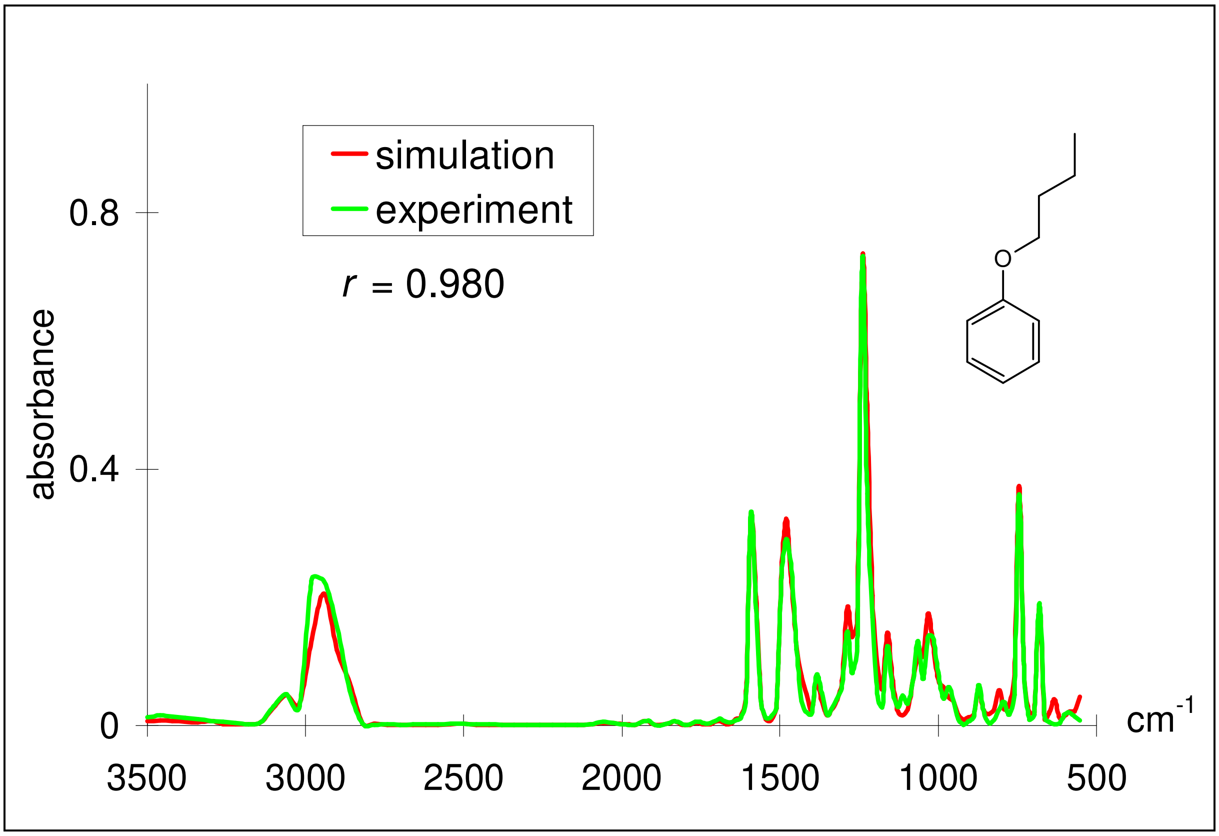


Abbildung 37: Zweitbeste IR-Spektrensimulation des Testdatensatzes für Butoxybenzol mit einem Korrelationskoeffizienten von 0.980 zwischen experimentellem und simuliertem Spektrum.

Das simulierte IR-Spektrum von Butoxybenzol weist nur geringe Intensitätsunterschiede gegenüber dem experimentellem Spektrum auf sowie die Verschiebung eines kleinen Peaks bei 625 cm^{-1}.

Das Interessante an den beiden folgenden Simulationen, der dritt- und sechstbesten, ist, daß für beide Simulationen dasselbe Neuron des neuronalen Counterpropagation-Netzes genutzt wurde, Neuron (3,7). Die experimentellen IR-Spektren von Essigsäurebenzylester und Essigsäure-(4-methyl-benzyl)-ester unterscheiden sich nur gering, so daß mit demselben vorhergesagten

IR-Spektrum eine gute Simulation für beide Verbindungen erreicht werden kann. Das vorhergesagte Spektrum basiert auf dem *meta*-Derivat, dem Essigsäure-(3-methyl-benzyl)-ester. Auf den ersten Blick fallen die Unterschiede zwischen den Spektren kaum auf. Auf den zweiten Blick fällt die etwas geringere Intensität des IR-Spektrums von Essigsäure-(4-methyl-benzyl)-ester auf, was auf die symmetrische Substitution dieses Benzolderivates zurückzuführen sein dürfte. Alle Spektren weichen ferner im Bereich von 675-800 cm^{-1} voneinander ab, da die Banden in diesem Bereich aber gegenüber den anderen Schwingungen der Benzylester relativ intensitätsschwach sind, fallen die Abweichungen kaum auf.

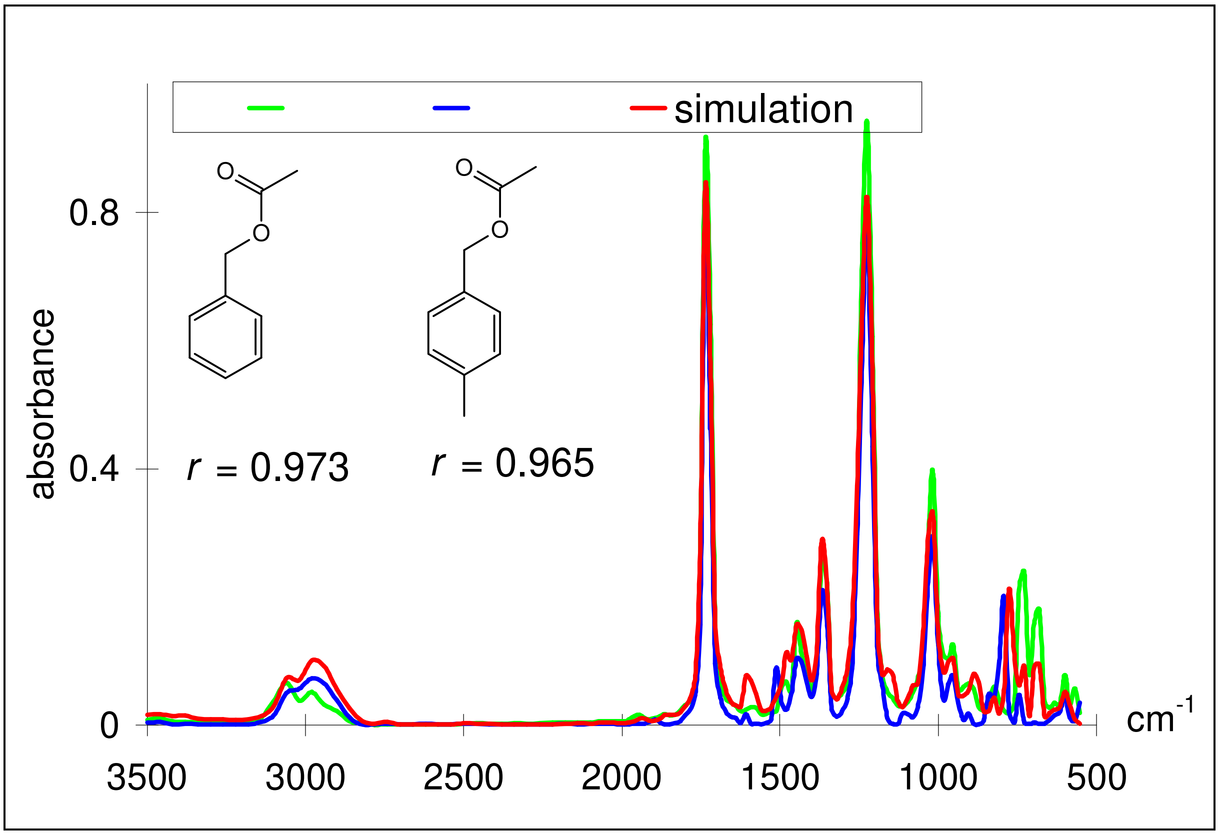


Abbildung 38: Dritt- und sechstbeste Simulation des Testdatensatzes. Beide Verbindungen fallen im Test auf Neuron (3,7), von dem das simulierte Spektrum stammt.

6.6.4.4 Testbeispiele aus den 25 besten Simulationen - Wasserstoffbrückenbindungen und andere Simulationsschwierigkeiten

Wasserstoffbrückenbindungen, insbesondere intermolekulare, stellen bei der Simulation von IR-Spektren eine besondere Herausforderung dar, da nicht einfach von einem aciden Proton auf Anzeichen für Wasserstoffbrückenbindungen im IR-Spektrum geschlossen werden kann. Wasserstoffbrückenbindungen sind im IR-Spektrum durch Verbreiterung und Absinken der

Wellenzahl für XH-Schwingungsbanden erkennbar. Ob eine Verbindung Wasserstoffbrückenbindungen ausbildet, hängt von der Molekülgeometrie sowie den Temperatur- und Druckverhältnissen ab. Die Molekülgeometrie und damit die 3D-Strukur eines Moleküls entscheidet über die prinzipielle Möglichkeit zur Ausbildung von Wasserstoffbrückenbindungen. Die Temperatur- und Druckverhältnisse, unter denen das IR-Spektrum aufgenommen wird, entscheiden darüber ob Wasserstoffbrückenbindungen ausgebildet werden. So fehlen beispielsweise in GC/IR-Spektren häufig die in KBr- und Flüssigkeitsspektren vorhandenen Wasserstoffbrükkenbindungen. Deshalb müssen bei der Simulation von IR-Spektren die Aufnahmebedingungen berücksichtigt werden. Im Fall des Benzoldatensatzes geschieht dies implizit, da in der Regel für ähnliche Verbindungen, die über das CPG-Netz das Simulationsergebnis bestimmen, auch ähnliche Meßbedingungen gewählt wurden.

Als erstes Beispiel für die Simulation eines Infrarotspektrums, das Anzeichen für Wasserstoffbrückenbindungen enthält und dessen Simulation zu den besten Simulationen des Testdatensatzes gehört, wird das Simulationsergebnis für 1-Phenyl-butanol vorgestellt. Dieser sekundäre Alkohol zeigt mit der OH-Bande bei 3400 cm^{-1} deutlich den Einfluß einer Wasserstoffbrückenbindung. Die OH-Bande liegt hier ca. 200 cm^{-1} tiefer als die Bande einer Hydroxygruppe, die nicht an einer Wasserstoffbrückenbindung beteiligt ist. Wie der erreichte Korrelationskoeffizient von 0.961 zeigt, wurde nahezu das gesamte Infrarotspektrum korrekt wiedergegeben. Kleinere Abweichungen finden sich lediglich im Bereich der CH-Streckschwingung und bei 1050 cm^{-1}, die OH-Banden bei 3400 cm^{-1} sind dagegen nahezu identisch.

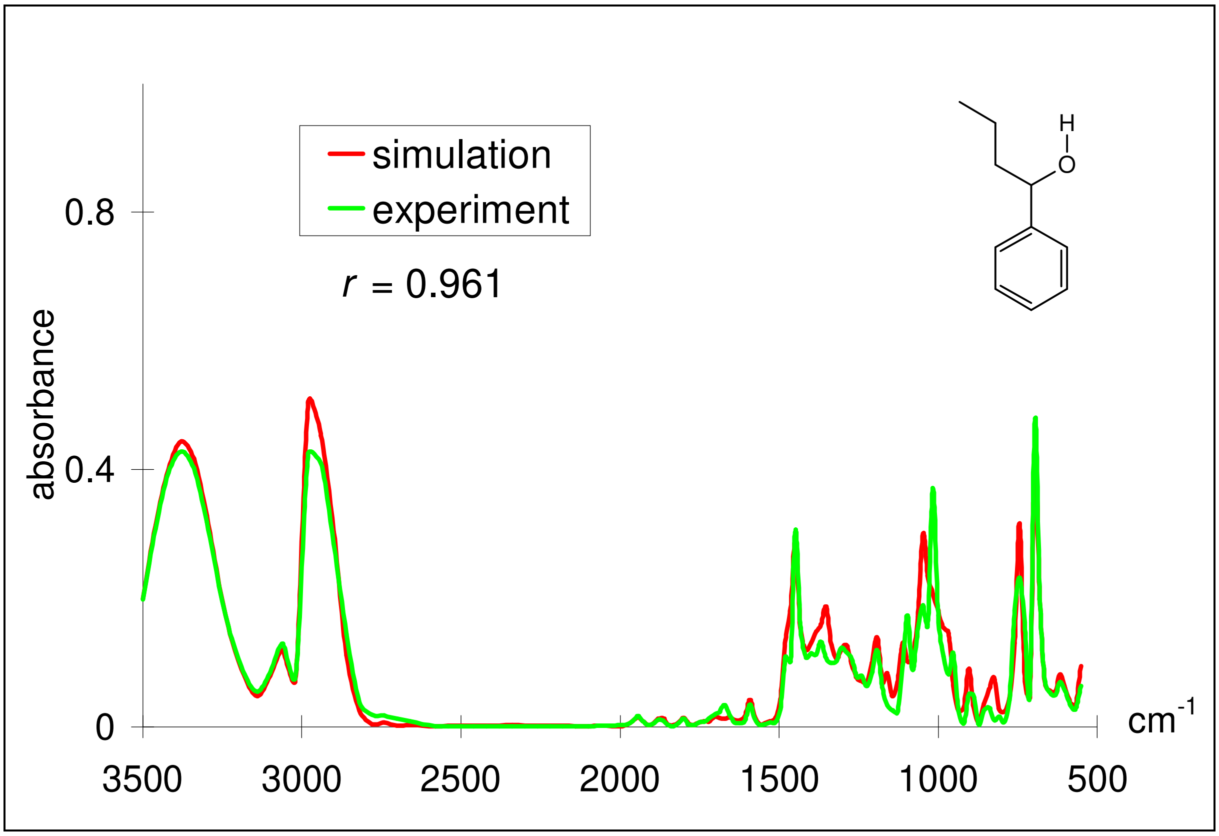


Abbildung 39: Eine der 25 besten Simulationen aus dem Testdatensatz für einen sekundären Alkohol, als erstes Beispiel für ein Molekül, daß unter den Meßbedingungen Wasserstoffbrückenbindungen ausbildet.

Neben der OH-Gruppe bilden Amine vielfach Wasserstoffbrücken aus, die häufig im IR-Spektrum sichtbar sind. Unter den 25 besten Simulationen des Testdatensatzes für die Infrarotspektrensimulation von substituierten Benzolen findet sich folgendes Beispiel für ein sekundäres Amin. N-Methylbenzylamin zeigt die breite Bande der N-H-Streckschwingung bei 3300 cm^{-1}. Dies stimmt mit dem von Hesse et al. angegeben Bereich für Flüssig- und Festphasenspektren überein.[45] Bei der Simulation des Spektrums konnten alle Banden korrekt simuliert werden, die einzigen Abweichungen sind eine Grundliniendrift oberhalb von 2500 cm^{-1} und eine etwas andere Form der CH-Bande bei 3000 cm^{-1}. Die andere Form der CH-Bande erklärt sich aus der Assoziation von N-Ethylbenzylamin aus dem Trainingsdatensatz mit Neuron (11,15), das für die Simulation des IR-Spektrums von N-Methylbenzylamin verwendet wurde.

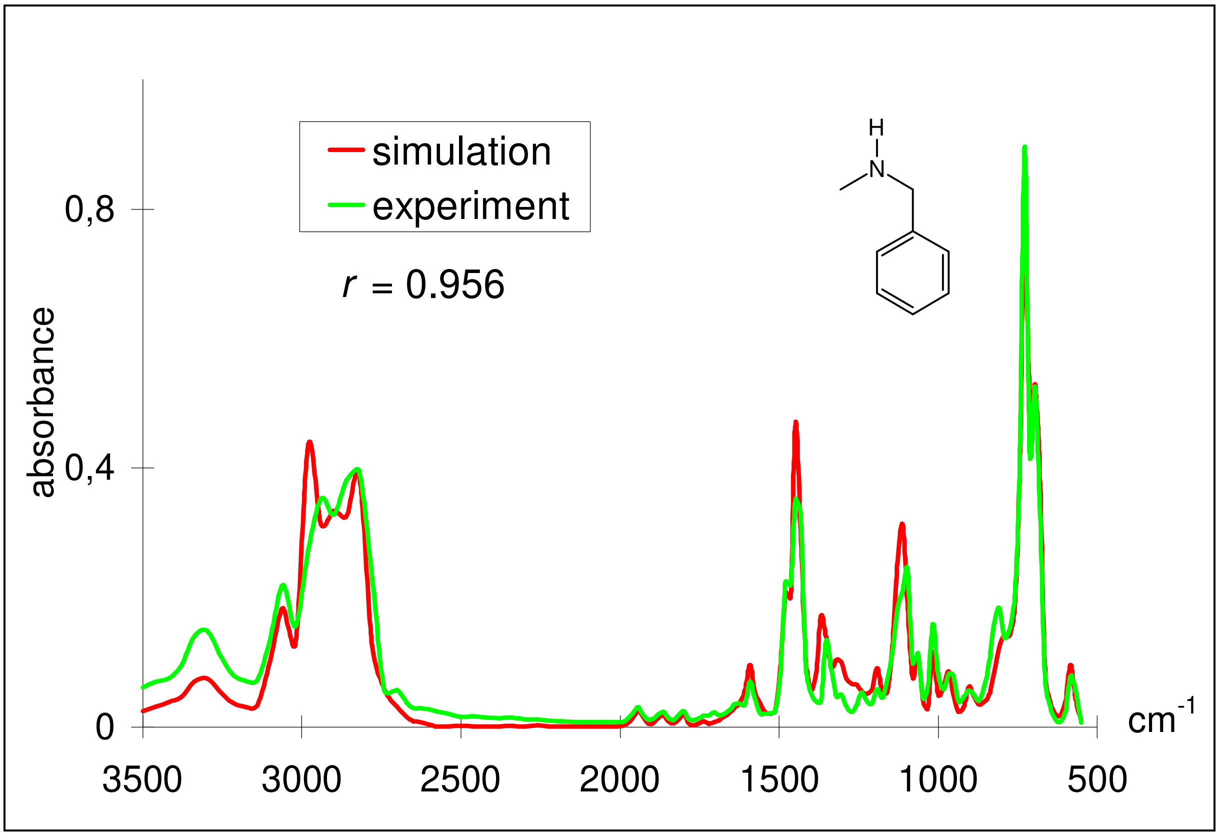


Abbildung 40: Simuliertes und experimentelles Infrarotspektrum von N-Methylbenzylamin als Beispiel für ein sekundäres Amin mit Wasserstoffbrükkenbindungen.

Neben dem obigen Beispiel für ein sekundäres Amin findet sich folgendes Beispiel für ein primäres Amin auf Platz fünf unter den besten Simulationen des Testdatensatzes. 4-(1,1-Dimethylethyl)-benzolamin vermutlich besser bekannt als *para*-tert. Butylanilin zeigt im Infrarotspektrum nicht die zwei typischen Banden der symmetrischen und der asymmetrischen NH-Streckschwingung, sondern eine dreifach aufgespaltene Bande mit Absorptionsspitzen bei 3460, 3380 und 3220 cm^{-1}. Die Spitzen bei 3460 und 3380 cm^{-1} sind typisch für Anilin-Derivate ohne Wasserstoffbrückenbindungen, während die dritte Absorption bei 3220 cm^{-1} typisch für durch Wasserstoffbrücken verbundene Amine ist. Im Fall von *para*-tert. Butylanilin liegen die Moleküle sowohl mit Wasserstoffbrücken als auch ohne vor, was nicht untypisch für Amine ist und, da das CPG-Netz von Beispielen lernt, auch korrekt simuliert wurde.

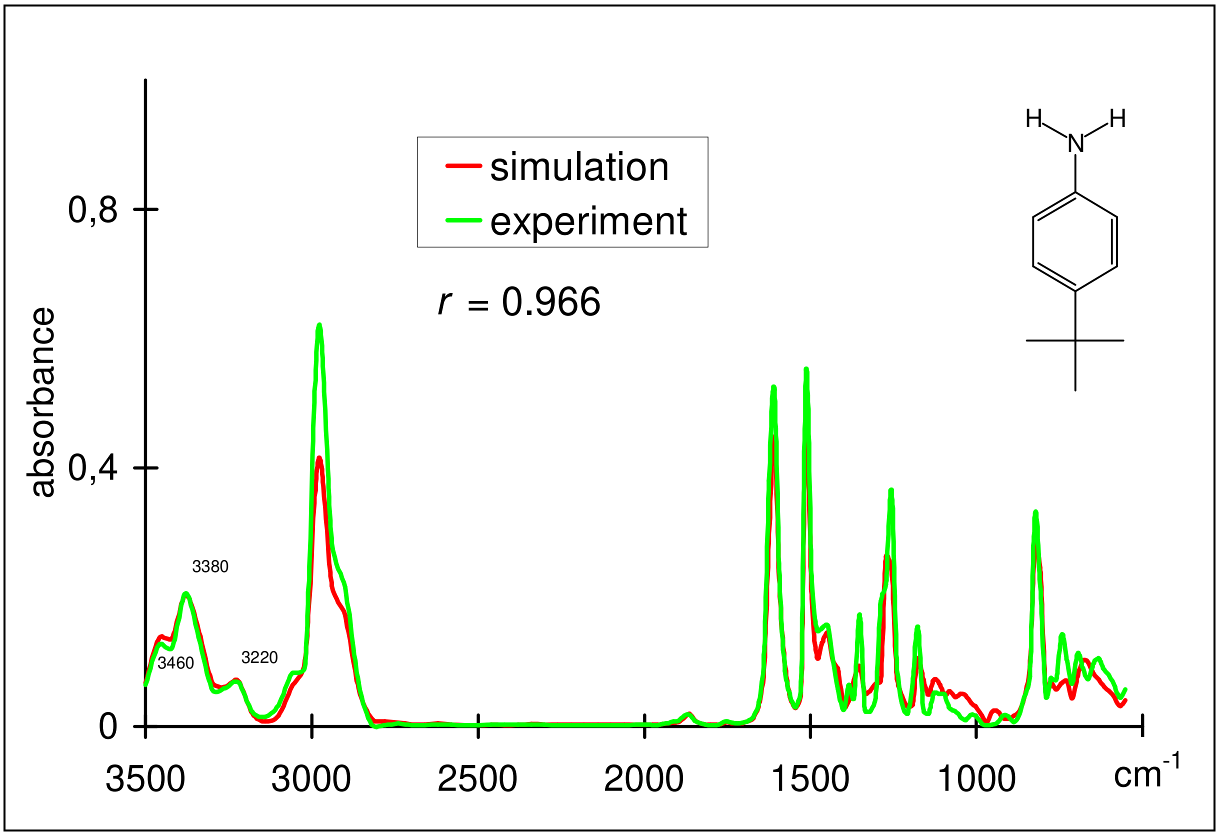


Abbildung 41: Simuliertes und experimentelles IR-Spektrum von para-^{tert}Butyl-Anilin als Beispiel für ein primäres Amin unter den 25 besten Simulationen des Testdatensatzes. Auffällig ist die dreifache Aufspaltung der NH-Bande, die auf das Vorliegen von durch Wasserstoffbrücken verbundenen Molekülen neben freien Molekülen hinweist.

Am deutlichsten zeigen sich die Auswirkungen von Wasserstoffbrückenbindungen im Spektrum von Carbonsäuren, wo die OH-Bande meist sehr stark verbreitert ist. Wie bei allen anderen Beispielen für Wasserstoffbrückenbindungen kann diese Verbreiterung einer Bande nicht berechnet werden und das Lernen von experimentellen Ergebnissen bleibt hier die einzige Methode zur Vorhersage. Das folgende Beispiel einer Carbonsäure mit Wasserstoffbrückenbindungen findet sich unter den 25 besten Simulationen des Trainingsdatensatzes. *para*-Chlorbenzoesäure weist im IR-Spektrum die typische breite Absorption der OH-Gruppe von Carbonsäuren über den Bereich von ca. 3200 bis 2200 cm^{-1} auf. Das simulierte Spektrum zeichnet diese breite Bande mit geringen Intensitätsabweichungen nach und ist im Fingerprintbereich nahezu identisch mit dem experimentellen Spektrum, bis auf eine Bande mittlerer Intensität bei 1080 cm^{-1}, die im simulierten Spektrum fehlt. Die Absorptionsbande des *para*-Chlorsubstituenten bei 744 cm^{-1} ist klar erkennbar. Insgesamt eine gelungene Simulation, die durch die Anwesenheit von *para*-Brombenzoesäure im Trainingsdatensatz möglich wurde.

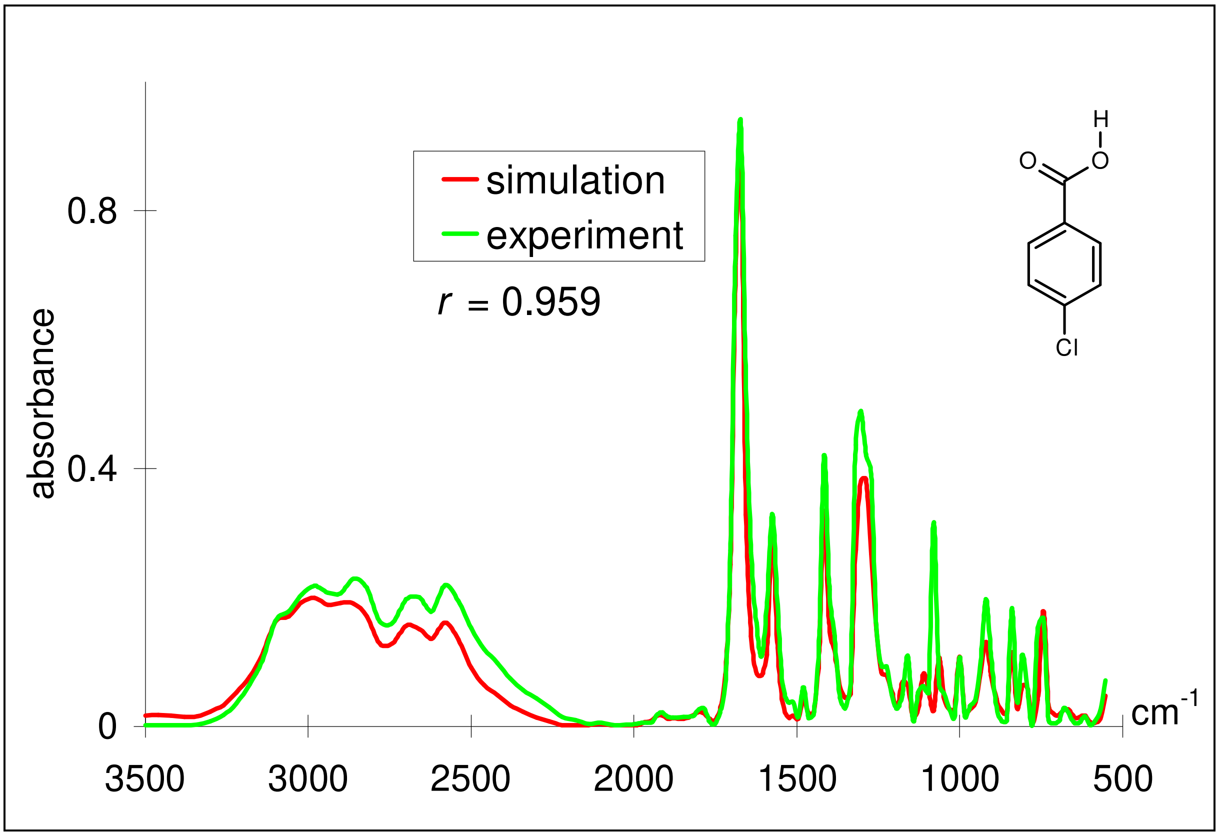


Abbildung 42: Experimentelles und simuliertes Infrarotspektrum von *para*-Chlorbenzoesäure als Beispiel für ein korrekt simuliertes IR-Spektrum einer typischen Carbonsäure aus den 25 besten Simulationen des Testdatensatzes für die Simulation von mono- bis trisubstituierten Benzolderivaten.

Die Infrarotspektren von Aldehyden waren lange Zeit mit quantenmechanischen Methoden nur schlecht zu berechnen, da Kopplungsphänomene zwischen der Schwingung der CO-Doppelbindung und der CH-Streckschwingung des Aldehydprotons nicht korrekt behandelt werden konnten. Erst DFT-Methoden konnten in den letzten Jahren die Probleme lösen. Der hier verwendete Ansatz umgeht auch diese Problematik vollständig, da die korrekte Behandlung der Kopplungsphänome implizit aus den Beispielen des Trainingsdatensatzes gelernt wird. Es ist darum nicht verwunderlich, wenn sich unter den 25 besten Simulationen des Testdatensatzes auch vier Simulationen für Aldehyde finden. Die Simulation von 3-Phenylpropanal gehört dazu. Das simulierte Spektrum gibt das experimentelle Spektrum, bis auf zwei zusätzliche intensitätsschwache Peaks bei 1112 und 795 cm^{-1}, wieder. Die wesentlichen Merkmale des experimentellen Spektrums wie CH-Bande, Carbonylbande und die CH-Deformationsschwingungen des Benzolrings bei 728 und 696 cm^{-1}, wie sie typisch für monosubstituierte Benzole sind, werden korrekt simuliert.[45]

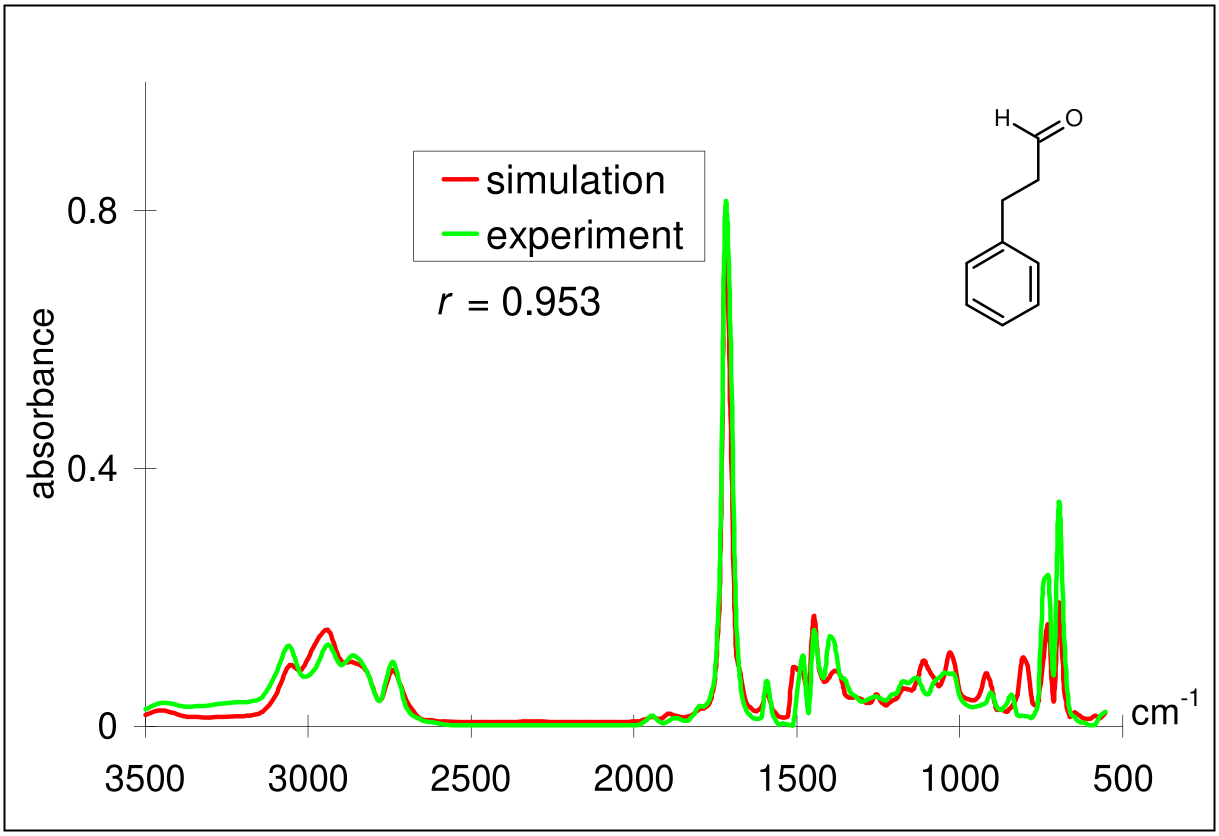


Abbildung 43: Experimentelles und simuliertes IR-Spektrum von 3-Phenylpropanal mit einem Korrelationskoeffizient von 0.953 zwischen experimentellem und simuliertem Spektrum.

Der Aufwand für quantenmechanische Berechnungen von IR-Spektren steigt mit der Anzahl der Atome im Molekül je nach Methode quadratisch, kubisch oder gar noch stärker an. Im Gegensatz dazu ist der Rechenaufwand für die hier verwendete Methode der IR-Spektrensimulation mittels 3D-Strukturcode und CPG-Netz nahezu unabhängig von der Molekülgröße, da die Molekülgröße nur in den schnellen Schritt der Codegenerierung eingeht, aber nicht in den langsameren Schritt der IR-Spektrensimulation mit Hilfe des neuronalen Netzes. Das größte Molekül des Testdatensatzes ist das Diphenylmethanderivat Diethoxy-di-(4-chlorphenyl)-methan. Für dieses Molekül wurde ein Korrelationskoeffizient von 0.932 zwischen experimentellem und simuliertem IR-Spektrum erreicht. Damit steht diese Simulation an siebzehnter Stelle unter den 25 besten Simulationen des Testdatensatzes. Abweichungen vom experimentellen Spektrum sind nur im Bereich der CH-Schwingungen knapp unterhalb von 3000 cm^{-1} und bei 1000 cm^{-1} zu erkennen. Die Ursache hierfür ist klar: die Simulation beruht auf dem Dimethoxyderivat im Trainingsdatensatz.

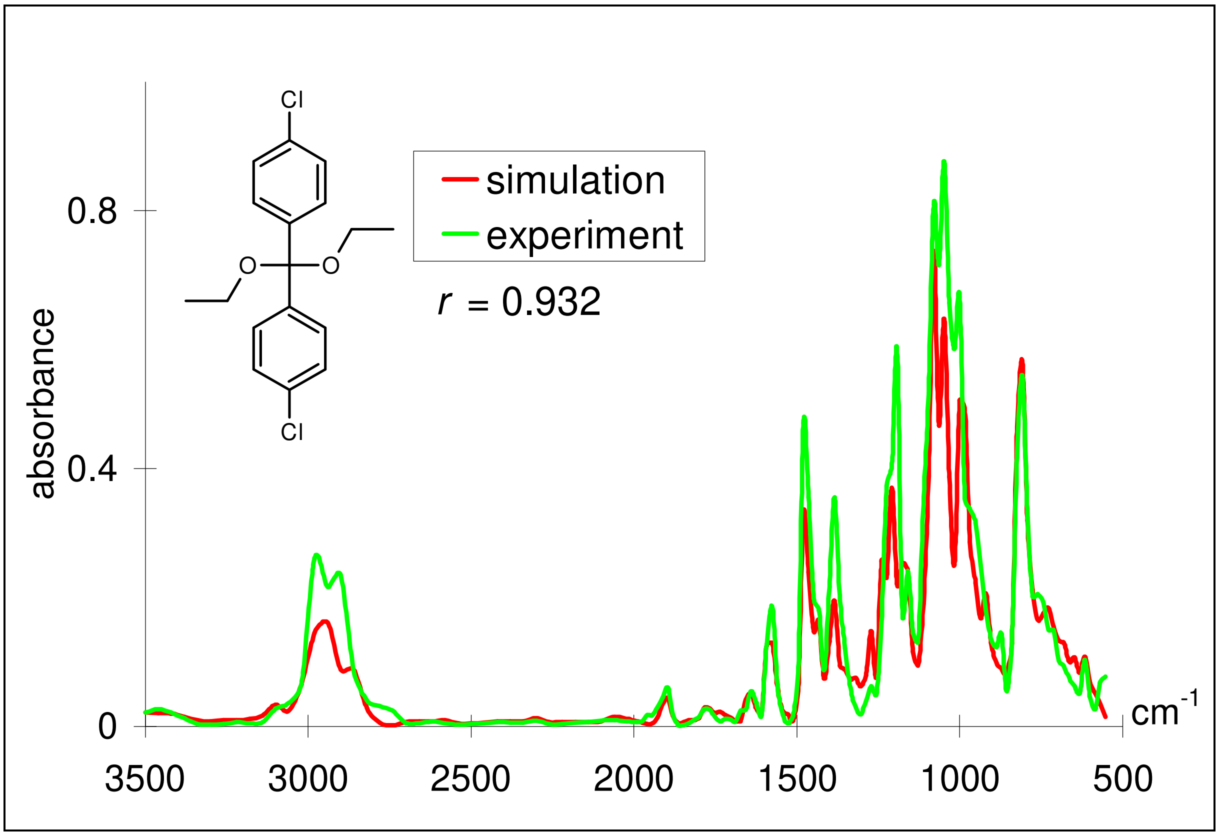


Abbildung 44: Simuliertes und experimentelles Infrarotspektrum von Diethoxydi-(4-Chlor-phenyl)-methan, dem größten Molekül im Testdatensatz. Die Abweichungen gehen auf die Verwendung des Dimethoxyderivates als Trainingsbeispiel zurück.

Atome mit hohen Ordnungszahlen, wie z.B. Brom, führen bei *ab initio* Berechnungen immer zu einem erhöhten Rechenaufwand. Die im Rahmen dieser Arbeit entwickelte Methode zur Simulation von IR-Spektren ist dagegen, im Hinblick auf die Rechenzeit, unabhängig von der Ordnungszahl der Atome im Molekül. Das folgende Beispiel eines Bromderivates soll verdeutlichen, daß diese Unabhängigkeit von der Ordnungszahl nicht mit einem Verlust an Simulationsqualität einhergeht. Für 4-(Brommethyl)-benzoesäuremethylester wurde ein Korrelationskoeffzient von 0.965 zwischen experimentellem und simuliertem IR-Spektrum erreicht. Der optische Vergleich der Spektren zeigt die typische, geringere Intensität des simulierten Spektrums sowie das Fehlen zweier schwacher Absorptionsbanden zwischen 1000 und 800 cm^{-1}.

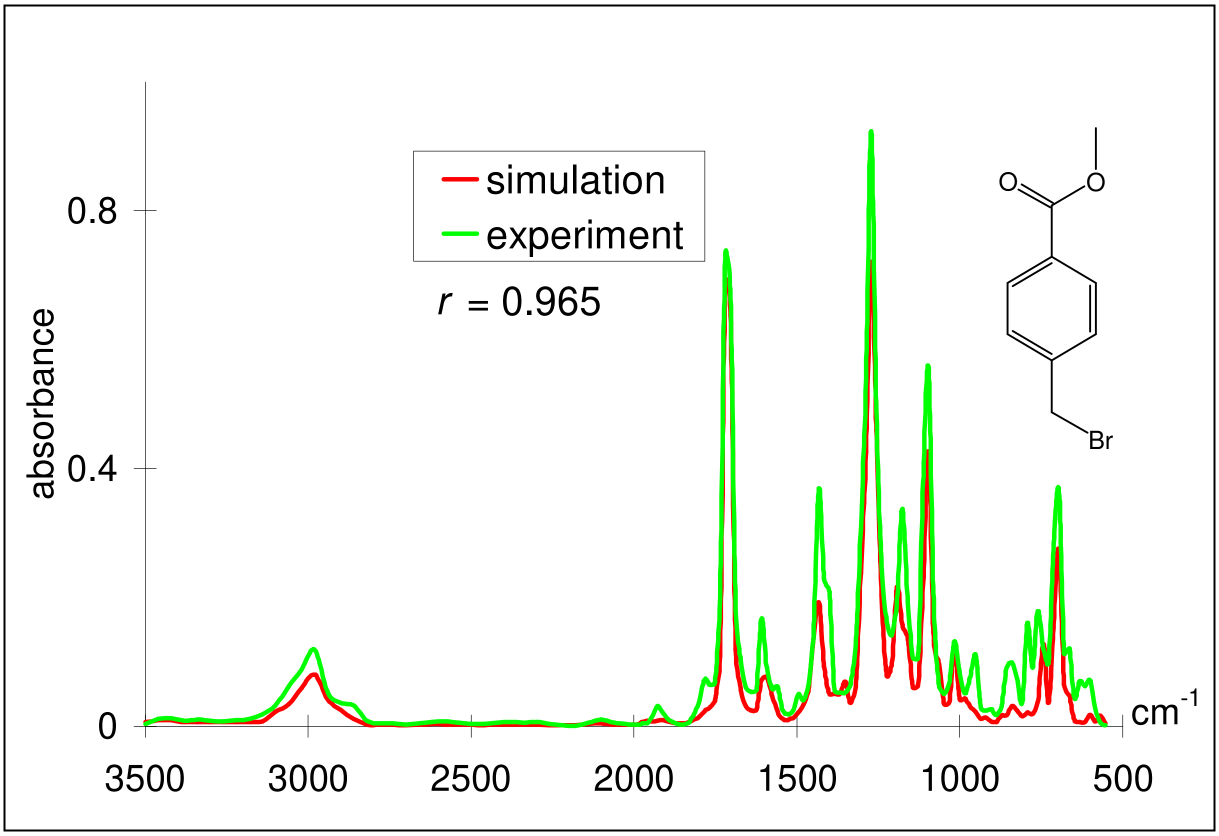


Abbildung 45: Experimentelles und simuliertes Spektrum von 4-(Brommethyl)-benzoesäuremethylester; mit einem Korrelationskoeffizient von 0.965 steht diese Simulation an siebter Stelle unter den 25 besten Simulationen.

6.6.4.5 Beispiele für Simulationen mit niedriegeren Korrelationskoeffizienten

Um die Qualität der Simulationen mit einem niedrigeren Korrelationskoeffizienten zu illustrieren, folgen Beispiele für Simulationen mit Korrelationskoeffizienten von ca. 0.9, 0.8, 0.7 usw.. Anhand dieser Beispiele soll das Verhältnis von Korrelationskoeffizient zur strukturellen Abweichung zwischen der Testverbindung und der oder den Trainingsverbindungen, deren Infrarotspektren zur Simulation genutzt wurden, diskutiert werden.

Die Simulation von 2-Phenylpentannitril (**N1**) mit einem Korrelationskoeffizient von 0.903 dient als Beispiel für Simulationen mit einem Korrelationskoeffizienten von ca. 0.9. Wie Abbildung 46 zeigt, unterscheiden sich simuliertes und experimentelles Spektrum im wesentlichen nur durch Intensitätsdifferenzen und unterschiedliche Bandenformen: zum einem bei den CH-Streckschwingungen um 3000 cm^{-1}, zum anderen bei einem Bandenkomplex zwischen 1500 und 1400 cm^{-1}. Die Ursache für die Abweichungen im simulierten Spektrum, das Neuron (4,15) des Counterpropagationen-Netzes entnommen wurde, liegt im Training des Netzes. Dem Neuron (4,15) wurde im Erinnerungstest das Isomere 5-Phenylpentannitril (**N2**) assoziiert und das in Neuron (4,15) gespeicherte Spektrum gleicht im wesentlichen dem Spektrum dieser Verbindung.

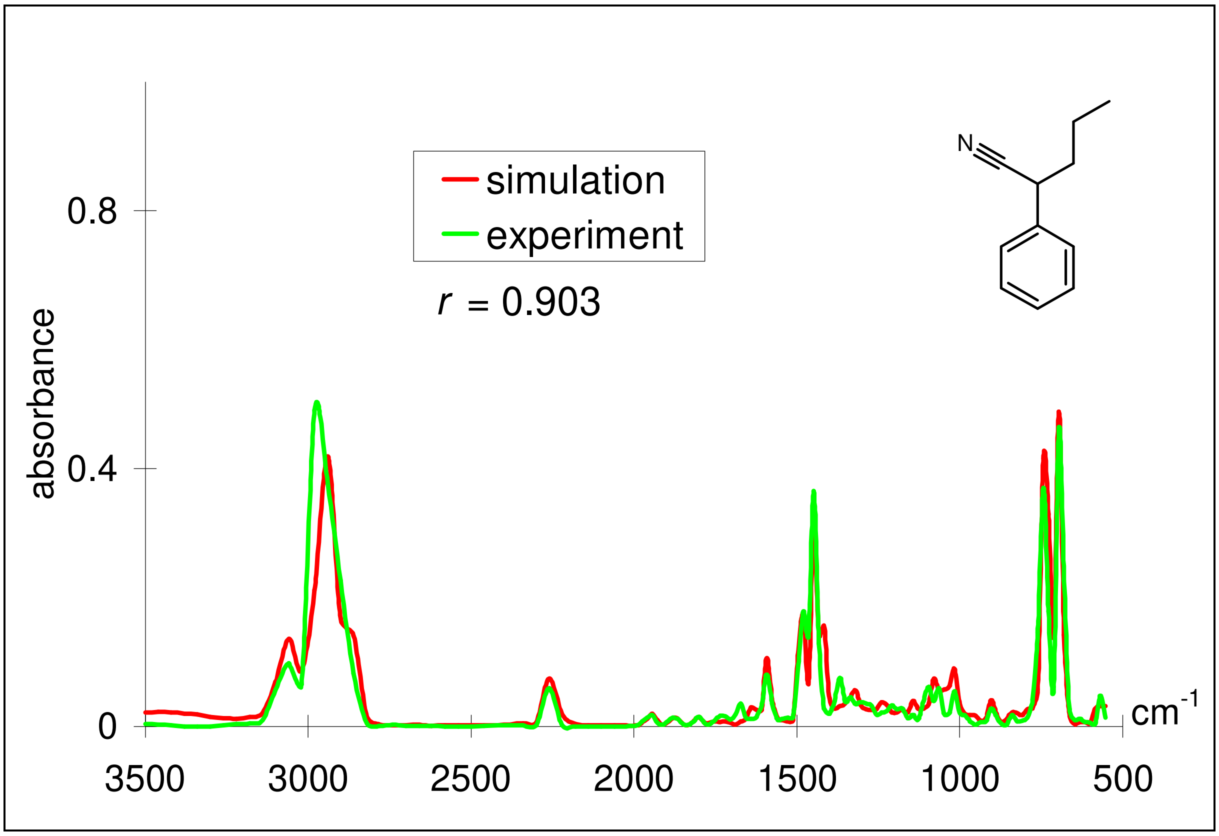
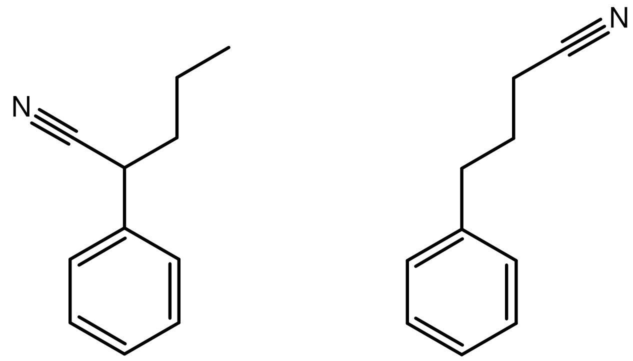


Abbildung 46: Experimentelles und simuliertes Infrarotspektrum von 2-Phenylpentannitril als Beispiel für eine Simulation mit einem Korrelationskoeffizienten von ca. 0.9.

Das Isomerenpaar von **N1** und **N2** (Schema 2) aus Trainings- und Testdatensatz weist die Stellungsisomerie einer funktionellen Gruppe in der Seitenkette der Benzolderivate als eine mögliche Differenz zwischen Molekülstrukturen aus, die für einen Korrelationskoeffizienten von 0.9 verantwortlich ist.



N1 **N2**

Schema 2: Isomerenpaar 2- (**N1**) und 5-Phenylpentannitril (**N2**)

Als Beispiel für einen Korrelationskoeffizienten von etwa 0.8 wurde die Simulation von (3,4-Dichlorphenyl)-methanol ausgewählt. Experimentelles und simuliertes Spektrum unterscheiden sich hier schon deutlicher und ein bis zwei der mittleren Banden im Spektrum fehlen oder erscheinen zusätzlich. Bei der Simulation von (3,4-Dichlorphenyl)-methanol, **CM1**, fallen vor allem die gegenüber dem experimentellem Spektrum viel stärkere Bande bei 1560 cm^{-1} und die Verschiebung der Banden bei 1120 und 872 cm^{-1} zu etwas niedrigeren Wellenzahlen auf. Daneben gibt es Intensitätsdifferenzen in allen Bereichen des Spektrums. Das simulierte Infrarotspektrum stammt vom Neuron (17,18) des Counterpropagation Netzes. Im Training wurde das Neuron (17,18) im wesentlichen von (4-Chlorphenyl)-methanol, **CM2**, und (3,5-Dichlorphenyl)-methanol, **CM3**, beeinflußt, die beide im Erinnerungstest diesem Neuron assoziiert wurden. Das mit Hilfe von Neuron (17,18) erhaltene Simulationsspektrum stellt somit ein Mischspektrum aus den beiden Trainingsmolekülen dar.

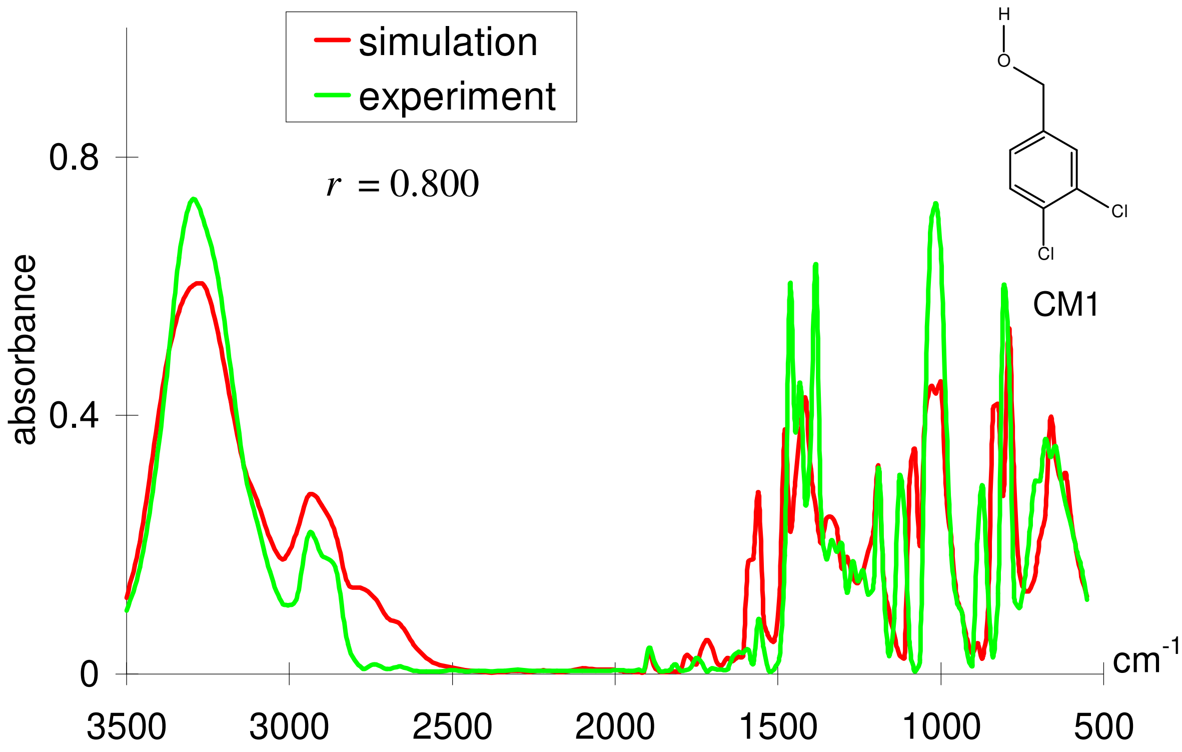
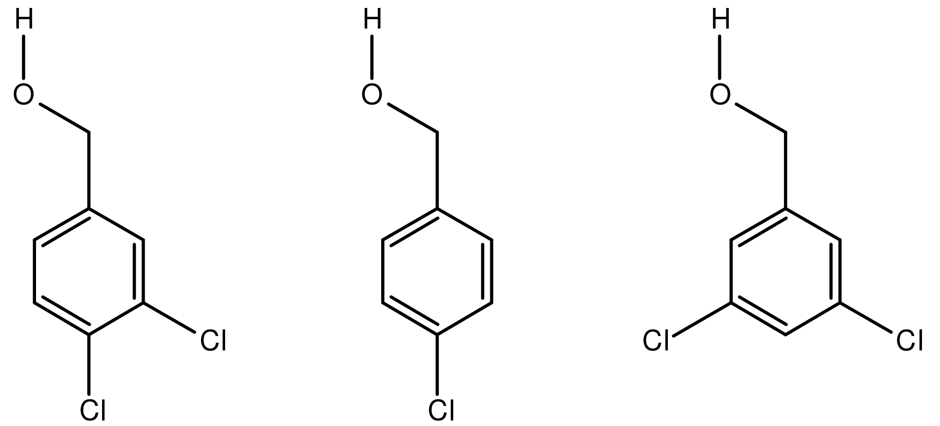


Abbildung 47: Simuliertes und experimentelles Spektrum von **CM1** als Beispiel für eine Simulation mit einem Korrelationskoeffizienten von 0.8.

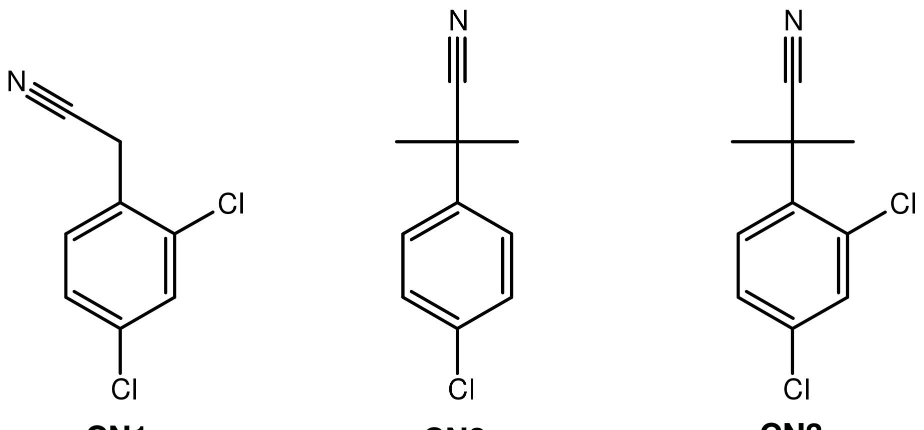


CM1 **CM2** **CM3**

Schema 3: (3,4-Dichlorphenyl)-methanol **CM1**, (4-Chlorphenyl)-methanol **CM2**, (3,5-Dichlorphenyl)-methanol **CM3**

Wie das Beispiel von (3,4-Dichlorphenyl)-methanol zeigt, kann der Grund für einen Korrelationskoeffizienten von 0.8 in der Stellungsisomerie bzw. dem Fehlen eines Chlorsubstituenten liegen. Wie stark der Einfluß von Stellungsisomerie bzw. der Ab- oder Anwesenheit eines zusätzlichen Substituenten auf das Infrarotspektrum ist, hängt von verschiedenen Faktoren ab - an erster Stelle von der Art des Substituenten. Hier fallen Substituenten, die ein zweites Mal im Molekül vorkommen, deutlich weniger ins Gewicht als gänzlich neue. Änderungen bei Methylgruppen und Halogenatomen sowie insbesondere bei der Anzahl von CH_2-Einheiten in längeren Alkylketten, haben im allgemeinen eine geringe Auswirkung auf das Infrarotspektrum, im Gegensatz zu polaren funktionellen Gruppen mit markanten Valenzschwingungen, wie z.B. der Carbonylgruppe. Dies wird sich bei den folgenden Beispielen deutlich zeigen.

Das folgende Beispiel für einen Korrelationskoeffizienten um 0.7, die Simulation von (2,4-Dichlorphenyl)-acetonitril, **CN1**, zeigt neben vielen Gemeinsamkeiten im Spektrum schon deutliche Abweichungen. Neben den Intensitätsunterschieden von wechselnder Stärke im gesamten Spektralbereich, fällt vor allem die zusätzliche mittelstarke Bandengruppe bei 1200 cm^{-1} und der veränderte Spektrenverlauf unterhalb von 700 cm^{-1} negativ auf. Es ist typisch für einen Korrelationskoeffizienten von ca. 0.7, daß im simulierten Spektrum entweder eine starke oder zwei mittlere Banden zusätzlich auftreten oder fehlen. Das simulierte Spektrum beruht auf den Spektren von 2-(2,4-Dichlorphenyl)-2-methylpropionitril und 2-(4-Dichlorphenyl)-2-methyl-propionitril aus dem Trainingsdatensatz, wobei das im Neuron gespeicherte Spektrum dem Spektrum der erstgenannten Verbindung wesentlich ähnlicher ist.



CN1 **CN2** **CN3**

Schema 4: Das Testmolekül (2,4-Dichlorphenyl)-acetonitril, **CN1**, und die bei Moleküle aus dem Trainingsdatensatz und 2-(4-Dichlorphenyl)-2-methyl-propionitril, **CN2**, und 2-(2,4-Dichlorphenyl)-2-methylpropionitril, **CN3**.

Der wesentliche strukturelle Unterschied, der in dem Korrelationskoeffizienten von 0.7 zum Ausdruck kommt, sind die zwei CH_x-Gruppen in der Nähe der polarsten funktionellen Gruppe des Moleküls, der Cyanogruppe.

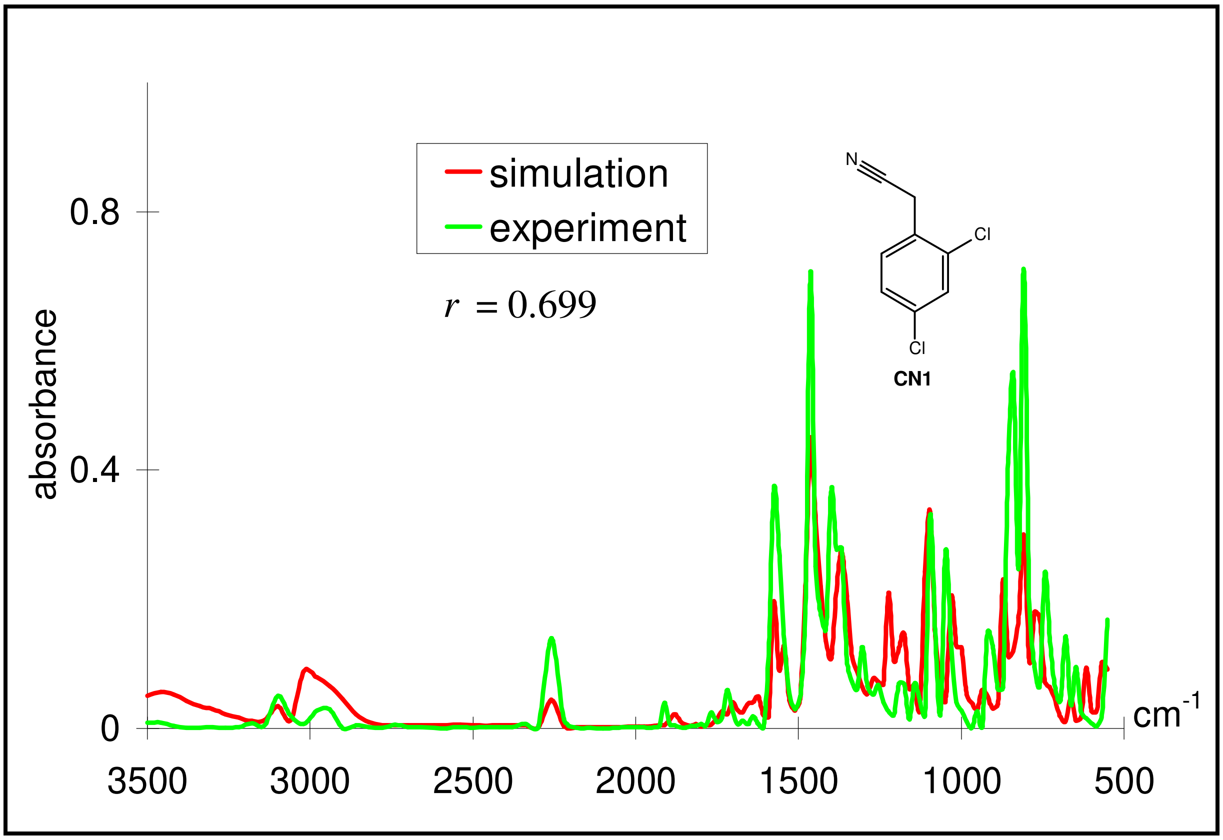


Abbildung 48: Simuliertes und experimentelles Infrarotspektrum von **CN1** als Beispiel für eine Simulation mit einem Korrelationskoeffizienten von 0.7.

Das folgende Beispiel für einen Korrelationskoeffizienten von 0.6 ist bereits an der Grenze für verwendbare IR-Spektrensimulation im Sinne einer Strukturaufklärung, da unter 0.6 bereits große Unterschiede in den zugrundeliegenden Molekülstrukturen möglich sind. Gewählt wurde als Beispiel die Simulation von 3,5-Dichlorbenzylamin. Experimentelles und simuliertes Spektrum unterscheiden sich in den Bereichen des Benzolkerns stark, während im Bereich der NH-Streckschwingungen gute Übereinstimmung besteht. Dies gibt einen ersten Hinweis darauf, daß sich Test- und Trainingsmolekül(e) im Bereich des Benzolkerns unterscheiden, nicht aber im Bereich der funktionellen Gruppe (Aminogruppe). Das für die Simulation genutzte Neuron (13,15) wurde im Lauf des Trainings wesentlich durch 4-Chlorbenzylamin geprägt. Damit unterscheiden sich in diesem Beispiel die Moleküle aus Trainings- und Testdatensatz durch die Position und Zahl der Chloratome am Benzolring. Dies ist nicht untypisch für Simulationen mit einem Korrelationskoeffizienten um 0.6.

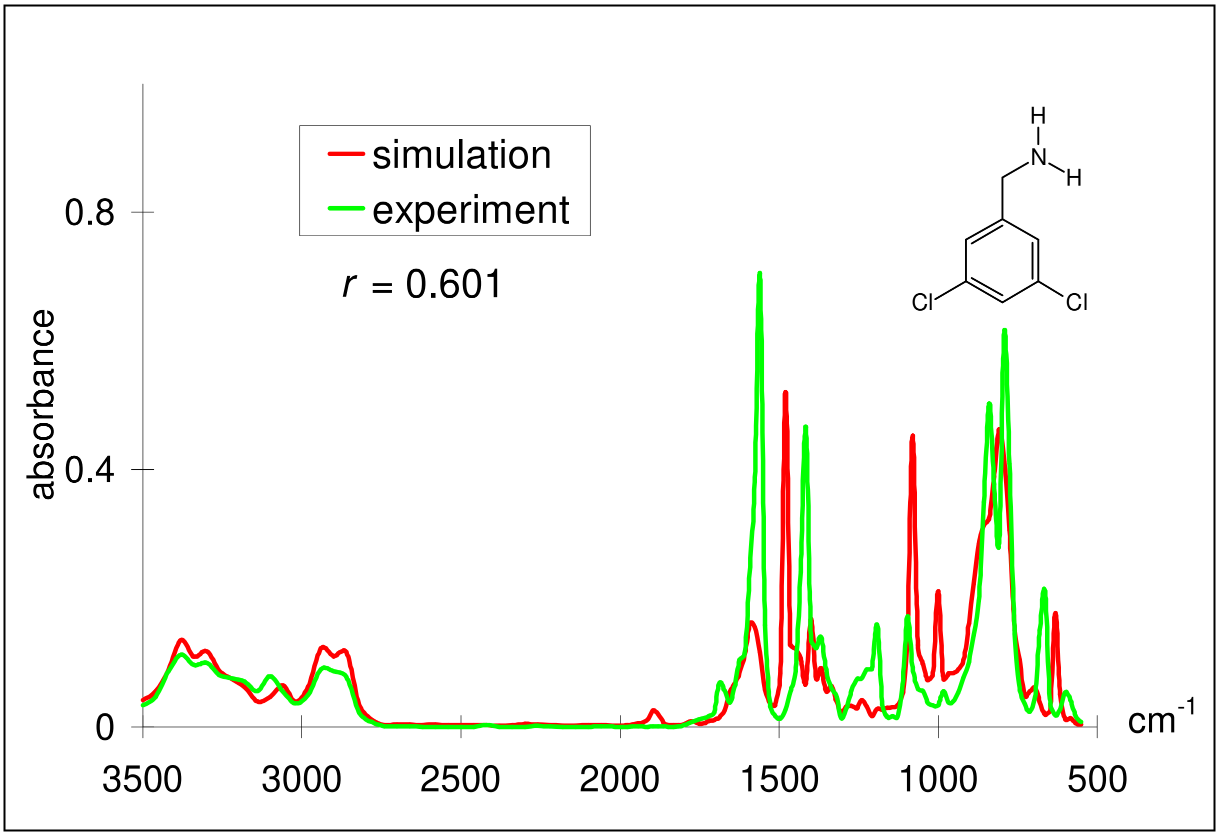
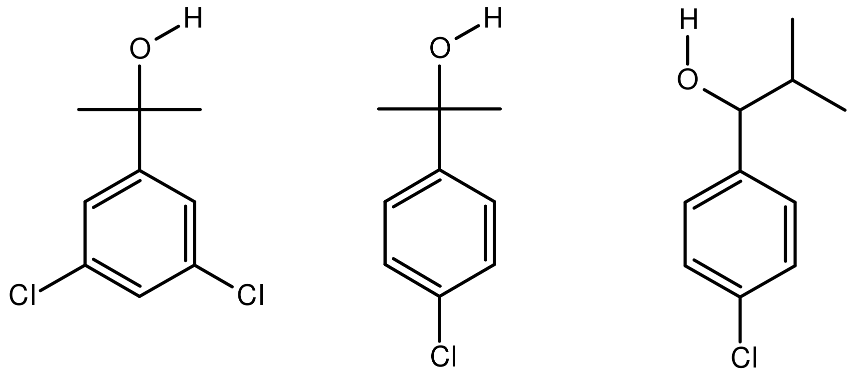


Abbildung 49: Experimentelles und simuliertes Infrarotspektrum von 3,5-Dichlorbenzylamin als Beispiel für einen Korrelationskoeffizienten von 0.6.

Die Simulation von 2-(3,5-dichlorphenyl)-propan-2-ol, **CO1**, dient als Beispiel für einen Korrelationskoeffizienten von 0.5. Die Unterschiede zwischen experimentellem und simuliertem Spektrum sind in bezug auf Intensität und Bandenlage bereits deutlich sichtbar. Intensitätsunterschiede gibt es vor allem bei der CH-Streckschwingung um 3000 cm^{-1} sowie den Banden des Benzolrings zwischen 1600 und 1500 cm^{-1}. Bei den Bandenlagen finden sich Abweichungen durch eine zusätzliche Bande bei 1000 cm^{-1} sowie eine fehlende Bande bei 650 cm^{-1}. Der Grund für diese Abweichungen liegt in den strukturellen Unterschieden von **CO1** zu den beiden Trainigsdatensatzmolekülen, 2-(4-Chlorphenyl)-propan-2-ol, **CO2**, und 1-(4-Fluorphenyl)-2-methylpropanol, **CO3**, die zu annähernd gleichen Teilen das Simulationsspektrum prägten.



CO1 **CO2** **CO3**

Schema 5: Die Testverbindung 2-(3,5-Dichlorphenyl)-propan-2-ol, **CO1**, von Neuron (11,24) und die Trainingsverbindungen (**CO2** und **CO3**) die diesem Neuron assoziiert wurden.

Der strukturelle Unterschied zwischen Test- und Traingsdatensatzmolekülen, die hier einem Neuron assoziiert wurden, umfaßt die Stellungsisomerie und das Fehlen eines Halogenatoms bzw. einer Methylengruppe sowie den Austausch von Chlor durch Fluor. Glücklicherweise treten Abweichungen dieser Größenordnung nur bei weniger als 1/8 der Simulationen des Testdatensatzes auf und liegen oft im Fehlen ähnlicher Benzolderivate im Trainingsdatensatz begründet.

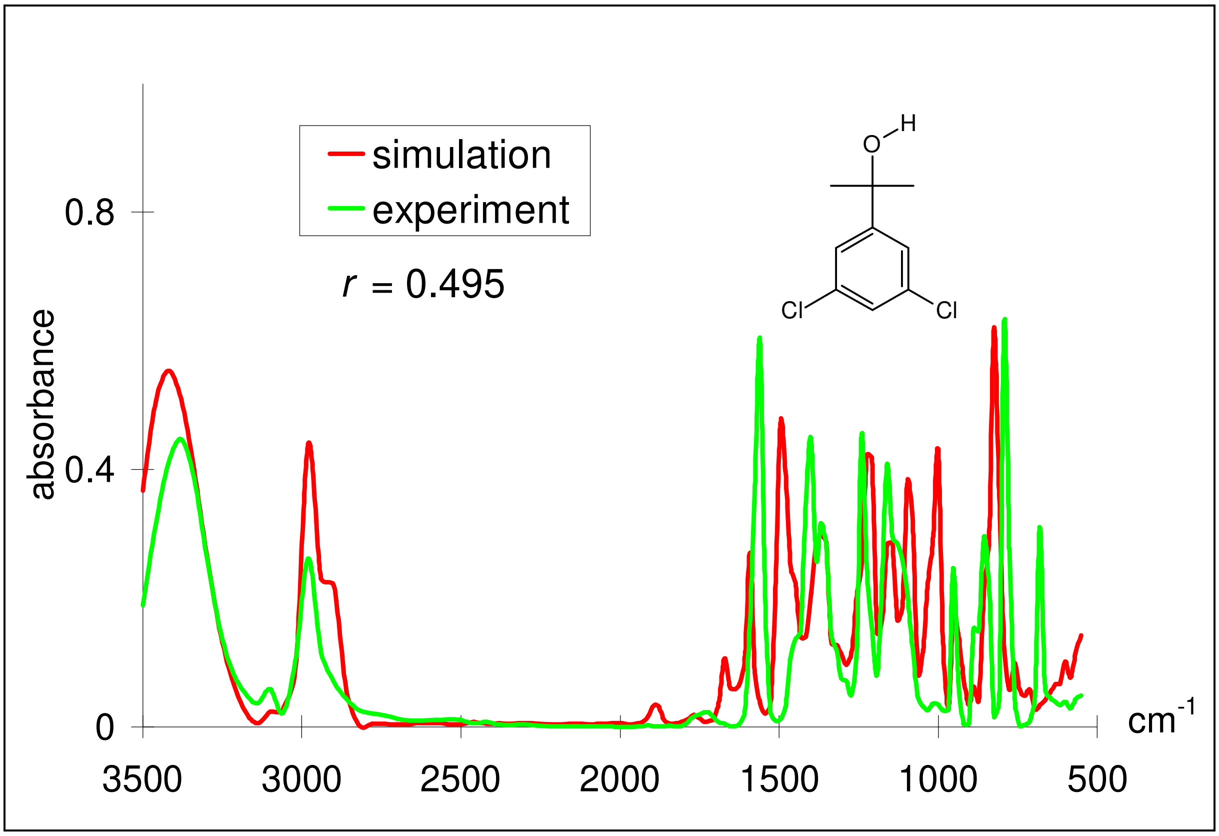
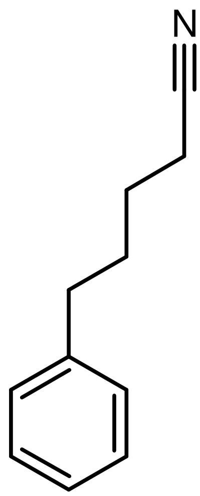


Abbildung 50: Experimentelles und simuliertes Spektrum von 1-Methyl-1-(3,5-dichlorphenyl)-ethanol als Beispiel für eine Simulation mit einem Korrelationskoeffizienten von 0.5.

Bei einem Korrelationskoeffizienten von deutlich unter 0.5 sind die Abweichungen zwischen simuliertem und experimentellem Infrarotspektrum so groß, daß eine Verwendung dieser Simulationen nur noch sehr eingeschränkt möglich sein wird. An den folgenden zwei Beispielen soll untersucht werden, was die Gründe für solcherart niedrige Korrelationskoeffizienten sein können.

Das erste dieser Beispiele ist die Simulation von 6-Chlor-2-(propanoxy-2-en)-benzonitril. Der Korrelationskoeffizient für diese Simulation beträgt 0.315. Interessanterweise hat dieses Testdatensatzmolekül viele Strukturmerkmale mit dem einzigen Trainingsdatensatzmolekül 5-Phenoxypentannitril gemeinsam, das in dasselbe Neuron projiziert wurde.



Schema 6:
5-

Die gemeinsamen Merkmale sind: Phenolethergruppe, Nitrilgruppe, und eine aliphatische Kohlenstoffkette. Allerdings sind diese unterschiedlich angeordnet und bei dem Molekül aus dem Trainingsdatensatz fehlt der Chlorsubstituent am Benzolring. Diese strukturellen Unterschiede sind für die Abweichungen im IR-Spektrum verantwortlich. Eine Verbesserung der Simulationsqualität setzt strukturell ähnlichere Verbindungen im Trainingsdatensatz als 5-Phenoxypentannitril voraus. Zudem legt die unterschiedliche Anordnung von Cyanogruppe und aliphatischer Kette in den beiden Strukturen den Verdacht nahe, daß im Code die aliphatische Kette nicht sehr gut repräsentiert wurde.

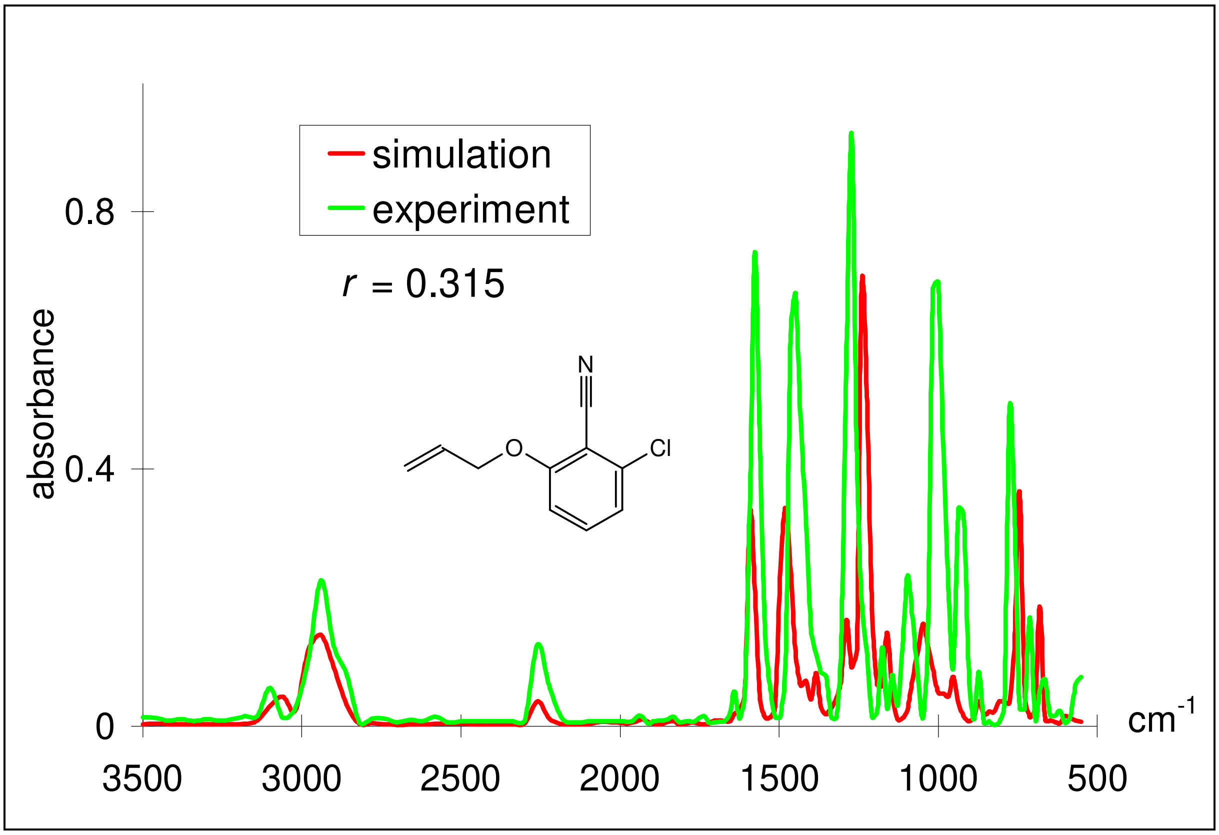
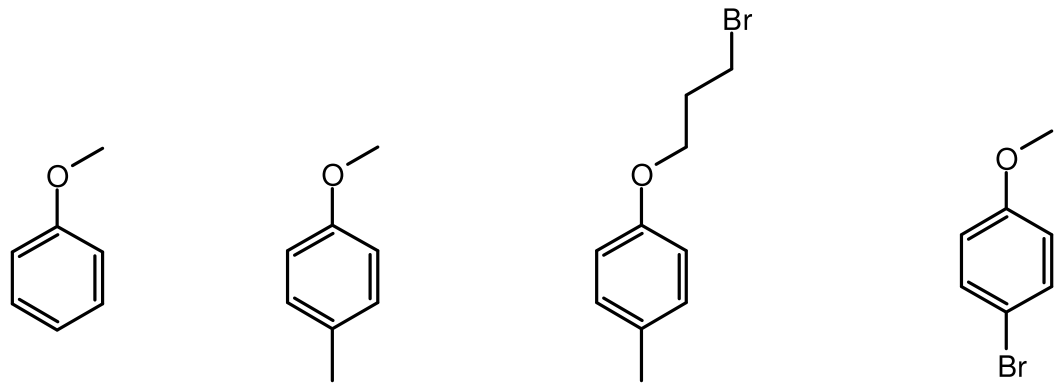


Abbildung 51: Experimentelles und simuliertes Spektrum von 6-Chlor-2-(propanoxy-2-en)-benzonitril.

Als letztes Beispiel soll die schlechteste Simulation des Testdatensatzes analysiert werden. Diese Stellung nimmt die Simulation von Anisol ein. Nicht zufällig handelt es sich um die Simulation eines einfachen Moleküls, da bei jedem einfachen Molekül, wie eben Anisol (Methoxybenzol) oder gar Benzol selber, jedes zusätzliche oder fehlendes Atom eine viel deutlichere Spur im Infrarotspektrum hinterläßt, als beispielsweise die Verlängerung einer aliphatischen Seitenkette von zwei auf drei CH_2-Einheiten.

Schema 1 zeigt neben Anisol aus dem Testdatensatz die drei Verbindungen des Trainingsdatensatzes, die in dasselbe Neuron wie Anisol projiziert wurden, zusammen mit den Korrelationskoeffizienten zwischen dem experimentellen und dem auf dem Neuron gespeicherten Infrarotspektrum für die Simulation. Wie die Korrelationskoeffizienten in Verbindung mit den Strukturen zeigen, ist der Unterschied, IR-spektroskopisch gesehen, zwischen Methodxybenzol , **mb1**, und 4-Methylmethoxybenzol, **mb2**, wesentlich größer als zwischen **mb2** und (3-Brompropanoxy)-4-methylbenzol, **mb3**. Dies obwohl **mb1** und **mb2** sich „nur“ in der zusätzlichen Methylgruppe am Benzolring unterscheiden, während bei **mb3** die aliphatische Seite des Phenolethers zwei zusätzliche CH_2-Einheiten und einen zusätzlichen Bromsubstituenten aufweist. Dies hat aber auf das Infrarotspektrum einen wesentlich geringeren Einfluß.



r=0.113	*r*=0.811	*r*=0.910	*r*=0.911
Methoxybenzol	4-Methylmethoxybenzol	(3-Brompropanoxy)-4-methylbenzol	4-Brommethoxybenzol
mb1	**mb2**	**mb3**	**mb4**
Testdatensatz	Trainingsdatensatz	Trainingsdatensatz	Trainingsdatensatz

Schema 7: Anisol und die drei Verbindungen aus dem Trainingsdatensatz, die in dasselbe Neuron wie Anisol projiziert wurden. Angegeben sind die Korrelationskoeffizienten zwischen experimentellen Spektren und dem auf dem Neuron gespeicherten Simulationsspektrum.

Das Beispiel von **mb1** zeigt klar die Grenzen dieser Methode zur Simulation von IR-Spektren. Die Zuordnung von **mb1** zu Neuron (3,21), dem u. a. **mb2** aus dem Trainingsdatensatz assoziiert wurde, kann im Sinne der Methode nicht als Fehler gewertet werden, führt aber dennoch zum schlechtesten Ergebnis des gesamten Versuchs, da jedes zusätzliche Atom das Spektrum von **mb1** stark verändert hätte. Mit anderen Worten, es gibt keine Verbindung, die im Sinne der Infrarotspektroskopie ähnlich zu **mb1** ist. Jeder zusätzliche Substituent am Benzolring, jede Veränderung an der aliphatischen Kette wird sich im Infrarotspektrum auswirken, da diese Änderungen in der Nähe der höchsten Ladung des Moleküls liegen. Dabei wirken sich die Effekte einer Änderung am Benzolring, da sie, aufgrund der Konjugation des aromatischen Systems, weit mehr Bindungen betreffen, stärker auf das Infrarotspektrum aus, als es Änderungen in der Seitenkette tun. So beträgt der Korrelationskoeffizient zwischen den experimentellen Spektren von **mb2** und **mb1** 0.048, bei vorhandenen aber unterschiedlichen *para*-Substituenten (zwischen **mb2** und **mb4**) 0.648 und bei identischen *para*-Substituenten aber unterschiedlicher aliphatischer Seitenkette (zwischen **mb2** und **mb3)** 0.778.

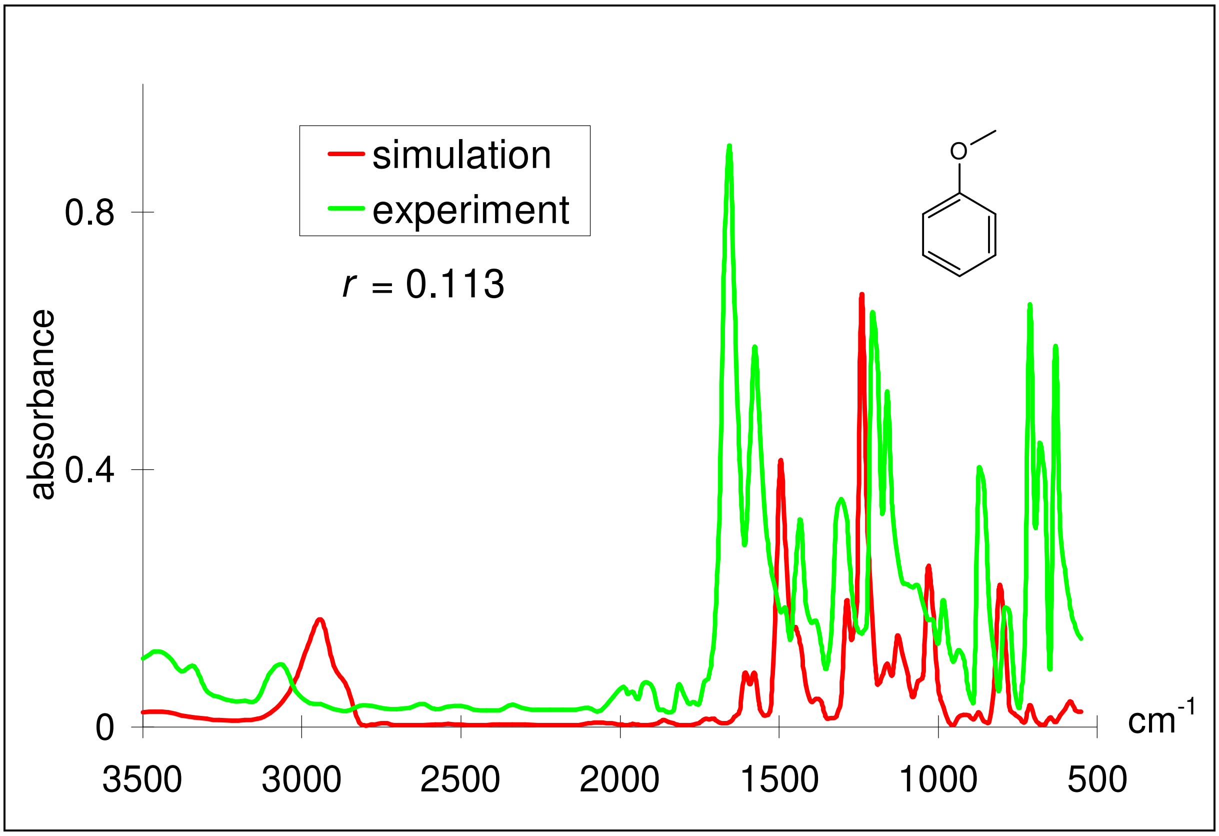


Abbildung 52: Anisol, die schlechteste Simulation unter allen Benzolderivaten. Grund - das Molekül ist zu einfach. Jedes Atom mehr oder weniger würde das Spektrum stark ändern.

6.6.5 *Spezialfälle - Verbindungen mit einem zweiten Ringsystem*

Um die Verteilung von Verbindungen im neuronalen Counterpropagation-Netz durch den kompetetiven Lernprozeß zu analysieren, wurden alle Verbindungen ausgewählt, die zusätzlich zum Benzolring einen zweiten fünf- oder sechsgliedrigen Ring aufwiesen. Insgesamt 17 Verbindungen, **R1** - **R17**, wurden so erhalten, die neben dem zweiten Ring eine Vielzahl weiterer struktureller Merkmale aufweisen. Abbildung 53 zeigt die Verteilung dieser Verbindungen im zweidimensionalen Netzwerk (Das CPG-Netz aus Abbildung 25 wird hierfür von oben betrachtet.). Knapp außerhalb der Netzmitte liegen die unpolaren Kohlenwasserstoffe und 1-Phenylpyrrole (**R12**-**R17**) auf den Neuronen (7,14), (8,14) und (8,15).

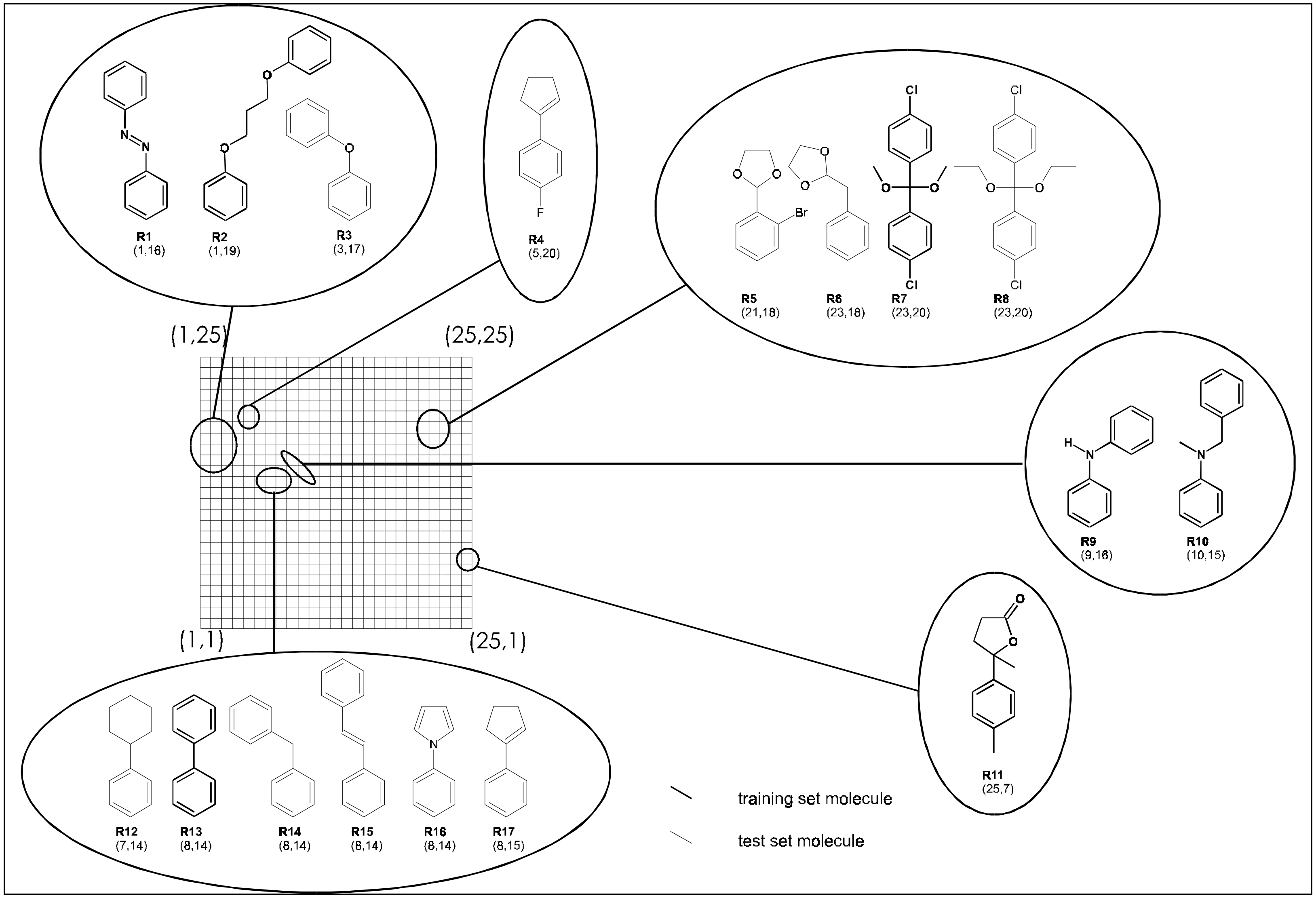


Abbildung 53: Verteilung der Benzolderivate mit einem zusätzlichen Ring aus fünf- oder sechs Atomen im CPG-Netz

Die Nutzung der partiellen Atomladung $q_{tot,i}$ bei der Codierung dieser Verbindungen mittels des 3D-MoRSE Codes betont die strukturelle Ähnlichkeit dieser Verbindungen, die sich auch in der Ähnlichkeit ihrer Infrarotspektren widerspiegelt. Dies zeigt Abbildung 54 mit dem gemittelten IR-Spektrum, das in Neuron (8,14) gespeichert wurde, sowie den experimentellen Spektren von Biphenyl und Diphenylmethan mit Korrelationskoeffizienten von 0.750 bzw. 0.744.

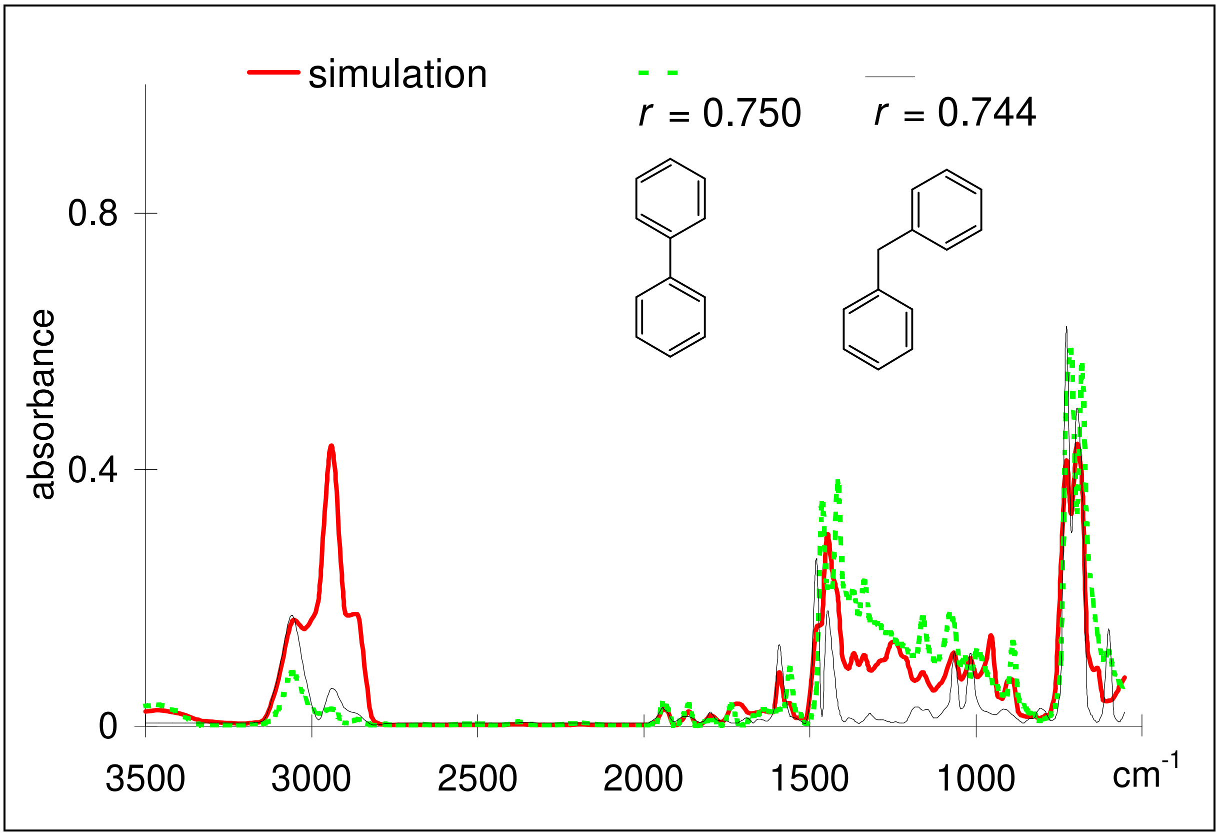


Abbildung 54: Experimentelle Infrarotspektren von Biphenyl und Diphenylmethan sowie das gemeinsame Simulationspektrum für beide Verbindungen von Neuron (8,14).

Die größte Abweichung zwischen simuliertem und experimentellen Spektren tritt bei 2900 cm^{-1} auf. Die stärkere Absorption in diesem Bereich geht im wesentlichen auf das Spektrum von Pent-3-enylbenzol aus dem Trainingsdatensatz zurück, welches ebenfalls diesem Neuron assoziiert wurde. Das Netz mittelt das gespeicherte Spektrum aus den Spektren von Biphenyl und Pent-3-enylbenzol sowie zu einem geringeren Anteil aus den Spektren der Verbindungen, die den Nachbarneuronen assoziiert wurden. Dies zeigt, daß das neuronale Counterpropagation-Netz nicht einfach ein Informationsspeicher ist, sondern der Lernalgorithmus eine Summierung und Wichtung der Informationen vornimmt. Bei den anderen Abweichungen im Spektrum, überwiegend zwischen 1400 und 900 cm^{-1}, liegt das simulierte Spektrum zwischen dem von Biphenyl und Diphenylmethan.

Das Speichern von Verbindungen, wie **R13-R16**, und ihrer Infrarotspektren im selben Neuron führt zu einer starken Kompression der Information. In diesem Fall ist die alleinige Nutzung der partiellen Atomladung als Atomeigenschaft im 3D-MoRSE Code eine zu starke Vereinfachung. Um zwischen diesen Verbindungen unterscheiden zu können, sind zusätzliche Atomvariablen im 3D-MoRSE Code notwendig bzw. die Aneinanderreihung zweier 3D-MoRSE Codes, von denen der zweite unpolare Atome gleichberechtigt zu geladenen codiert. So wäre es beispielsweise denkbar, 32 3D-MoRSE Codewerte mit $A_i = q_{tot,i}$ und weitere 32 3D-MoRSE Codewerte mit $A_i = VDW\text{-}Radius$ als Codierung zu nutzen.

In einem Nachbarneuron, ist das Spektrum von Phenylcyclopenten-1 (**R17**) gespeichert. Die Einführung eines para-Fluor-Substituenten in (**R17**), die (**R4**) ergibt, führt zu einer Verlagerung der Verbindung weg von dem Cluster der reinen Kohlenwasserstoffe hin zu Neuron (5,20). Der elektronegative Fluorsubstituent hat einen deutlichen Einfluß auf Ladungsverteilung im Molekül, zumal er an der Konjugation der Elektronen durch den Benzolring bis hinauf in die Doppelbindung des Cyclopentens beteiligt ist.

Die Einführung eines Stickstoffatoms direkt am Benzolring, wie in **R9** und **R10**, bringt die Verbindungen in einen Cluster aus den Neuronen (9,16) und (10,15) in direkter Nachbarschaft zu den Kohlenwasserstoffen und dem 1-Phenylpyrols. Die simulierten Infrarotspektren von **R9** und **R10** haben Korrelationskoeffizienten von 0.910 bzw. 0.998.

Ein Sauerstoffsubstituent am Benzolring hat einen weitaus drastischeren Einfluß auf die Ladungsverteilung, so daß **R2** und **R3** weiter entfernt vom Cluster der Kohlenwasserstoffen, in die Neuronen (1,19) und (3,17) gemappt werden. Für **R2** wird dabei ein Korrelationskoeffizient zwischen experimentellem und simulierten Spektrum von 0.999 und für **R3** von 0.639 erhalten. Die experimentellen Spektren für drei Verbindungen **R1** - **R3** sind in Abbildung 55 abgebildet. Azobenzol (**R1**) mit einem Korrelationskoeffizienten von 0.995 fällt, aufgrund der Polarität der Azogruppe, in dieselbe Region des Netzes.

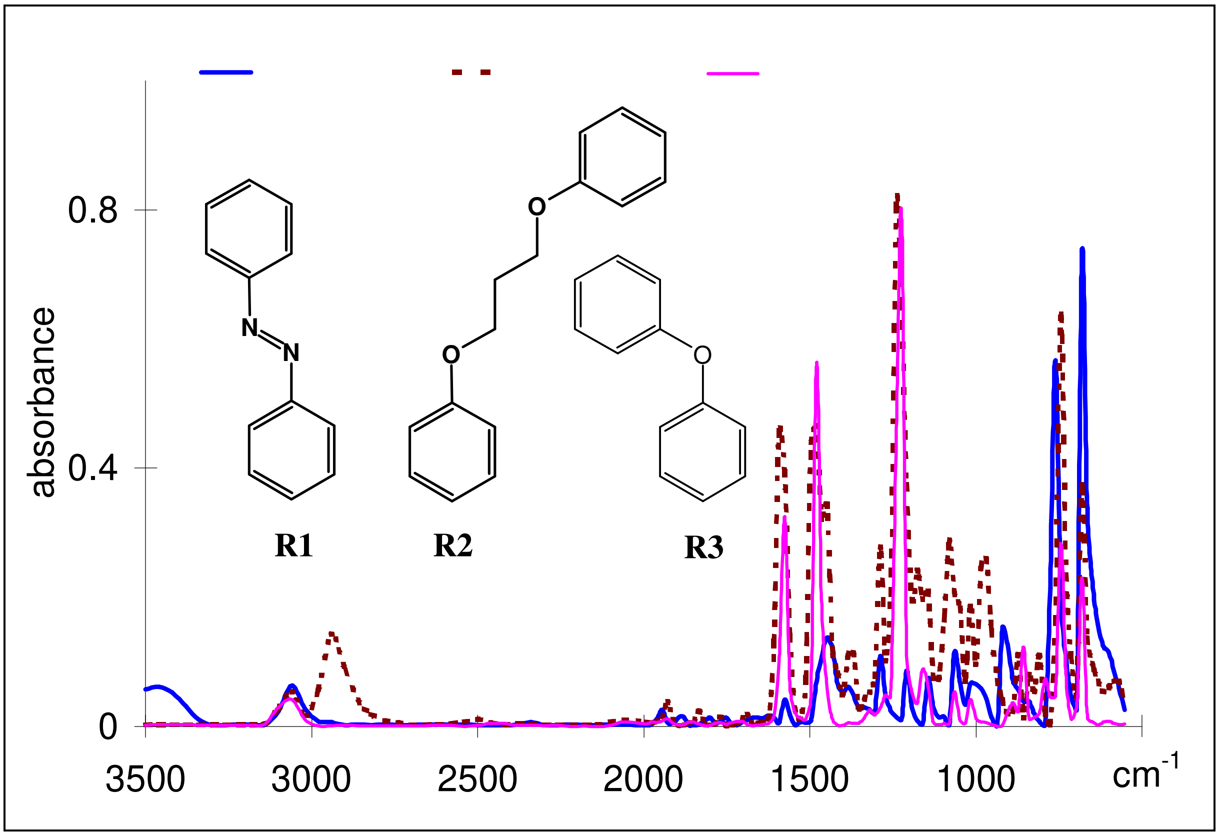


Abbildung 55: Die experimentellen Spektren von Diphenyl-azen (R1), Diphenylether (R2) und 1,3-Diphenoxy-propan (R3), die hier gemeinsam abgebildet sind, weil alle drei Verbindungen in einem Cluster an der linken oberen Kante des Netzes landen.

Acetale und Ketale (**R5**-**R8**) formen einen eigenen Cluster, deutlich separiert von den Stoffklassen, die zuvor diskutiert wurden. Verständlich, wenn man den starken Einfluß der zwei Sauerstoffatome auf die Ladungsverteilung im Molekül bedenkt, obwohl diese nur induktiv wirkt, im Gegensatz zu dem mesomeren Effekt der Heteroatomsubstituenten in den vorstehend behandelten Verbindungen. Die große strukturelle Ähnlichkeit von **R7** und **R8** wird vom CPG-Netz erkannt, weshalb sie demselben Neuron (23,20) assoziiert werden. Konsequenterweise zeigen die Infrarotspektren große Ähnlichkeit, die in Korrelationskoeffizienten zwischen experimentellen und simuliertem Spektrum von 0.998 bzw. 0.932 ihren Ausdruck findet.

Das Lacton **R11** unterscheidet sich von allen zuvor diskutierten Strukturen so stark, daß es in einen völlig anderen Bereich des Netzes gemappt wird, als alle anderen Verbindungen zuvor. Der Korrelationskoeffizient für **R11** zwischen experimentellem und simuliertem Spektrum beträgt 0.850.

6.6.6 *Beobachtungen im Netzwerk*

Ein mit 3D-MoRSE Codes und Infrarotspektren von Verbindungen trainiertes neuronales Counterpropagation-Netz sollte zwei wesentliche Eigenschaften zeigen: die nicht lineare Interpolation sowie die gewichtete Summation von IR-Spektren. Beide Eigenschaften sollen im folgenden Abschnitt detailliert untersucht werden.

6.6.6.1 *Interpolation und gewichtete Summierung von Infrarotspektren durch das neuronale Counterpropagation-Netz*

Die Lernmethode der CPG-Netze hat zwei Konsequenzen:

1. Wenn zwei Verbindungen aus dem Trainingsdatensatz demselben Neuron assoziiert werden, ist das im Neuron gespeicherte Spektrum eine Mischung (gewichtete Summierung) aus den experimentellen Spektren der zwei Verbindungen.

2. Auch die Gewichte von Neuronen sind signifikant denen im Erinnerungstest keine Verbindung assoziiert wurde (leere Neuronen).

Beide Effekte gewichtete Summierung von Spektren und die Signifikanz leerer Neuronen, welche die CPG-Netze zur Interpolation befähigt, sollen anhand von Beispielen analysiert werden.

<u>Gewichtete Summierung von IR-Spektren.</u> Ein Beispiel hierfür ist Neuron (23,6). Aus dem Trainingsdatensatz werden drei Verbindungen: Benzoesäureethylester (**M1**), 3-(Chlormethyl)-benzoesäuremethylester (**M2**) und 4-Methylbenzoesäuremethylester (**M3**) in dieses Neuron gemappt, was zu einer gewichteten Summierung ihrer IR-Spektren führt. Die Korrelationskoeffizienten zwischen den experimentellen Spektren der drei Verbindungen und dem simuliertem Spektrum, das im Neuron (23,6) gespeichert wurde betragen 0.948, 0.899 bzw. 0.956.

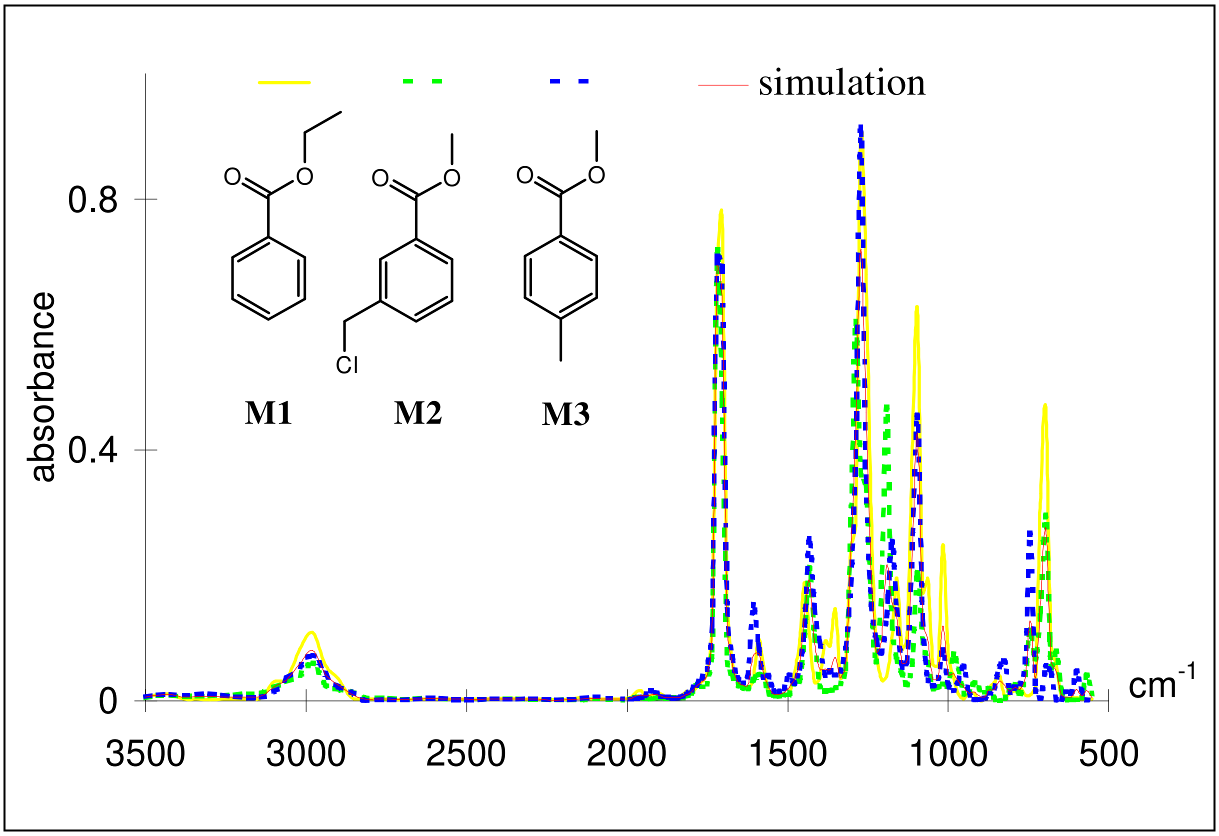


Abbildung 56: Experimentelle Spektren von Benzoesäureethylester (M1), 3-(Chlormethyl)-benzoesäuremethylester (M2) und 4-Methylbenzoesäuremethylester (M3) aus dem Trainingsdatensatz, die im Erinnerungstest dem Neuron (23,6) assoziiert wurden und das zugehörige Simulationsspektrum.

Für 4-(Brommethyl)-benzoesäuremethylester, **M4**, aus dem Testdatensatz, der ebenfalls in dieses Neuron projiziert wird, beträgt der Korrelationskoeffizient zwischen experimentellem und simuliertem IR-Spektrum 0.965. Damit ist der Korrelationskoeffizient für **M4** aus dem Testdatensatz besser als für die drei Verbindungen des Trainingsdatensatzes. Abbildung 57 zeigt noch einmal dieses interessante Ergebnis.

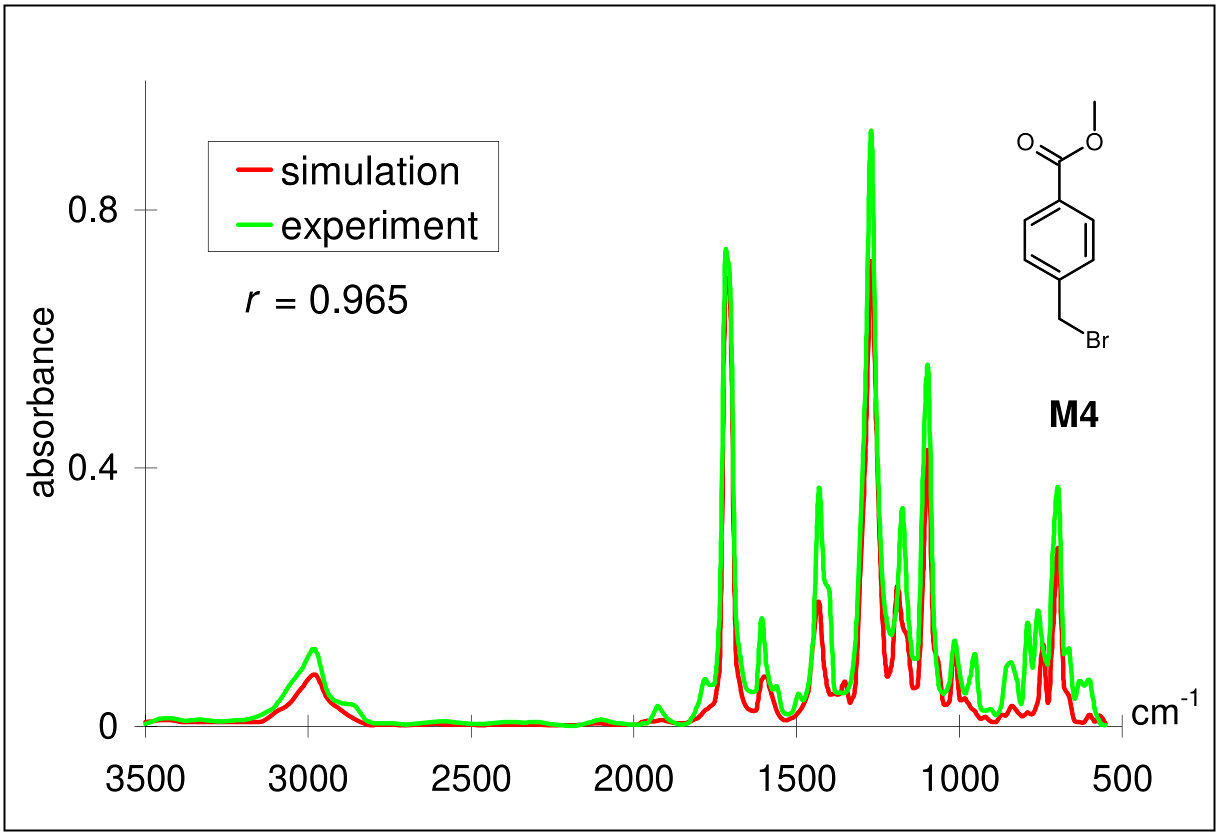


Abbildung 57: Ergebnis der IR-Spektrensimulation für 4-(Brommethyl)-benzoesäuremethylester (M4) aus dem Testdatensatz mit einem höheren Korrelationskoeffizienten als für alle drei Verbindungen des Trainingsdatensatzes, die ebenfalls in dieses Neuron projiziert wurden.

Interpolation von Spektren. Bei der Abfrage des trainierten Netzes mit dem Strukturcode von 1-(4-Methoxyphenyl)-prop-2-en-1-ol aus dem Testdatensatz wurde die beste Übereinstimmung zwischen dem Strukturcode und Eingabegewichten des Neurons (18,19) gefunden. Diesem Neuron wurde kein Molekül aus dem Trainingsdatensatz assoziiert, trotzdem ist im Ausgabeteil des Neurons ein IR-Spektrum gespeichert und so ist eine Simulation des IR-Spektrums für 1-(4-Methoxyphenyl)-prop-2-en-1-ol möglich. In der Tat stimmt das simulierte Spektrum mit dem experimentellen Spektrum in einem vernüftigen Maß überein, wie es der Korrelationskoeffizient von 0.847 zwischen experimentellem und simuliertem Spektrum anzeigt. Abbildung 58 ermöglicht den visuellen Vergleich von simuliertem und experimentellem Spektrum.

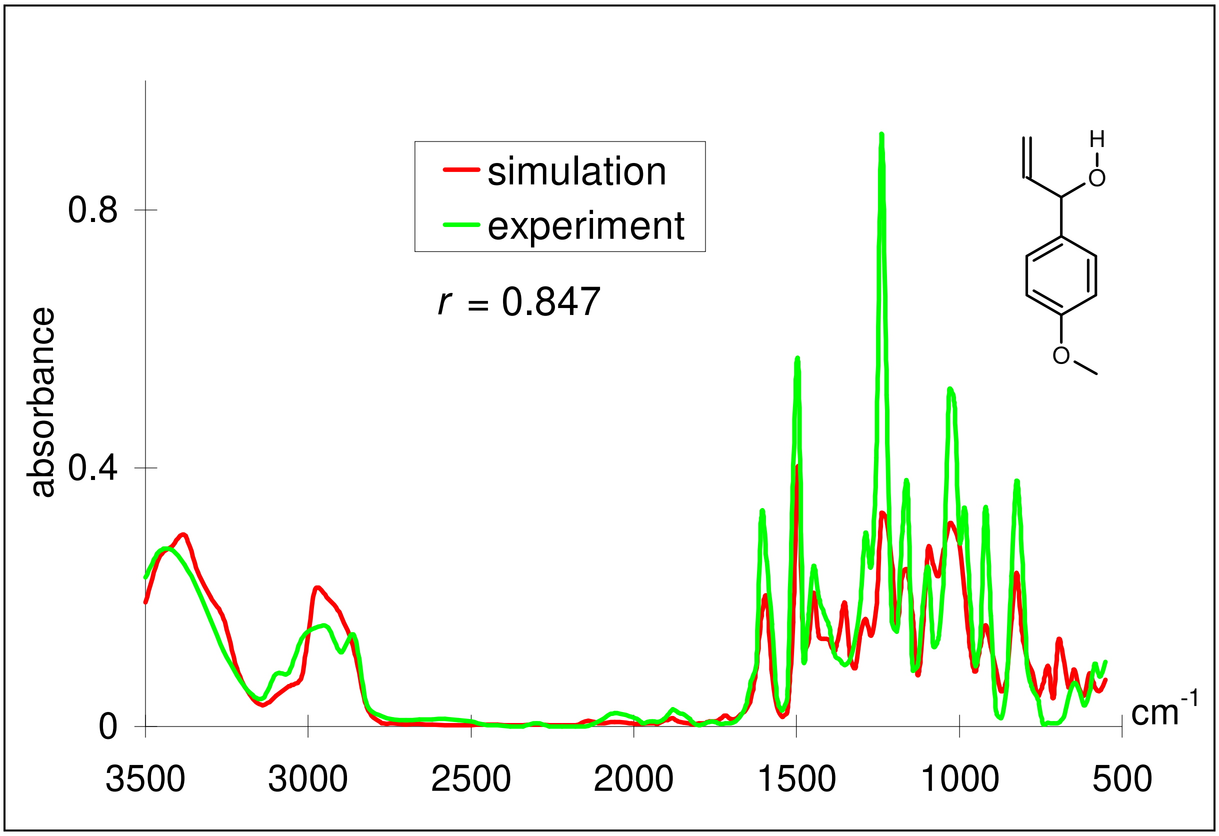


Abbildung 58: Experimentelles und mittels Interpolation simuliertes Spektrum von 1-(4-Methoxyphenyl)-prop-2-en-1-ol.

Die Gewichte des Neurons (18,19) wurden durch die Daten von allen Verbindungen des Trainigsdatensatzes beeinflußt. Wobei den Molekülen aus dem Trainigsdatensatz eine herausragende Bedeutung zukommt, die in die direkt benachbarten Neuronen des Netzes projiziert wurden. Abbildung 59 zeigt die Strukturen der Verbindungen, die in sechs der acht Neuronen der ersten Nachbarschaftssphäre des Neurons (18,19) projiziert wurden. Es ist interessant zu sehen, wie die strukturellen Merkmale von 1-(4-Methoxyphenyl)-prop-2-en-1-ol auf die Verbindungen der Nachbarneuronen verteilt sind. So nimmt die Bedeutung des Methylethers von oben nach unten zu, während die Funktionalität des Prop-2-en-1-ols eher auf der linken Seite des Netzausschnitts zu finden ist.

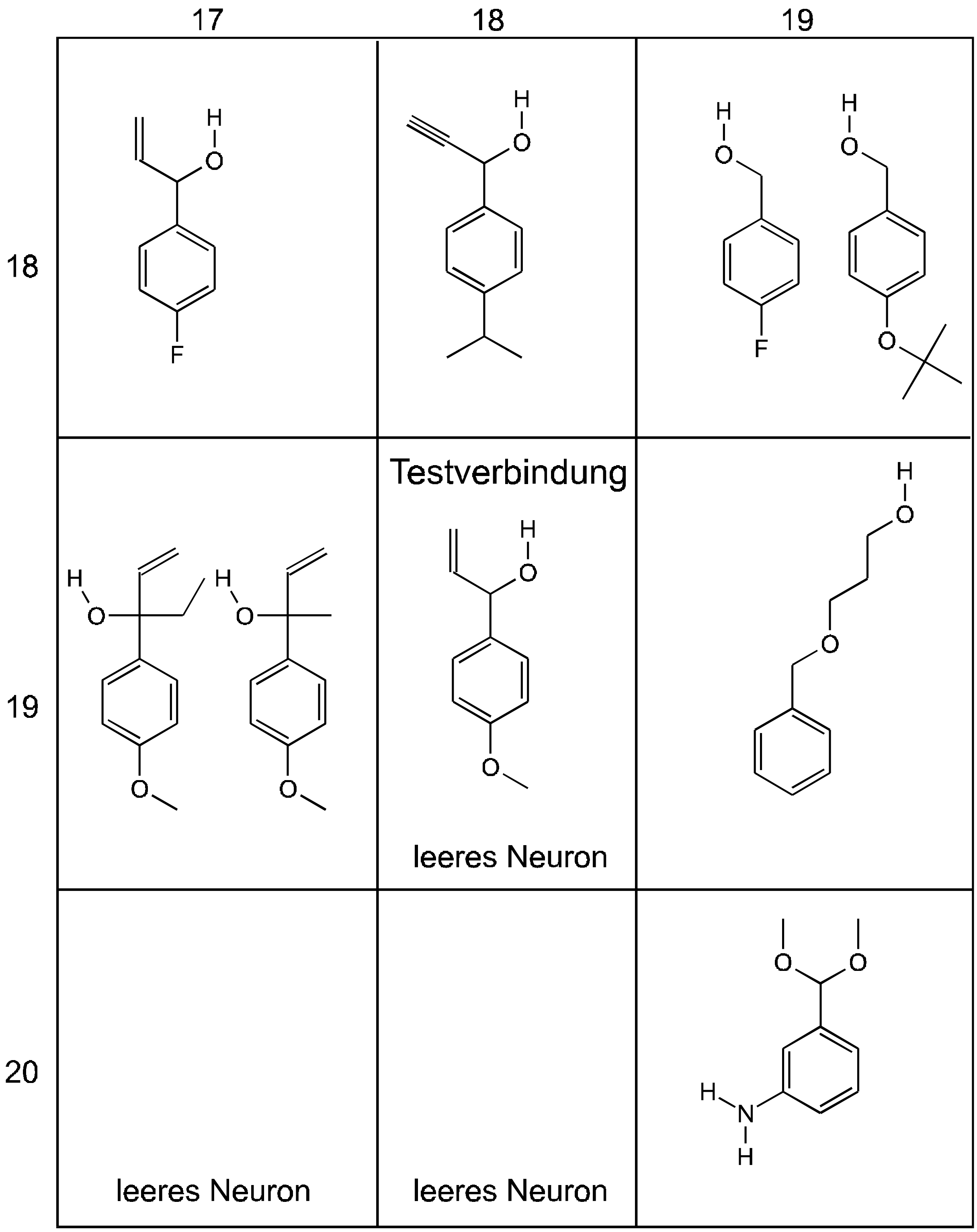


Abbildung 59: Verteilung der Verbindungen des Trainingsdatensatzes im trainierten CPG-Netz, deren IR-Spektren den wesentlichen Einfluß auf die Interpolation des Infrarotspektrums für die Testverbindung hatten.

6.6.7 Resümee der Simulationsergebnisse für mono-, di- und trisubstituierte Benzolderivate

Leere Neuronen, denen im Erinnerungstest kein Molekül des Trainingsdatensatzes assoziiert wurden, erlauben die Interpolation von IR-Spektren für Simulationszwecke. Dies macht die Simulation von Infrarotspektren für Verbindungen auch dann möglich, wenn unmittelbar strukturell verwandte Moleküle im Trainingsdatensatz fehlen. Zum anderen erlaubt die gewichtete Summierung von IR-Spektren auf Neuronen aus Bereichen mit einer hohen Datendichte oftmals gleich für eine ganze Anzahl von Verbindungen ausgewogene Vorhersagen.

Überraschend schlechte Ergebnisse beim Erinnerungstest, wie im Fall der bereits vorgestellten schlechtesten Simulation des Trainingsdatensatzes für 2-Methyl-*2H*-chromen-3-carbaldehyde (siehe auch Abbildung 34) sollten gründlich untersucht werden. Die Ursache für eine solche schlechte Simulation könnte auch in einem Fehler der Datenbasis liegen.

Bei Nutzung dieser Simulationsmethode sollte immer daran gedacht werden, daß sie auf einem Analogieschluß beruht. Kann es zu einem Molekül im Hinblick auf das IR-Spektrum keine analogen Moleküle oder ähnliche Verbindungen geben, so ist eine Simulation mit der hier vorgestellten Methode sinnlos. Dies zeigt das Beispiel von Anisol eindrucksvoll. Dies bedeutet, daß die Simulation von IR-Spektren chemischer Grundstrukturen, wie im Fall der Benzolderivate des Benzols, prinzipiell mit dieser Methode nicht möglich ist. Selbst Verbindungen wie Toluol, Phenol und Vanillin lassen Schwierigkeiten bei der Simulation erwarten. In solchen Fällen ist die quantenmechanische Berechnung der Infrarotspektren dieser Methode vorzuziehen, zumal die meist hohe Symmetrie der Grundstrukturen wie des Benzols, aber auch des Methans, Cyclopropans, Cubans und Naphthalins, die quantenmechanische Berechnung der Spektren erleichtert. Trotz dieser Einschränkung reichen die Strukturanalogien in vielen Fällen aus, um gute bis sehr gute Simulationen zu erlauben.

Ein anderer Punkt, der bei den Simulationen von 2-Methyl-*2H*-chromen-3-carbaldehyde und 6-Chlor-2-(propanoxy-2-en)-benzonitril (siehe auch Abbildung 33 und Abbildung 51) zum Tragen kommt, ist, daß nicht genug ähnliche Verbindungen im Trainingsdatensatz vorhanden sind. Einfach, weil sie in der Datenbasis fehlen oder das Molekül am Rande des Datenraums liegt, wie im Fall von 2-Methyl-*2H*-chromen-3-carbaldehyde. Damit stellt sich die Frage, wie die Datenbasis eines trainierten CPG-Netzes kontinuierlich verbessert werden kann und die dafür notwendigen experimentellen Daten verfügbar werden. Wie die Problematik von Molekülen umgangen werden kann, die an den Rändern des für eine bestimmte Stoffklasse festgelegten

Trainingsdatensatzes liegen. Beispielsweise hätte die Definition des hier verwendeten Datensatzes Heptylbenzol zugelassen Octylbenzol hingegen nicht. Damit ist aber die Interpolation den Infrarotspektrums von Heptylbenzol schlicht unmöglich, da die Definition zwar Hexylbenzol im Trainingsdatensatz erlaubt hätte, nicht aber Octylbenzol.

Nicht zuletzt ist auch die Definition einer Stoffklasse nicht unproblematisch, wie die Moleküle in Abbildung 53 zeigen. Die Festlegung der Grenze einer Stoffklasse: sind Biphenyle oder Diphenylmethane noch Benzolderivate und wie ist es mit Cyclohexylbenzol oder Naphthalin?

Die Lösung, der wir uns in den folgenden Kapiteln langsam annähern wollen, wurde von Steinhauer et al. aufgezeigt.[29,71] Er verzichtete bei der Nutzung von CPG-Netzen vollständig auf definierte Datensätze für Stoffklassen, sondern suchte für jede Abfrage eines Netzes zunächst einen passenden Datensatz aus einer Datenbank zusammen, trainierte mit diesem ein kleines dediziertes Netz, und fragte dieses dann ab. Damit fällt die Problematik der Definition und der Ränder von stoffklassenbezogener Datensätze weg, weil unabhängig von der Definition einer Stoffklasse, die gemäß der 3D-Strukturcodierung ähnlichsten Moleküle einer Datenbank zum Netztraining genutzt werden. Auch ist damit die Verbesserung der Datenbasis des Netzes einfach über die Erweiterung der genutzten Datenbasis zu erreichen.

Der hier kurz beschriebene Ansatz, wird in Kapitel 6.8 als anfrageorientierte Simulation eingehend vorgestellt.

6.7 Cyclohexene

Nach den substituierten Benzolderivaten, deren Gerüst, der Benzolring, planar ist, sollte sich die Untersuchung eines Datensatzes mit einem dreidimensionalen nicht planaren Strukturgerüst anschließen. Hierfür boten sich Cyclohexenderivate an. Zum einen schränkt die endocyclische Doppelbindung die konformative Flexibilität gegenüber einer offenkettigen Verbindung oder Cyclohexan doch etwas ein, zum anderen ist der Cyclohexenring aber immer uneben. Die Doppelbindung ist zwar eben, der Rest des Ringes bildet aber die Sesselkonformation partiell aus (vgl. auch Abbildung 60).

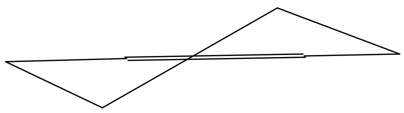


Abbildung 60: 3D-Struktur von Cyclohexen mit der Verteilung der Elektronendichte (AM1)[72].

6.7.1 Der Datensatz

In Analogie zu den Benzolderivaten wurden die Grenzen für den Cyclohexendatensatz wie folgt definiert:

- jede Molekülstruktur muß einen Cyclohexenring enthalten
- Substituenten dürfen keine Kette mit mehr als fünf aufeinanderfolgenden Nicht-Wasserstoffatomen aufweisen
- die Verbindungen dürfen nur aus den Elementen H, C, N, O und Cl bestehen.

Von den 13373 Verbindungen der SpecInfo-Datenbank[11] erfüllten 125 Verbindungen die vorstehenden Auswahlkriterien. Dies sind deutlich weniger als bei den Benzolderivaten (871 Verbindungen) und das, obwohl die Zahl der Substituenten nicht beschränkt wurde und zusätzlich exo-cyclische Doppelbindungen am Ringsystem möglich waren. Trotz der wesentlich kleineren Anzahl von Verbindungen im Datensatz, hat so die chemische Varianz im Datensatz, gegenüber dem vorstehend beschriebenen Datensatz der mono-, di- und trisubstituierten Benzolderivate, zugenommen.

6.7.1.1 Eine Stichprobe als Test

Schon bei den Benzolderivaten zeigten sich Probleme mit Lücken in der Datenbasis. Weitere Arbeiten von P. Selzer[73] über kleine Datensätze an Naphthalin-, Pyridin- und Chinolinderivaten in Kombination mit Isochinolinderivaten führten zu ähnlichen Ergebnissen. Um dies nicht zu wiederholen, war eine Änderung des Verfahrens zur Auswahl von Test- und Trainingsdatensatz notwendig. Zeitgleich zu dem Auftreten dieser Problematik stellte V. Steinhauer sein Konzept der anfrageorientierten Nutzung von Counterpropagation neuronalen Netzen vor.[29] Das Konzept sieht vor, für eine Anfrage, hier der Simulation eines IR-Spektrums, aus einer Datenbank zuerst einen Datensatz mit ähnlichen Eigenschaften auszuwählen, mit diesem dann erst ein spezielles CPG-Netz für die einzelne Anfrage zu trainieren und dieses Netz dann mit den Anfragedaten abzufragen. Anfragedaten wären im Fall der IR-Simlulation, die Werte des 3D-Strukturcodes der Anfragestruktur. Da in diesem Fall der Datensatz bereits vorlag bot es sich an, in Anlehnung an die anfrageorientierte Methodik, aus den 125 Strukturen des Cyclohexendatensatzes einige für den Testdatensatz auszuwählen, deren Verbindungsklasse gut im Gesamtdatensatz repräsentiert war. Die Wahl fiel auf die folgenden drei Strukturen:

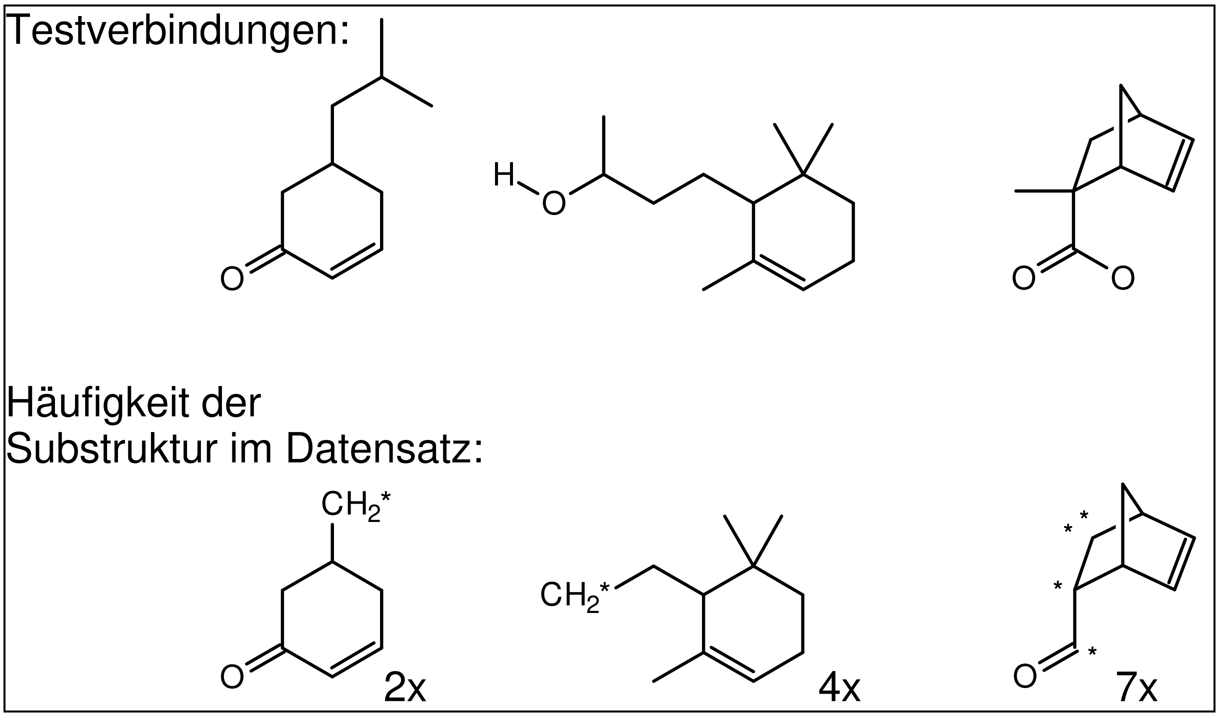


Abbildung 61: Die drei Cyclohexenderivate, die zum Testen der IR-Spektrensimulation ausgewählt wurden (erste Reihe). Die zweite Reihe zeigt, wie oft die entsprechende Substruktur unter allen 125 Cyclohexenderivaten vorkam.

Damit enthält der Trainingsdatensatz 122 Verbindungen und der Testdatensatz besteht aus den drei Verbindungen 5-(2-Methylpropyl)-cyclohex-2-enone, 4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-ol und 2-Methyl-bicyclo[2.2.1]hept-5-en-2-carbonsäure. Die entsprechenden

Substrukturen sind zwei-, vier- bzw. siebenmal im Gesamtdatensatz enthalten. Allerdings ist dabei auf die Definition der Substrukturen zu achten (vgl. Abbildung 61). So ist zwar die 5-Methylen-cyclohex-2-enon - Substruktur nur zweimal im Gesamtdatensatz enthalten, die Cyclohex-2-enon - Substruktur jedoch 35 mal.

Die Auswahl der Testverbindungen erlaubt es, diesen Versuch als Vorversuch für eine anfrageorientierte Simulation anzusehen. Und in der Tat, die hier erzielten Ergebnisse sind mit den Ergebnissen, die später mit Hilfe der anfrageorientierten Simulation erzielt wurden, nahezu identisch.[30]

6.7.1.2 Die Codierung

Zur Codierung der Cyclohexenderivate wurde der 3D-MoRSE Code mit folgenden Parametern verwendet: n=64, $A_i = q_{tot,i}$, s_{max} = 15.5 $Å^{-1}$. Gegenüber den Benzolderivaten wurde die Zahl der Werte auf 64 erhöht und das Maximum für das Maß des Beugungswinkels, s_{max}, auf 15.5 $Å^{-1}$ herabgesetzt. Damit wird der Einfluß der funktionellen Gruppen auf den Code herabgesetzt, indem die Bedeutung kleiner Abstandsänderungen sinkt und die Bedeutung größere Abstände steigt. Dies schien notwendig angesichts des Simulationsfehlers für 6-Chlor-2-(propanoxy-2-en)-benzonitril (Abbildung 51), das auf demselben Neuron landete wie 5-Phenoxypentannitril.

Inwieweit die Änderung der Codierungsparameter gut war, wird anhand der Ergebnisse dieses Versuchs diskutiert werden.

6.7.2 Das Netztraining und die Größe des Netzes

Zur Darstellung und Trainings des CPG-Netzes im Computer wurde das Programm *kmap*[70] mit den folgenden Parametern genutzt:

- Netztopologie: toroidal
- erste Lernrate: 0.95
- automatisch Anpassung von Lernrate und Korrekturentfernung
- Netzgröße und maximale Korrekturentfernung siehe Text

Da aufgrund der hohen Varianz im Datensatz eine genaue Abschätzung der benötigten Netzgröße nicht möglich war, wurde die Kantenlänge des Netzes zwischen fünf und 25 Neuronen in Schritten von fünf Neuronen variiert. Dabei wurde die maximale Korrekturentfernung vom nächstliegenden Neuron auf ein Drittel der Kantenlänge des Netzes festgelegt. Bei allen Netzen wurde eine toroidale (Ringoberfläche) Netztopologie gewählt, damit keine Stoffklasse an den Rand des Netzes gedrängt werden konnte. Das Training eines jeden Netzes benötigte rund 30000 Iterationen. Zur Auswahl des besten Netzes, dessen Ergebnisse anschließend diskutiert werden sollen, wurde jedes Netz mit den drei Testmolekülen getestet. Aus dem anschließenden Vergleich der Ergebnisse schnitt das Netz mit einer Kantenlänge von 20 Neuronen am besten ab. Die mit diesem Netz erhaltenen Ergebnisse sollen nachfolgend diskutiert werden.

Nur die beiden Netze mit einer Kantenlänge von 10 bzw. 20 Neuronen können je eines der drei Testmoleküle so vorhersagen, daß das simulierte Spektrum dem experimentellem des Testmoleküls ähnlicher ist, als allen anderen Spektren des Datensatzes. Bei dem CPG-Netz mit einer Kantenlänge von 10 Neuronen beträgt der mittlere Korrelationskoeffizient für die drei Testmoleküle 0.922 bei dem CPG-Netz mit einer Kantenlänge von 20 Neuronen beträgt dieser 0.937. Warum das Netz mit 15 Neuronen Kantenlänge schlechter ist als die Netze mit 10 und 20 Neuronen Kantenlänge, läßt sich aus der Verteilung der Korrelationskoeffizienten für die drei Testverbindungen vermuten. Bei 10 Neuronen Kantenlänge variieren die Korrelationskoeffizienten um 0.01 und das Netz mußte die Information aus 122 Trainingsverbindungen in 100 Neuronen speichern. Damit hat das Netz mit einer Kantenlänge von 10 Neuronen vermutlich die Information gut verallgemeinert. Bei 20 Neuronen Kantenlänge liegen die Korrelationskoeffizienten 0.857, 0.966 und 0.988. Dies weißt darauf hin, daß bei den Molekülen, wo eine gute Simulation durch sehr eng verwandte Verbindungen möglich war (vgl. Kapitel 6.7.3), die Chance dazu durch eine stärkere Differenzierung im Netzwerk genutzt wurde, während im Fall von 2-Methyl-bicyclo[2.2.1]hept-5-en-2-carbonsäure sich die mangelnde Verallgemeinerung im Netzwerk negativ auf die Simulationsqualität auswirkte. Das CPG-Netz mit 15 x 15 Neuronen liegt nun dazwischen und wurde vermutlich von den negativen Auswirkungen der beiden Extreme mangelnde Verallgemeinerung und mangelnde Differenzierung in Bezug auf die drei Testmoleküle beeinflußt.

6.7.3 Gute Ergebnisse für den Testdatensatz der Cyclohexenderivate

Die Ergebnisse des Erinnerungstestes für den Trainingsdatensatz sind, angesichts der bewußten Einschränkung dieses Versuches auf die drei Testmoleküle, von untergeordneter Bedeutung und sollen hier nicht im Detail diskutiert werden. Der mittlere Korrelationskoeffizient zwischen experimentellen und simulierten Spektren von 0.997 für den Trainingsdatensatz zeigt, daß die Beziehung zwischen 3D-Struktur und Spektrum für die Verbindungen des Testdatensatzes vom Netz gelernt wurde. Bei einer maximalen Belegung der Neuronen mit zwei Molekülen des Trainingsdatensatzes, verbunden mit einem niedrigsten Korrelationskoeffizienten von 0.947, können auch keine Anzeichen für eine zu starke Kompression des Datenraumes gefunden werden.

Die beste Simulation des Testdatensatzes wurde für 5-(2-Methylpropyl)-cyclohex-2-enon erreicht. Die Korrelationskoeffizienten für die drei Testmoleküle finden sich in Tabelle 8.

Tabelle 8: Die Korrelationskoeffizienten zwischen experimentellem und simuliertem Spektrum für die drei Testverbindungen des Cyclohexendatensatzes.

Testverbindung	5-(2-Methylpropyl)-cyclohex-2-enone	4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-ol	2-Methyl-bicyclo[2.2.1]hept-5-en-2-carbonsäure
r	0.988	0.966	0.857

6.7.3.1 5-(2-Methylpropyl)-cyclohex-2-enon

Die Simulation von 5-(2-Methylpropyl)-cyclohex-2-enon erfolgte mittels Neuron (13,16). Dieses Neuron wurde im wesentlichen von 5-(1-Methylethyl)-cyclohex-2-enon beeinflußt, dessen experimentelles IR-Spektrum mit dem im Neuron gespeicherten Spektrum identisch ist. Der Unterschied zwischen dem Test- und Trainingsmolekül beträgt eine CH_2-Einheit in der Seitenkette. Dieser Unterschied ist im IR-Spektrum kaum sichtbar, wie schon der Korrelationskoeffizient von 0.988 zeigt. Der einzig sichtbare Unterschied ist, wie Abbildung 62 zeigt, die schwächer ausgeprägte Struktur der CH-Bande um 3000 cm^{-1}. Ansonsten sind simuliertes und experimentelles Spektrum nahezu identisch. Der Unterschied in der CH-Bande beruht nicht auf unterschiedlichen Meßbedingungen, wie man leicht vermuten könnte. Beide Spektren wurden laut SpecInfo-Datenbank als Kuevettenspektren gemessen.

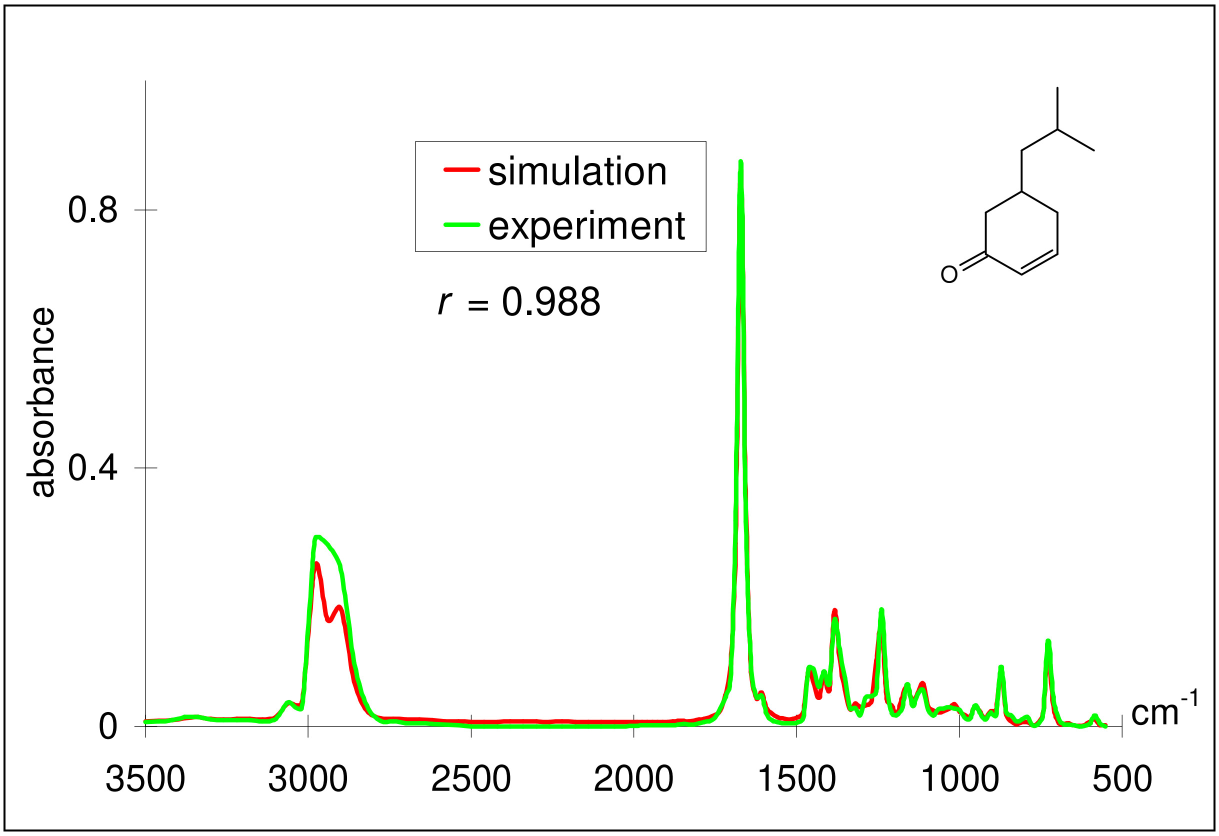


Abbildung 62: Experimentelles und simuliertes IR-Spektrum von 5-(2-Methylpropyl)-cyclohex-2-enon, die beste Simulation des Cyclohexendatensatzes.

Dieses Ergebnis ist erfreulich, wenn auch nur wenig verwunderlich, da mit 5-(1-Methylethyl)-cyclohex-2-enon nur ein zweites eng verwandtes Molekül im Trainingsdatensatz vorhanden war (siehe auch Abbildung 61). Daß der geringe Unterschied der beiden Moleküle von einer Methylengruppe erkannt und zur Simulation genutzt wurde, ist ganz im Sinne der Methode.

6.7.3.2 4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-ol

Beim zweiten Testmolekül 4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-ol ist die Situation anders. Dem Neuron (8,3), das dem Testmolekül nach dem Training am ähnlichsten war, wurde kein Molekül des Trainingsdatensatzes assoziiert. Das bedeutet, das Simulationspektrum wurde im Fall von 4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-ol vom trainierten CPG-Netz mittels eines leeren Neurons interpoliert. Beeinflußt wurde das leere Neuron (8,3) im Training hauptsächlich von folgenden Molekülen des Trainingsdatensatzes (in Klammern das Neuron dem sie assoziiert wurden): 1-Bicyclo[2.2.1]hept-5-en-2-yl-ethanol (7,2), 4-Isopropylcyclohex-2-enol (7,4), 2,4,4-Trimethylcyclohex-2-enol (7,4), 1-(4-Isopropyl-1-methylbicyclo[2.2.2]oct-5-en-2-yl)-ethanol (8,2), 3-(4-Methylcyclohex-3-enyl)-butan-1-ol (9,2), 2-Methyl-4-(2,6,6-

trimethylcyclohex-2-enyl)-but-3-en-2-ol (9,4), 5-(2,6,6-Trimethylcyclohex-1-enyl)-3-methyl-pent-1-en-3-ol (9,4). Abbildung 63 veranschaulicht dies graphisch.

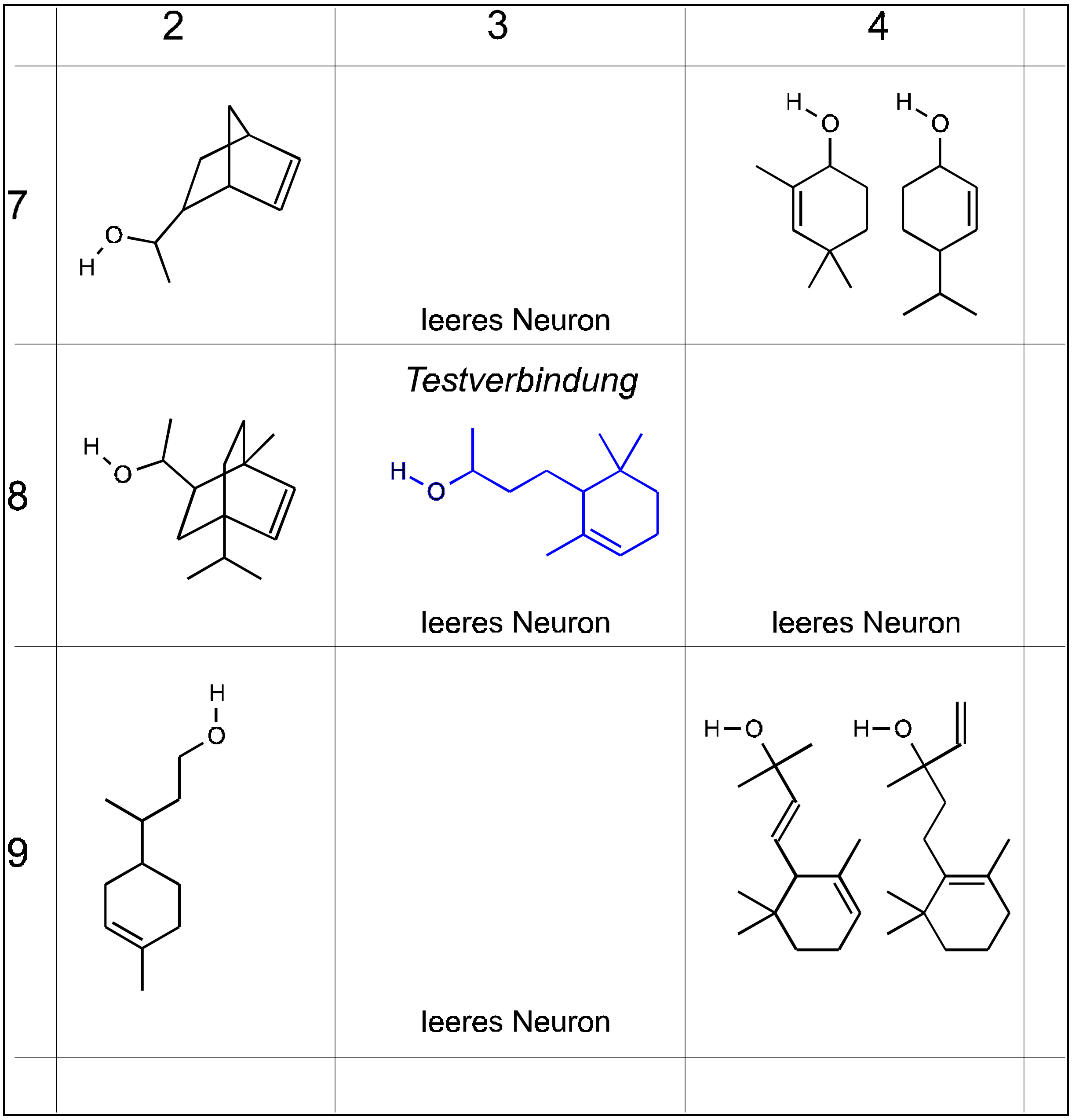


Abbildung 63: Verteilung der Trainingsmoleküle mit wesentlichem Einfluß auf das Neuron (8,3), das zur Simulation des IR-Spektrums von 4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-ol benutzt wurde.

Allen Molekülen gemeinsam ist die Alkoholfunktionalität, was für die Simulation wichtig ist, da die alkoholische Substruktur das Spektrum erheblich beeinflußt, während alle anderen Banden, außer der CH-Bande, nur eine schwache Intensität zeigen. Damit ist die polare OH-Bindung spektrenbestimmend, da eine Breite und intensive CH-Bande allen Cyclohexenderivaten gemeinsam ist. Insofern ist es erfreulich, daß auf den Neuronen rings um das Neuron (8,3) keine Verbindungen mit anderen Funktionalitäten als der OH-Gruppe zu finden sind. Eine genauere Betrachtung der Verbindungen, die den Neuronen um das zur Simulation genutzte Neuron (8,3) herum assoziiert wurden, läßt erkennen, wie die einzelnen Trainingsverbindungen Strukturmerkmale beisteuern um die Simulation zu ermöglichen. So findet sich die ß-ständige Methylgruppe auf den Neuronen (7,2) und (8,2), die lange gesättigte Kohlenwasserstoffkette auf Neuron (9,2) und der passend substituierte Cyclohexenring ist gleich zweimal auf Neuron (9,4) vorhanden. Getrübt wird das Bild etwas durch die Verbindungen auf Neuron (7,4), deren Hydroxygruppen als einzige direkt am Cyclohexenring sitzen statt an einer aliphatischen Seitenkette. Allerdings gibt es durchaus auch Brüche in CPG-Netzen, d.h. die gleiche Entfernung zwischen zwei Neuronen muß nicht unbedingt bedeuten, daß die Unterschiede zwischen den Neuronen gleich groß sind. Das Ergebnis der Simulation zeigt Abbildung 64.

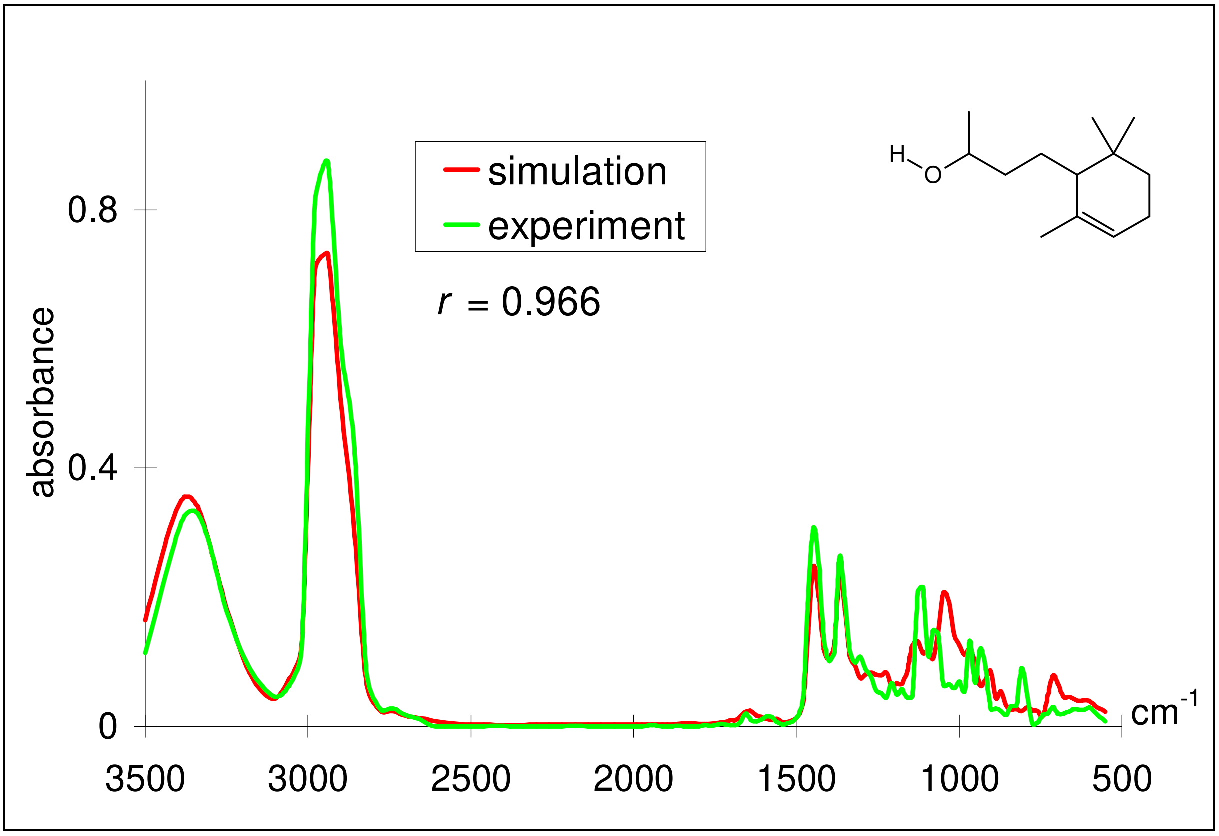


Abbildung 64: Interpoliertes simuliertes und experimentelles IR-Spektrum für 4-(2,6,6-Trimethyl-cyclohex-2-enyl)-butan-2-ol.

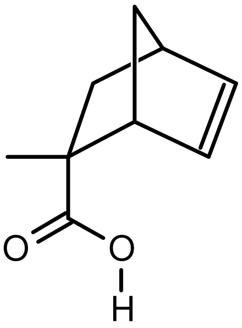
Analysiert man den gesamten Cyclohexendatensatz an, so stellt man fest, daß alle Verbindungen, die neben dem Cyclohexenring nur aliphatische Substituenten und die Alkoholfunktionalität besitzen, auf den Neuronen um Neuron (8,3) landen (insgesamt 18 Verbindungen auf den Neuronen (4,2) bis (12,7)). Damit ist aber auch klar, daß die exakte 3D-Struktur angesichts der Vielfalt der Substitutionsmuster nicht die Rolle spielte. Deutlich wird das auch durch die Tatsache, daß das Aminoderivat des Testmoleküls 4-(2,6,6-Trimethylcyclohex-2-enyl)-butan-2-amin außerhalb des Bereiches der Alkohole dem Neuron (8,20) assoziiert wird. Dieses Neuron ist aufgrund der toroidalen Netztopologie nur drei Neuronen von dem zur Simulation genutzten Neuron (8,3) entfernt und liegt diesem außerhalb des Bereiches der Alkohole am nächsten. Damit deutet sich eine Rangfolge für die Ähnlichkeitserkennung durch das Netz aber auch für die Infrarotspektren an:

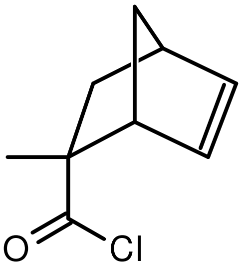
1. Identität der polaren funktionellen Gruppe(n)

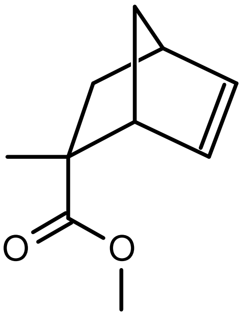
2. die Anordnung der funktionellen Gruppen im Raum (3D-Struktur)

6.7.3.3 2-Methyl-bicyclo[2.2.1]hept-5-en-2-carbonsäure

Zur Simulation des IR-Spektrums von 2-Methyl-bicyclo[2.2.1]hept-5-en-2-carbonsäure (**nc1**) wurde das Neuron (16,7) genutzt, dem im Erinnerungstest 1,4-Dimethylcyclohex-3-en-carbonsäure (**cc1**) assoziiert wurde. Damit wurde entgegen den Erwartungen nicht eine der sechs im Trainingsdatensatz vorhandenen Verbindungen mit der Bicyclo[2.2.1]hept-5-en-2-carbonyl-Grundstruktur (vlg. Abbildung 61) zur Simulation herangezogen. Der Grund hierfür ist, daß keine dieser Verbindungen in α-Stellung zur Carboxylgruppe ein quartärnäres Kohlenstoffatom aufweist. Die einzige Carbonsäure mit einem quartärnären Kohlenstoffatom in dieser Position im Datensatz ist **cc1**. Trainings- und Testverbindung stimmen zudem in der relativen Position der Doppelbindung zur Carboxylgruppe überein. Abbildung 65 zeigt Moleküle aus dem Trainingsdatensatz, die der Testverbindung ähnlich sind und gibt ihre Lage im trainierten neuronalen Counterpropagation-Netz an. Die ähnliche Lage der drei Carbonsäuren in Abbildung 65 läßt sofort das erste Sortierkriterium des neuronalen Netzes erkennen: Die polare funktionelle Gruppe muß identisch sein. Daß im nächsten Schritt die größere Ähnlichkeit des IR-Spektrums von **cc1** zum Infrarotspektrum von **nc1** mit Hilfe des 3D-MoRSE Codes richtig erkannt wurde, ist erfreulich. Denn erwarten würde man aufgrund des gleichen bicyclischen Gerüstes der beiden Norbornencarbonsäuren **nc1** und **nc2**, daß sich deren Infrarotspektren stärker ähneln. Dies bedeutet, daß in diesem Fall für das IR-Spektrum der Unterschied zwischen tertiärem und quartärnärem β-Kohlenstoffatom bzw. das Vorhandensein einer β-ständigen Methylgruppe entscheidender ist, als die Struktur des Cyclohexenrings.





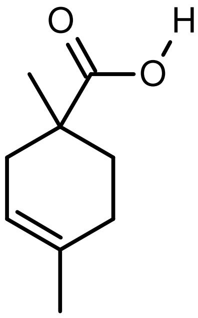


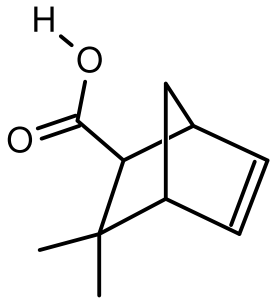
Testverbindung

nc1
Neuron: (16,7)
rms: 0.0
r: 1.0

nd1
Neuron: (3,10)
rms: 0.399
r: 0.055

nd2
Neuron: (20,12)
rms: 0.525
r: 0.458





cc1
Neuron: (16,7)
rms: 0.0913
r: 0.857

nc2
Neuron: (14,5)
rms: 0.186
r: 0.644

Abbildung 65: Die Testverbindung und ähnliche Moleküle aus dem Trainingsdatensatz mit ihrer Lage im CPG-Netz. Der *rms*-Wert zwischen ihrem 3D-MoRSE Code und dem 3D-MoRSE Code *nc1* und der Korrelationskoeffizient *r* zwischen dem IR-Spektrum der Struktur und dem Infrarotspektrum von *nc1* ist angegeben.

Der Vergleich der IR-Spektren der drei Carbonsäurederivate aus Abbildung 66 zeigt, daß lediglich ein Peak mittlerer Intensität bei 712 cm^{-1} auf das Norbornengerüst zurückgeht, da er beiden Norbornencarbonsäurederivaten, **nc1** und **nc2**, gemeinsam ist. Im gesamten übrigen Spektralbereich sind sich die Spektren der beiden β-Methylcarbonsäurederivate **cc1** und **nc1** ähnlicher, wie Abbildung 66 beweist.

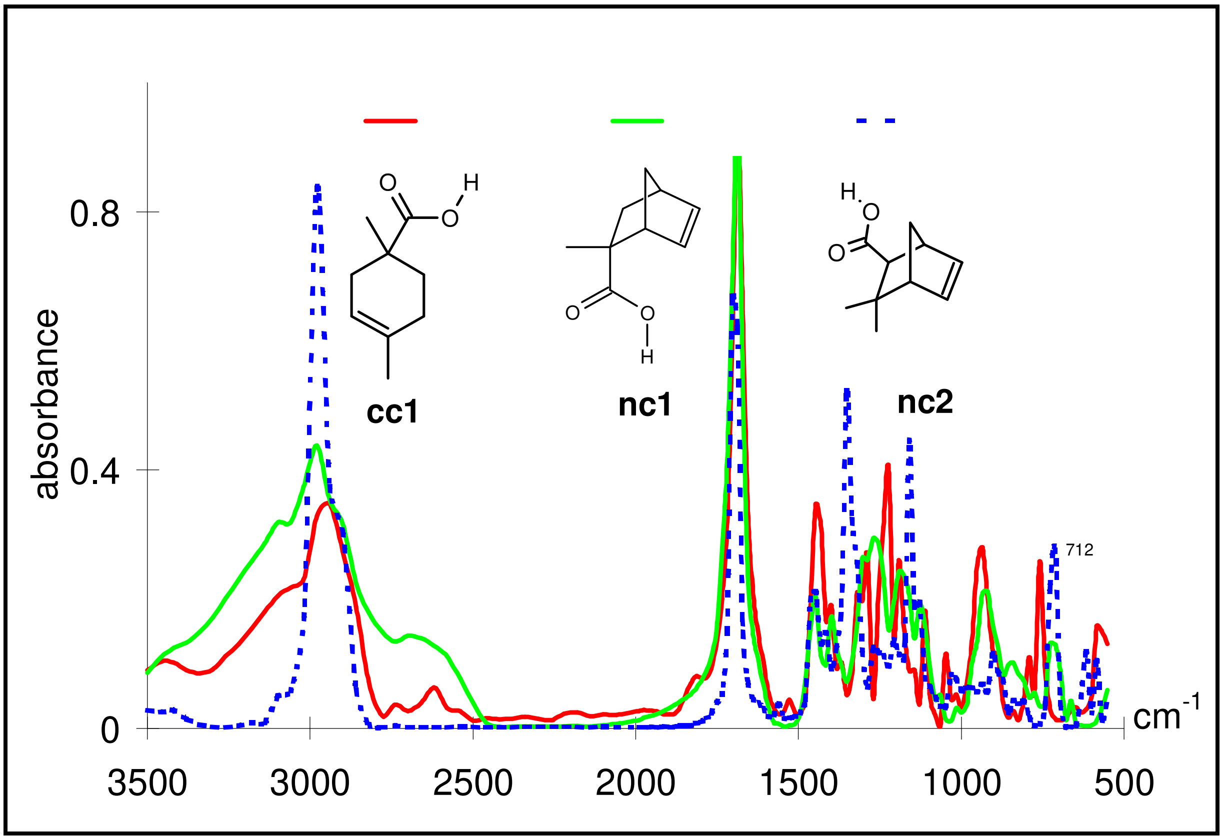


Abbildung 66: Experimentelle Infrarotspektren von Carbonsäurederivaten, die sich im Ringsystem bzw. in Substitution des β-Kohlenstoffatoms unterscheiden. Hierbei ist klar zu erkennen, daß die Substitution des β-Kohlenstoffatoms IR-spektroskopisch wichtiger ist als das Ringsystem.

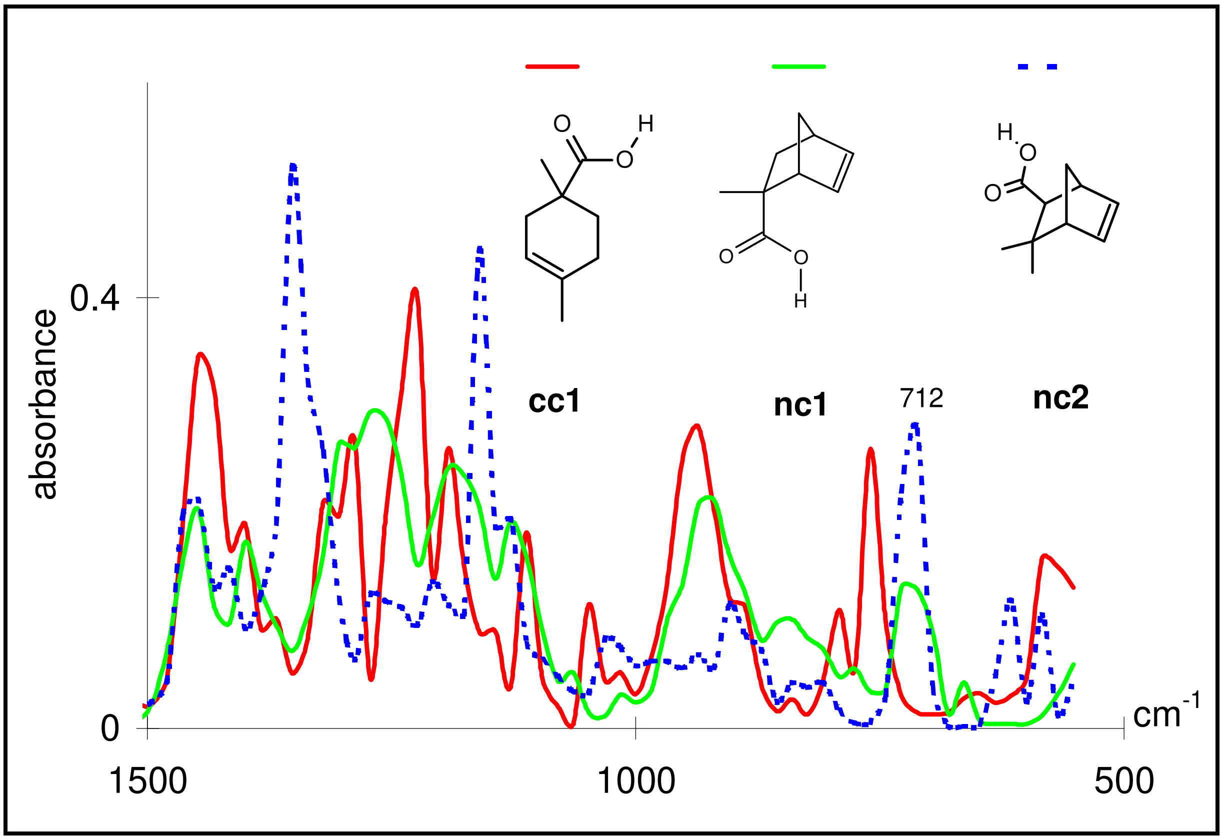


Abbildung 67: Fingerprintbereich aus Abbildung 67.

6.7.4 *Fazit der Simulation für drei Cyclohexenderivate*

Die Simulation von IR-Spektren kann erfolgreich durchgeführt werden, wenn genügend ähnliche Moleküle im Trainingsdatensatz vorhanden sind. Das Konzept der anfrageorientierten Nutzung von CPG-Netzen ist erfolgversprechend.

Die Ergebnisse zeigen, daß für die Spektrensimulation die Identität der polaren funktionellen Gruppe, im Fall der hier besprochenen Cyclohexenderivate der Hydroxy- und Carboxygruppe von größter Bedeutung ist, falls nur eine polare Funktion im Molekül vorhanden ist. In der Bedeutung folgen dann vermutlich die nächste Umgebung der polaren Funktion und dann erst das Molekülgerüst. Die erfolgreiche Spektrensimulation für die drei Testmoleküle spricht für eine weitere Verwendung des 3D-MoRSE Codes mit den für diesen Versuch genutzten Einstellungen: Anzahl der Werte n=64, Atomeigenschaft $A_i = q_{tot,i}$ und Maximalwert für den Beugungswinkelparameter s_{max} = 15.5 $Å^{-1}$.

6.8 Die anfrageorientierte Simulation - Methodik und erste Beispiele

Die anfrageorientierte Simulation von IR-Spektren beruht auf den folgenden Schritten:

- Definition der Anfragestruktur
- Berechnung von Atomeigenschaften und 3D-Struktur der Anfragestruktur
- Berechnung des 3D-Strukturcodes der Anfragestruktur
- Auswahl eines Satzes (Trainingsdatensatzes) von ähnlichen Molekülen auf der Basis der 3D-Strukturcodes aus einer Datenbank mit 3D-Strukturcodes und IR-Spektren
- Training eines neuronalen Counterpropagation-Netzes mit dem Trainingsdatensatz
- Abfrage des trainierten CPG-Netzes mit dem 3D-Strukturcode der Anfragestruktur
- Nutzung des im Ausgabeteil des ähnlichsten Neurons gespeicherten IR-Spektrums als simuliertes Infrarotspektrum.

Abbildung 68 veranschaulicht den Ablauf der anfrageorientierten IR-Spektrensimulation graphisch.

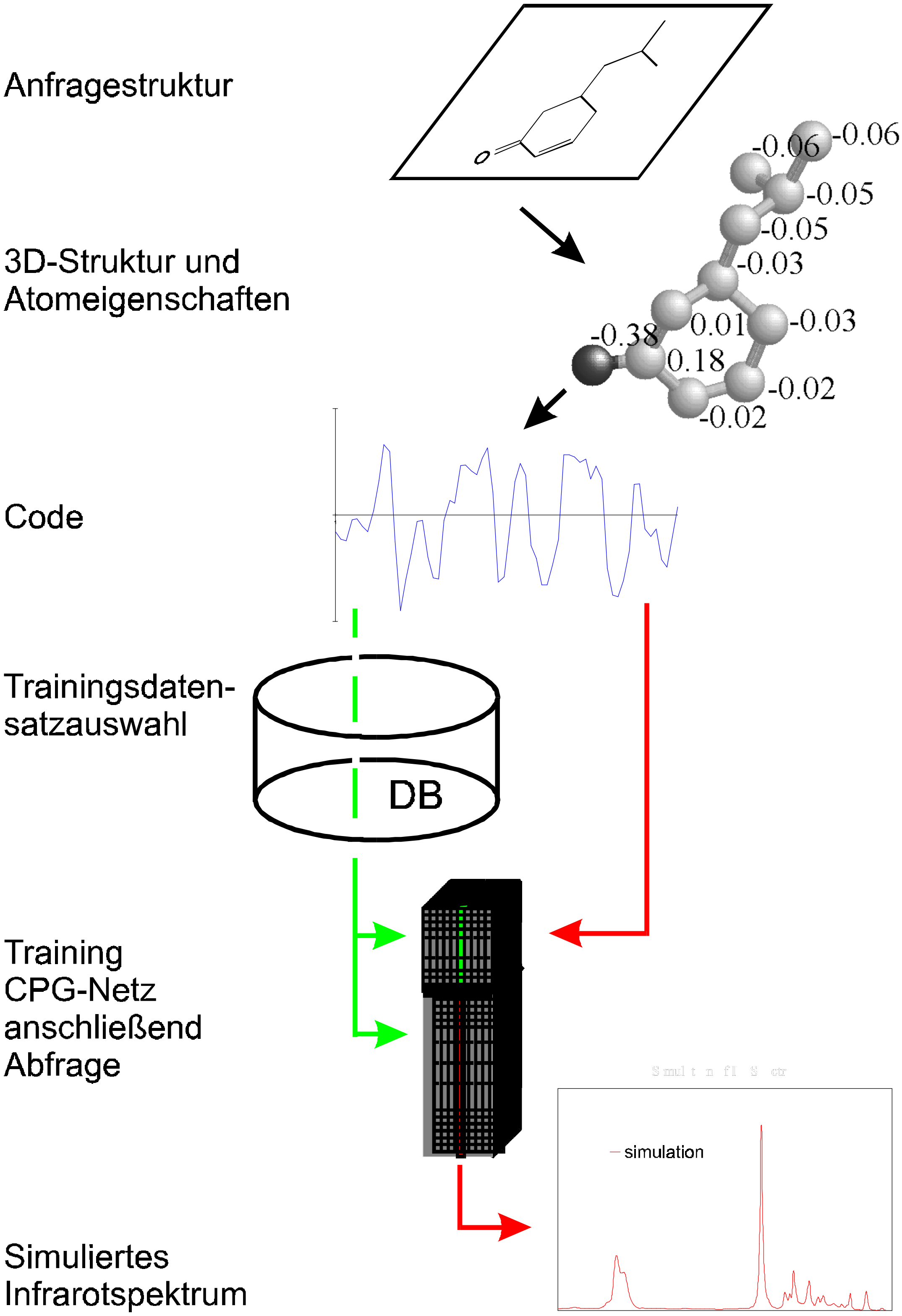


Abbildung 68: Die Methodik der anfrageorientierten Simulation. Wie die Abbildung zeigt, wird der aus eingegebenen Anfragestruktur generierte 3D-Strukturcode zweimal benötigt. Zuerst (hellgrau) um aus einer Datenbank den Trainingsdatensatz mit ähnlichen Molekülen für das CPG-Netz auszuwählen. Dann, nach erfolgtem Training, um das trainierte Netz mit dem Code der Anfragestruktur abzufragen und aus dem ähnlichsten Neuron das simulierte Infrarotspektrum zu extrahieren (dunkelgrau).

Die anfrageorientierte IR-Simulation wurde von P. Selzer und J. Schuur nach einer Idee von V. Steinhauer entwickelt.[30] Die anfrageorientierte Simulation von IR-Spektren bietet vor allem Flexibilität bei der Simulation von IR-Spektren. Die Notwendigkeit zur Definition von Datensätzen für Stoffklassen entfällt. Zudem muß eine Anfragestruktur keiner Stoffklasse mehr zugeordnet werden. Damit stellt sich zugleich das Problem von Verbindungen am Rande des Definitionsbereiches einer Stoffklasse nicht mehr. Da im Rahmen einer anfrageorientierten Simulation der zum Training des Netzes notwendige Trainingsdatensatz dynamisch aus einer Datenbank generiert wird, kann eine Verbesserung der Datenbasis ebenfalls flexibel über eine Erweiterung der Datenbank erfolgen. Die sonst notwendige und zudem mehr oder weniger subjektive Neuauswahl des Trainingsdatensatzes und das erneute Training großer CPG-Netze entfällt somit.

Zwei Probleme sollen nicht verschwiegen werden:

1. Die anfrageorientierte Simulation ist meistens rechenintensiver, da anstatt der Abfrage eines großen neuronalen Netzes ein Trainingsdatensatz aus einer Datenbank ausgewählt und ein kleines CPG-Netz speziell für die Anfrage trainiert werden muß (Zeitbedarf auf einer Sparc 10-40/Solaris 2.5.1 anfrageorientierte Simulation ca. 5 Minuten, Abfrage trainiertes Netz Benzoldatensatz 16.02 s davon 16 s für IO). Erst wenn die Zahl der Neuronen im großen Netz die Zahl der Iterationen beim Training für die anfrageorientierte Simulation um ein vielfaches übersteigt, ist die Rechenzeit gleich oder größer (bei der Verwendung des selben Programms unter Vernachlässigung von IO-Zeiten).

2. Die dynamische Generierung eines Trainingsdatensatzes läßt zwar hoffen, daß dieser genügend ähnliche Verbindungen für eine erfolgreiche Simulation enthält, bietet aber keine Gewähr hierfür. Dieses Problem resultiert zwar aus der Zusammensetzung der verwendeten Datenbank, da aber die Zusammensetzung der Datenbank kaum allen Anwendern der Methode bekannt sein wird, ist es auch ein Problem der Methode.

Das zweite Problem ist von erheblicher Bedeutung, wenn die Möglichkeit zur Automatisierung der anfrageorientierten Simulation genutzt werden soll. Fehlt eine Angabe, ob der dynamisch generierte Trainingsdatensatz genügend ähnliche Verbindungen für eine erfolgreiche Simulation enthält, so ist die Qualität des vorhergesagten IR-Spektrums nicht vorhersehbar. Insofern stellt sich die Frage, wie eine Qualitätskontrolle in den Prozeß der anfrageorientierten Simulation eingebaut werden kann. Ideen hierzu finden sich im Kapitel 6.11.

6.8.1 *Ablauf der automatischen anfrageorientierten Simulation*

Ausgangspunkt der anfrageorientierten Simulation bleibt die Eingabe der Anfragestruktur durch den Chemiker, gefolgt von der Berechnung der Atomeigenschaften durch das Programmpaket *PETRA*[69] und der Abschätzung der 3D-Struktur mit Hilfe des 3D-Strukturgenerators *CORINA*[59-62]. Die Atomeigenschaften und die 3D-Struktur bilden die Voraussetzung zur Berechnung des 3D-Strukturcodes der Anfragestruktur. Als 3D-Strukturcode wird in den folgenden Beispielen, soweit nicht ausdrücklich anders vermerkt, der 3D-MoRSE Code verwandt und zwar mit denselben Randbedingungen (n=64, $A_i = q_{tot,i}$, $s_{max} = 15.5$ Å^{-1}), die sich schon bei den Cyclohexenderivaten (Abschnitt 6.7) bewährt haben. Für die Suche nach ähnlichen Verbindungen wird vorausgesetzt, daß die Datenbank (für die nachfolgenden Untersuchungen alle Verbindungen der SpecInfo Datenbank, für die IR-Vollspektren existieren und die nur H, C, N, O und X (X = Halogene enthalten) vollständig in codierter Form vorliegt. Dabei muß für die Codierung der Datenbank derselbe 3D-Strukturcode mit denselben Randbedingungen verwendet werden, wie für die Anfragestruktur.

Mit dem 3D-Strukturcode der Anfragestruktur wird dann in der codierten Datenbank, nach den 50 Verbindungen mit den zur Anfragestruktur ähnlichsten Strukturcodes gesucht. Die im Rahmen der Suche gefundenen 3D-Strukturcodes und die zugehörigen IR-Spektren dienen als Trainingsdatensatz für das zur Simulation verwendete neuronale Netz. Für die anfrageorientierte Simulation hat sich ein neuronales Counterpropagation-Netz mit acht mal acht Neuronen und planarer Topologie bewährt, das unüberwacht trainiert wird. Über die Abfrage des Netzes mit dem 3D-Strukturcode der Anfragestruktur wird das simulierte IR-Spektrum erhalten. Abbildung 69 zeigt diese Vorgehensweise noch einmal schematisch.

Bei der anfrageorientierten Simulation kommen die folgenden Programme zum Einsatz:

CSED: CACTVS System Editor für die graphische Eingabe chemischer 2D-Strukturen[74]

PETRA: Parameter Estimation for the Treatment of Reactivity Applications[75], als Implementation der PEOE-Methode[63,65] zur Berechnung der partiellen Atomladung, $q_{tot,i}$.

CORINA 3D-Strukturgenetator der eine Konformation mit niedriger Energie aus der Bindungslisten einer Verbindung generiert.[58] Die verwendeten Algorithmen sind in den Referenzen 59 - 62 näher beschrieben.

Code3D Programm zur Codierung der 3D-Struktur von Molekülen,[76] Implementation des 3D-MoRSE Codes, des Radialcodes und der IDH-Codierung.

CHOOSE Programm zur Auswahl des Trainingsdatensatzes aus einer Datenbank, die aus Strukturcodes und Infrarotspektren besteht.[77]

kmap Implementation neuronaler Kohonen- und Counterpropagation-Netze mit der Möglichkeit zur graphischen Analyse der Netze und des Trainingsverlaufs.[70]

j2jcamp Programm zur Auswertung und Formatkonvertierung der Ergebnisse neuronaler Netze.[78]

MS Excel Tabellenkalkulation mit der Möglichkeit der Glättung von Linien in Diagrammen, was zur Darstellung der Infrarotspektren genutzt wurde.[79]

CSIR Cactvs System IR-Spectra Display[80] zur graphischen Anzeige von Infrarotspektren und 2D-Strukturen chemischer Verbindungen.

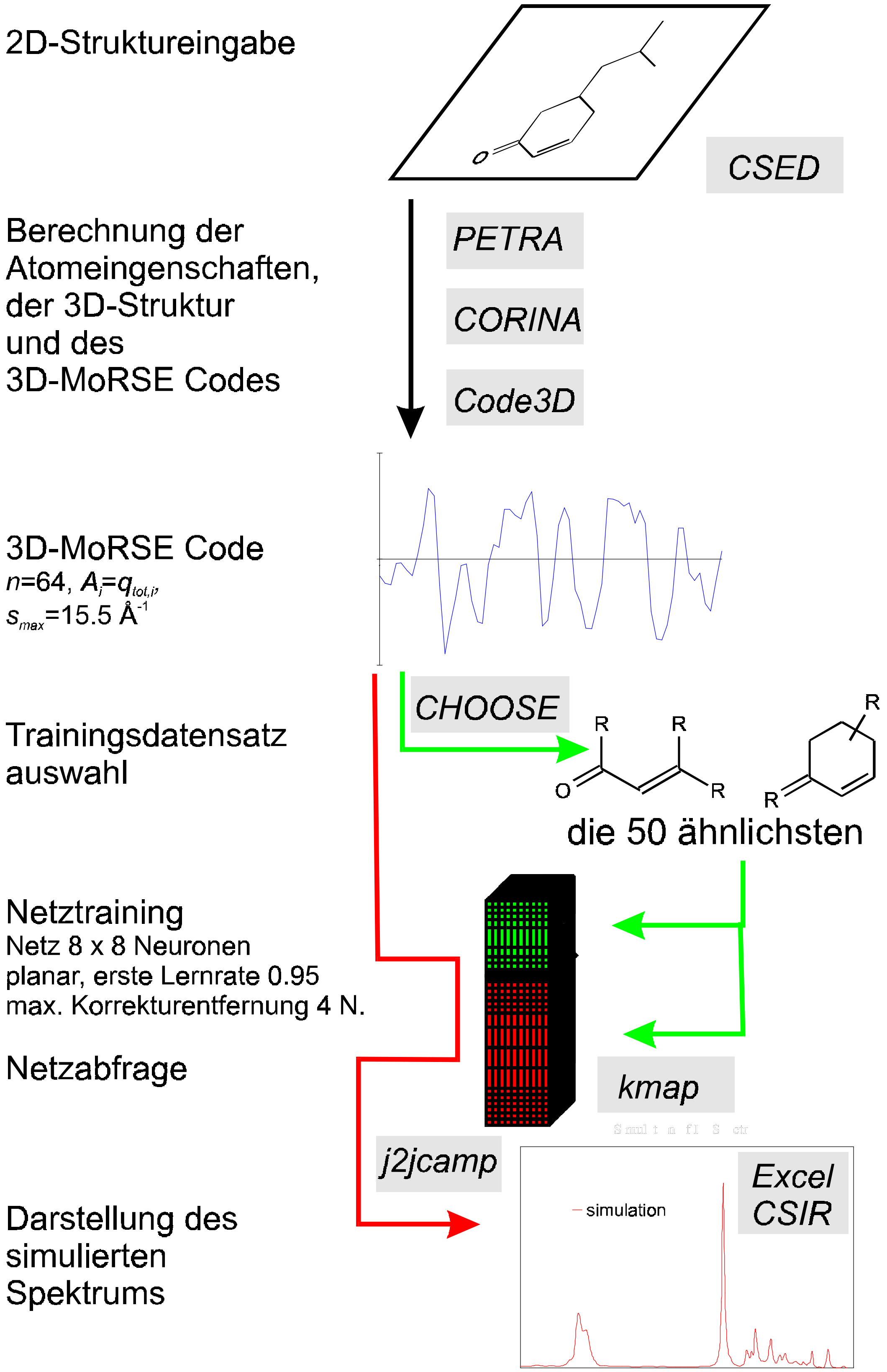


Abbildung 69: Der Ablauf der automatischen anfrageorientierten Simulation, wie er als Skript zu Simulation der Infrarotspektren der nachfolgend vorgestellten Verbindungen realisiert wurde. Grau hinterlegt die Namen der verwendeten Programme.

6.8.2 Anfragorientierte Simulation - ein erster Test mit Molekülen unterschiedlicher Größe

In diesem Kapitel sollen die Simulationen von (1-Methylpropyl)-harnstoff, 5-(2-Metylpropyl)-cyclohex-2-enon und dem Steroid Cholesterin vorgestellt werden. Diese drei Simulationen dienen als erster Test, ob Erfolge mit der anfrageorientierten Simulation unabhängig von der Molekülgröße möglich sind. Alle drei Simulationen wurden mit den Standardwerten für die anfrageorientierte Simulation durchgeführt, d.h. 50 Verbindungen im Trainingsdatensatz, planares CPG-Netz mit 8 x 8 Neuronen, erste Lernrate 0.95 und maximale Korrekturentfernung 4 Neuronen bei automatischer Anpassung von Lernrate und Korrekturentfernung während des CPG-Netztrainings. Das Training benötigte im Mittel rund 50000 Iterationen, die gesamte anfrageorientierte Simulation benötigte im Schnitt 5 min auf einer Sparc 10-40 unter SunOS 5.5.1.

6.8.2.1 (1-Methylpropyl)-harnstoff

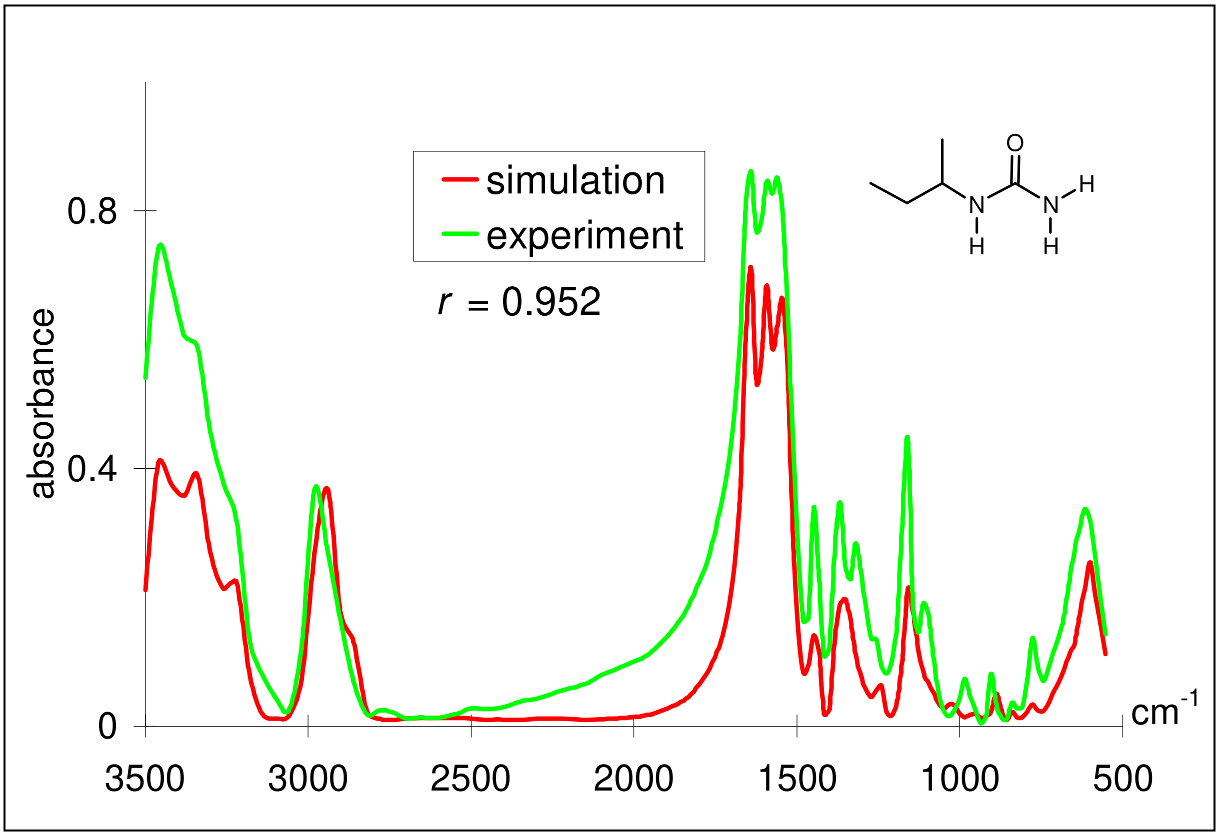


Abbildung 70: Simuliertes und experimentelles Infrarotspektrum von *sec*-Butylharnstoff mit einem Korrelationskoeffzienten von 0.952.

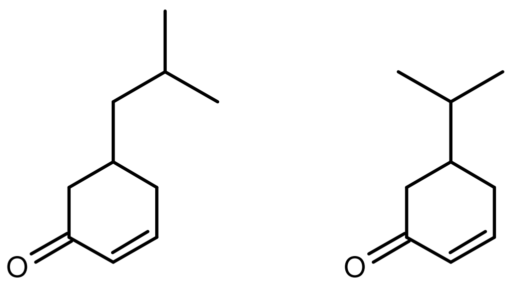
Die Simulation des IR-Spektrums für (1-Methylpropyl)-harnstoff erfolgt mit Hilfe von Neuron (4,1) des trainierten Netzes. Das in diesem Neuron gespeicherte Infrarotspektrum ist nahezu identisch mit den experimentellen Infrarotspektren der beiden Trainingsmoleküle (1,3-

Dimethylbutyl)-harnstoff und Cyclohexylharnstoff, die dem Neuron (4,1) assoziiert wurden. Die Korrelationskoeffizienten zwischen den experimentellen und dem simulierten Spektrum dieser Trainingsmoleküle betragen 0.992 bzw. 0.994.

Der graphische Vergleich des simulierten und des experimentellen Spektrums von *sec*-Butylharnstoff zeigt, daß die Unterschiede zwischen Experiment und Simulation im wesentlichen auf Intensitätsunterschieden beruhen. Einzig drei kleinere Peakschultern unterhalb von 1300 cm^{-1} konnten nicht wiedergegeben werden.

6.8.2.2 5-(2-Metylpropyl)-cyclohex-2-enon

Bei der IR-Spektrensimulation von 5-(2-Metylpropyl)-cyclohex-2-enon sind Ergebnis und Ursache dem vorstehenden Beispiel von (1-Methylpropyl)-harnstoff ähnlich. Der Korrelationskoeffizient für die Simulation ist mit 0.989 etwas höher als beim Harnstoffderivat und mit 5-(1-Methylethyl)-cyclohex-2-enon gibt es ein Trainingsmolekül, daß dem für die Simulation genutzten Neuron (4,1) im Erinnerungstest assoziiert wurde und dessen IR-Spektrum mit dem im Neuron gespeicherten identisch ist (Korrelationskoeffizient 1.00). Abbildung 71 zeigt wie wenig sich die um eine CH_2-Einheit kürzere Seitenkette im IR-Spektrum auswirkt. Lediglich in der Peakspitze der CH-Bande knapp unterhalb von 3000 cm^{-1} ist eine Intensitätsabweichung zwischen simuliertem und experimentellem Spektrum sichtbar. Damit ist das Ergebnis identisch bzw. minimal besser als das Ergebnis was mit einem Trainingsdatensatz aus Cyclohexenderivaten erreicht wurde (vgl. Seite 116 und Abbildung 62).



Schema 8: 5-(2-Metylpropyl)-cyclohex-2-enon und 5-(1-Methylethyl)-cyclohex-2-enon aus dem Trainingsdatensatz dessen Spektrum zur Vorhersage genutzt wurde.

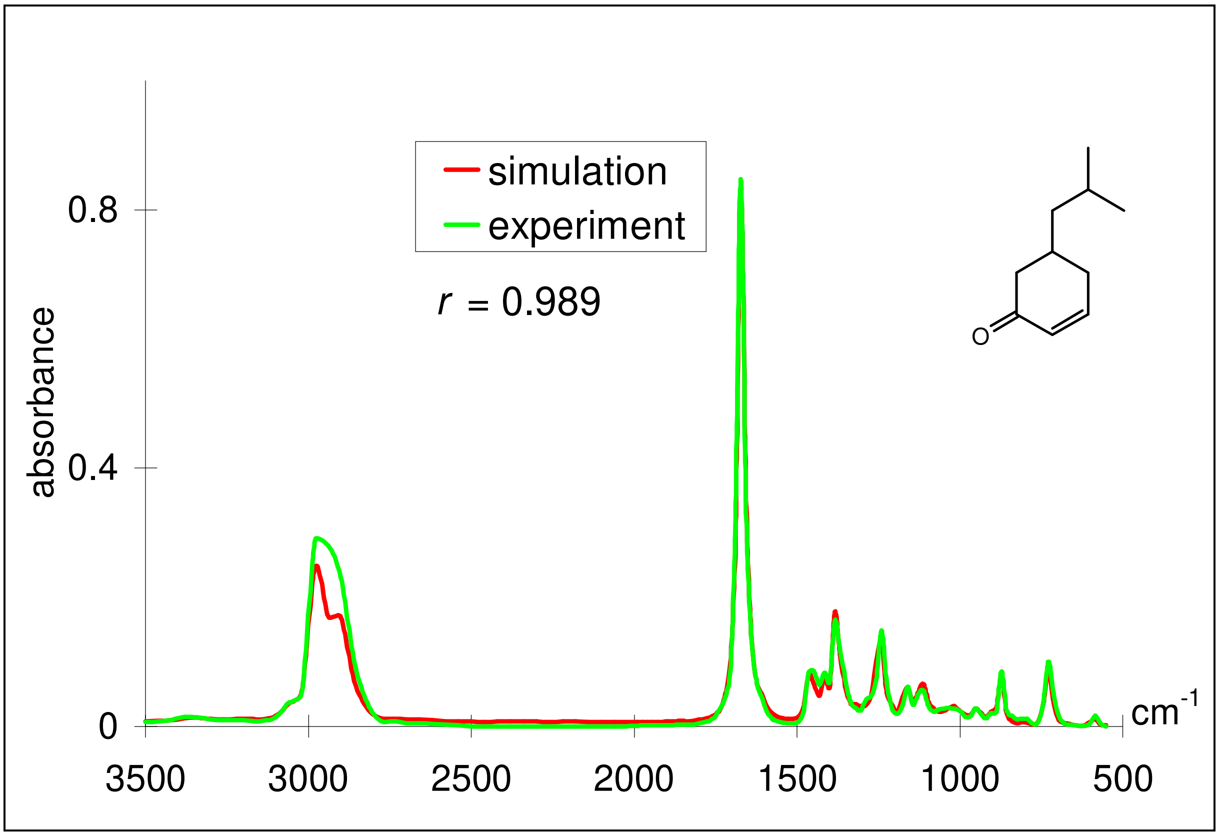


Abbildung 71: Mit Hilfe der anfrageorientierten Methodik simuliertes Infrarotspektrum von 5-(2-Metylpropyl)-cyclohex-2-enon und das experimentelle Spektrum des Moleküls.

6.8.2.3 Das Steroid Cholesterin

Steroide sind dank ihrer zahlreichen biologischen Wirkungen z.B. als Geschlechtshormone, von großem chemisch-pharmazeutischen Interesse. Allerdings macht die Molekülgröße Steroide für *ab initio* - Berechnungen zur Simulation von Infrarotspektren denkbar ungeeignet. Es ist deshalb von besonderem Interesse, ob auch für Moleküle dieser Größe mit Hilfe der anfrageorientierten Methodik Infrarotspektren simuliert werden können. Cholesterin, **TS**, dient als Testbeispiel für eine solche Simulation. Wie Abbildung 72 zeigt, gelingt diese Simulation recht gut. Im wesentlichen wird, mit etwas schwächerer Intensität, das experimentelle Infrarotspektrum wiedergegeben. Die Peaklagen können reproduziert werden.

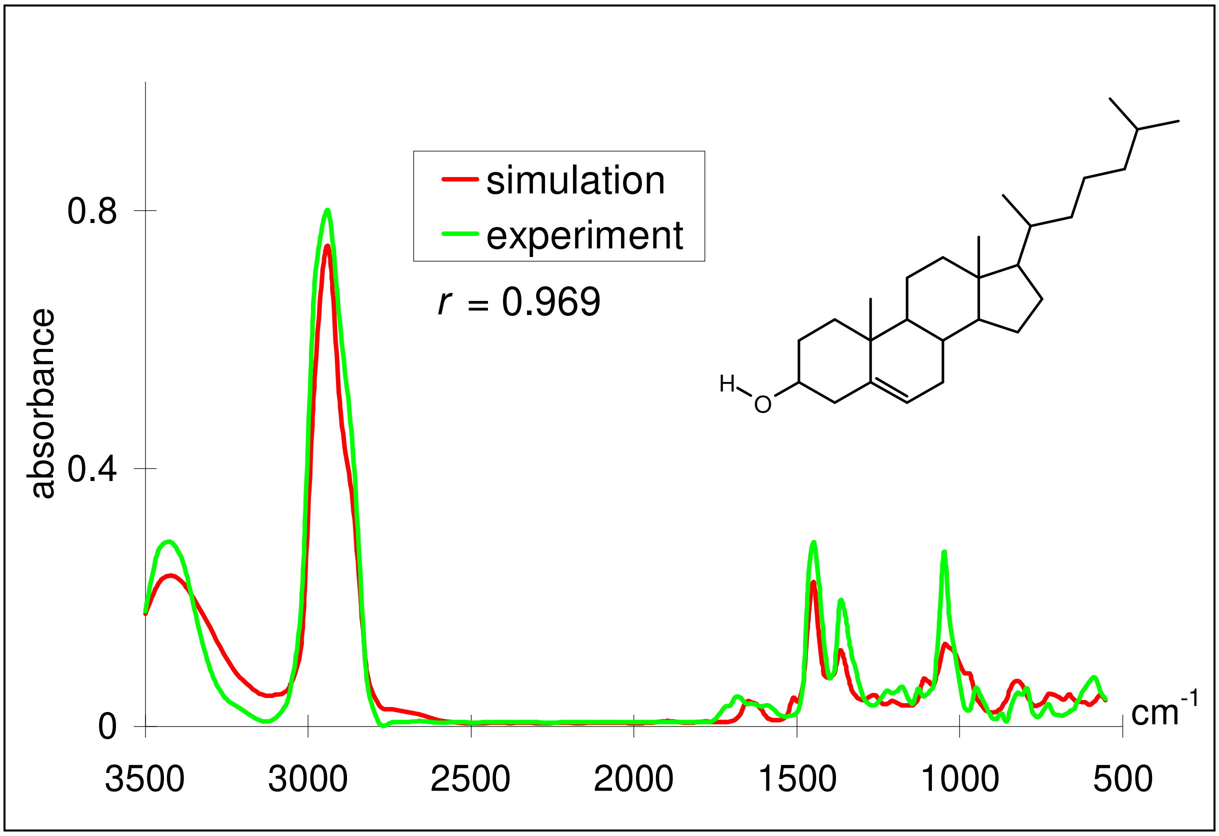


Abbildung 72: Simuliertes und experimentelles Infrarotspektrum des Steroidhormons Cholesterin.

Für die Simulation wurde das in Neuron (3,7) gespeicherte Infrarotspektrum genutzt. Neuron (3,7) waren im Erinnerungstest keine Trainingsmoleküle assoziiert worden. Vielmehr interpolierte das CPG-Netz das Simulationsspektrum für Cholesterin, **TS**, aus den Infrarotspektren der umliegenden Trainingsmoleküle **S1-S6**. Die Lage von **TS** und den Trainingsmolekülen im CPG-Netz zeigt Abbildung 73. Allen Trainingsmolekülen gemeinsam ist die Kombination einer Hydroxygruppe mit einem großen aliphatischen Teil. Daß die umliegenden Moleküle nicht sämtlich Steroide sind, liegt dabei sicherlich an der Tatsache, daß die Anzahl von Monohydroxysteroiden in der Datenbank beschränkt ist.

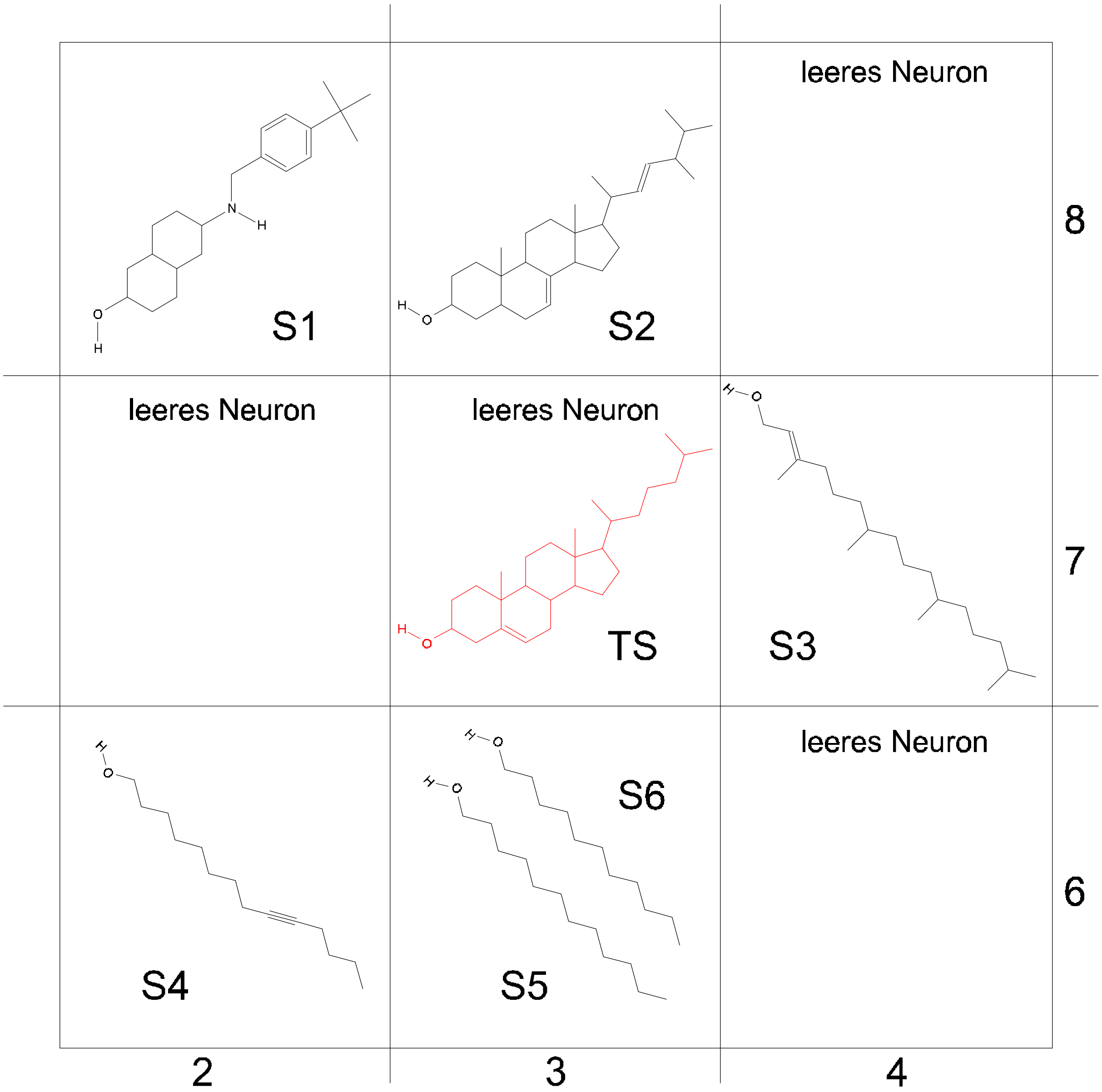


Abbildung 73: Interpolation des Simulationsspektrums von Cholesterin - die umliegenden Trainingsmoleküle deren Infrarotspektren für die Interpolation genutzt wurden.

6.9 Vorhersage des IR-Spektrums von primären Aminen unter besonderer Berücksichtigung der NH-Banden

Das Infrarotspektrum primärer Amine ist aus einem Grund besonders interessant - Gestalt und Lage der NH-Bande oberhalb von 3500 und 3070 cm^{-1} geben Aufschluß darüber, ob überhaupt, und wenn, zu welchem Prozentsatz, die H-Atome der Aminogruppen an Wasserstoffbrückenbindungen beteiligt sind.[45] Sind die Wasserstoffatome primärer Amine nicht an Wasserstoffbrückenbindungen beteiligt, so finden sich in der Regel im Infrarotspektrum zwei Banden zwischen 3500 und 3300 cm^{-1}.[81] Sind die H-Atome der Aminogruppen dagegen in Wasserstoffbrückenbindungen eingebunden, ist zumeist nur eine Bande bei etwa 3100 cm^{-1} zu beobachten. Neben diesen Reinformen für Wasserstoffbrückenbindungen, vorhanden oder nicht, kommen natürlich auch Mischformen vor, d.h. zum Teil sind drei Banden zu beobachten oder es finden sich Überlagerungen eines breiten Peaks, der von H-Atomen in Wasserstoffbrückenbindungen stammt, mit zwei schmaleren Banden, die von dem Anteil an freien Wasserstoffatomen in der Substanz stammen. Selbstverständlich hängt das resultierende Infrarotspektrum damit stark von Meßbedingungen ab, da beispielsweise in der Gasphase, insbesondere unter den Bedingungen der GC/IR-Kopplung, (intermolekulare) Wasserstoffbrückenbindungen selten sind, während sie in Kristallen, die möglicherweise noch Kristallwasser enthalten, am häufigsten sein dürften. So wird im Rahmen der Diskussion der Ergebnisse zu untersuchen sein, welche Ursachen zu den Abweichungen zwischen experimentellen und simulierten Infrarotspektren führen.

6.9.1 Die untersuchten Moleküle

Alle Moleküle wurden der SpecInfo-Datenbank entnommen. Die Auswahl der Strukturen wurde mittels einer Substruktursuche mit dem Programm *submtc*[82] nach der folgenden Substruktur im Datenbestand der SpecInfo-Datenbank vorgenommen. Die Definition von R mit Wasserstoff bzw. Kohlenstoff in ß-Position war notwendig, um z.B. Amide und Aminoacetale auszuschließen.

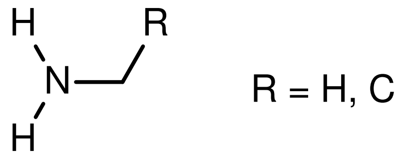


Abbildung 74: Substruktur mit der der Datensatz an primären Aminen ausgewählt wurde.

Die Suche ergab insgesamt 77 primäre Amine. Da außer der primären Aminfunktion keine weiteren Aussagen über das Molekül getroffen wurden, sind in der Auswahl eine Vielfalt von Kombinationen funktioneller Gruppen enthalten. Die Palette der gefundenen Verbindungen reicht von primären aliphatischen Aminen wie Pentylamin über 1,2-Diaminoethan und 3-Amino-2-hydroxypropionamid bis hin zu komplexen aromatischen Verbindungen mit 16 und mehr Kohlenstoffatomen.

6.9.2 Durchführung der Simulation

Alle Simulationen wurden nach der Methode der anfrageorientierten Simulation durchgeführt (s.a. Abschnitt 6.8.1). Dabei wurden die Standardwerte für die anfrageorientierte Simulation genutzt: 50 Verbindungen im Trainingsdatensatz, planares CPG-Netz mit 8 x 8 Neuronen, erste Lernrate 0.95 und maximale Korrekturentfernung 4 Neuronen bei automatischer Anpassung von Lernrate und Korrekturentfernung während des CPG-Netztrainings. Für eine der anfrageorientierten Simulationen wurden auf einer Sparc 10-40 unter SunOS 5.5.1 rund 8 min benötigt.

Für dieses Experiment wurden die dreidimensionalen Molekülstrukturen mit dem 3D-MoRSE Code unter den folgenden Randbedingungen codiert: 120 Werte pro Molekül, Atomeigenschaft $A_i = q_{tot,i}$ und Maximalwert des Beugungswinkelmaßes s_{max} = 30 Å^{-1}. Damit wurde die Anzahl der Werte n pro Molekül sowie der Wert für s_{max} gegenüber den vorausgegangenen Experimenten nahezu verdoppelt. Dies geschah, um die Auflösung des 3D-MoRSE Codes für kurze Distanzen zu erhöhen (s_{max} 15.5 -> 30 Å^{-1}) ohne dabei Informationen über größere Distanzen zu verlieren (Δs konstant, ≈ 0.25 Å^{-1}). So ergibt sich für diesen Versuch entsprechend Gleichung (12) ein Auflösungslimit für kurze Distanzen von 0.1 Å und für große Distanzen von 12 Å entsprechend Gleichung (13). Hiermit sollte gewährleistet werden, daß sowohl kleine Distanzänderungen in der Umgebung der Aminogruppe als auch große Distanzen zwischen weiter entfernten funktionellen Gruppen durch den Code adäquat berücksichtigt werden.

6.9.3 Ergebnisse

Eine Simulation der Infrarotspektren aller 77 primären Amine war möglich. Für alle 77 primären Amine wurden die Korrelationskoeffizienten zwischen experimentellem und simuliertem IR-Spektrum für das gesamte IR-Spektrum und für den Bereich der NH-Bande zwischen 3500 und 3060 cm^{-1} berechnet. Der mittlere Korrelationskoeffizient zwischen experimentellem und simuliertem Infrarotspektrum liegt für das gesamte Infrarotspektrum mit einem Wert von 0.766 höher als bei den substituierten Benzolderivaten. Der gegenüber dem Wert für das gesamte Infrarotspektrum gesteigerte mittlere Korrelationskoeffizient von 0.825 für den Bereich der NH-Bande zeigt, daß das Ziel dieses Versuchs, eine qualitativ hochwertige Simulation des Bereichs der N-H - Bande erreicht wurde, zumal der überwiegende Teil der schlechtesten Simulationen erklärbar ist. Abbildung 75 zeigt die Verteilung der Korrelationskoeffizienten für den Datensatz der 77 primären Amine, deren IR-Spektren mit der anfrageorientierten Methode simuliert wurden.

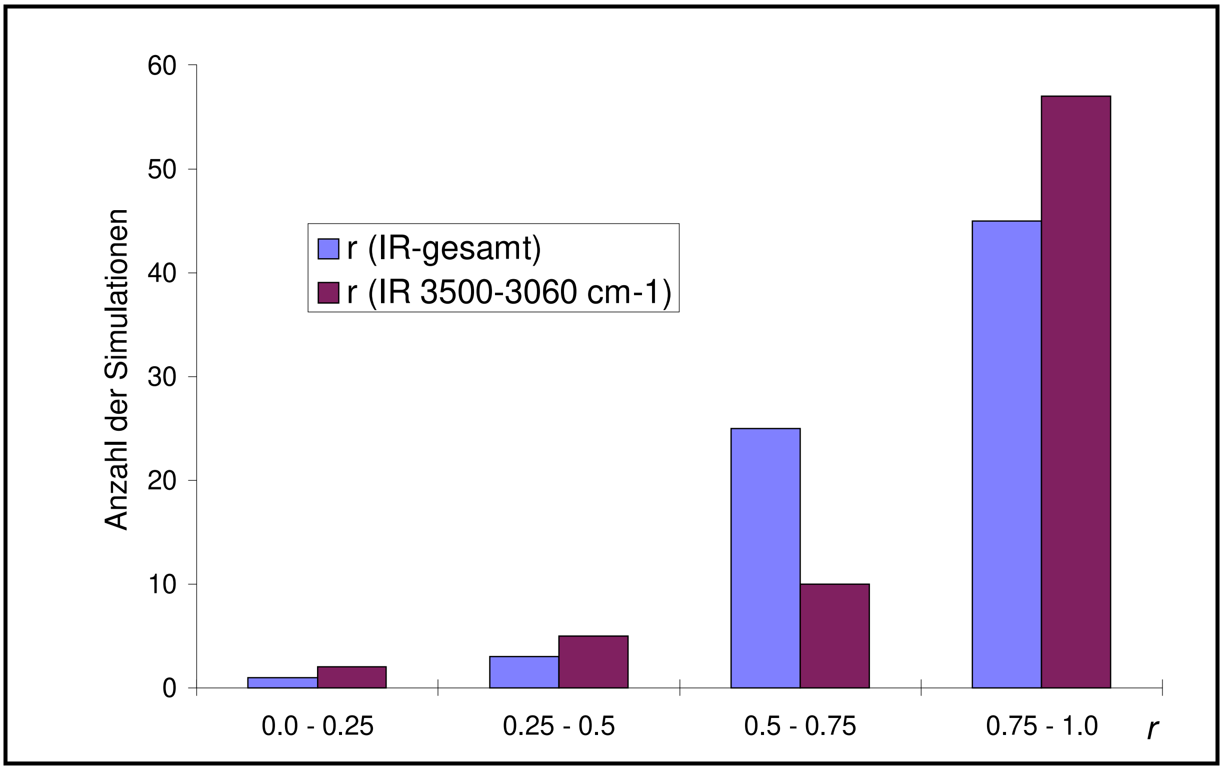


Abbildung 75: Verteilung der Korrelationskoeffizienten zwischen experimentellem und simuliertem Infrarotspektrum (hell: gesamtes Infrarotspektrum von 3500 - 552 cm^{-1}; dunkel: beschränkt auf den Bereich der Aminobanden zwischen 3500 und 3060 cm^{-1}) für die 77 primären Amine, deren Infrarotspektrum mit der anfrageorientierten Methode simuliert wurde.

Wie Abbildung 75 zeigt, konnte für knapp 75% (57 Amine, Bereich der Aminobanden) bzw. 60% (45 Amine, gesamtes Spektrum) der Amine ein Korrelationskoeffizient von über 0.75 zwischen experimentellem und simuliertem Spektrum erreicht werden. Das heißt bei nahezu drei Viertel der Spektren wurde der Bereich der NH-Bande einigermaßen korrekt wiedergegeben. Da, wie die folgende Analyse der Ergebnisse zeigt, die meisten schlechten Simulationen vorhersehbar schlecht waren, ist dies ein gutes Ergebnis.

6.9.4 Die Gründe für die Abweichungen der simulierten Spektren

Das hier zunächst die schlechtesten Simulationen diskutiert werden und nicht wie üblich die besten, liegt daran, daß hier dank der Methode erstmals die Ursachen für die schlechten Simulationen weitgehend aufgeklärt werden können. Wie sich zeigt, sind die Abweichungen zwischen simulierten und experimentellen IR-Spektren in den meisten Fällen nicht der Methode anzulasten.

Die nachstehende Simulation von 3-Aminopropanol ist mit Abstand die schlechteste Simulation benutzt man das Spektrum aus der SpecInfo-Datenbank als Referenz. Bei allen anderen Simulationen ist der Korrelationskoeffizient mit Werten von ca. 0.5 mindestens doppelt so hoch. Nachfolgend sollen alle dreizehn schlechtesten Simulationen besprochen werden, um die Quellen für die Abweichungen offenzulegen.

6.9.4.1 Die schlechteste Simulation: 3-Aminopropanol

3-Aminopropanol ist für eine Simulation schon aufgrund seiner Größe relativ ungeeignet, da jede Änderung an der Propankette bereits erhebliche Auswirkungen auf das IR-Spektrum haben dürfte. Hinzu kommt, daß in der SpecInfo-Datenbank kein wirklich ähnliches Molekül gefunden wurde und das Intensitätsverhältnis der NH- bzw. OH-Bande zwischen 3500 und 3100 cm^{-1} zur CH-Bande für ein aliphatisches Moleküle sehr ungewöhnlich ist, wie Abbildung 76 zeigt. Ein anders Vergleichspektrum, dessen NH- bzw. OH-Banden im übrigen sehr gut mit der Simulation übereinstimmen, läßt vermuten daß es sich hier eventuell um einen falschen Datenbankeintrag handelt. Eine, zugegeben einfache, semiempirische Berechnung des IR-Spektrums gibt mit einer ähnlichen Intensität für alle drei Banden einen weiteren Hinweis darauf, daß hier eventuell ein falsches Spektrum vorliegt. Die Berechnung des Infrarotspektrums erfolgte mit PM3/HyperChem[83]. Die niedrigste berechnete Frequenz bei 82.91 cm^{-1} belegt, daß ein Minimum vorliegt.

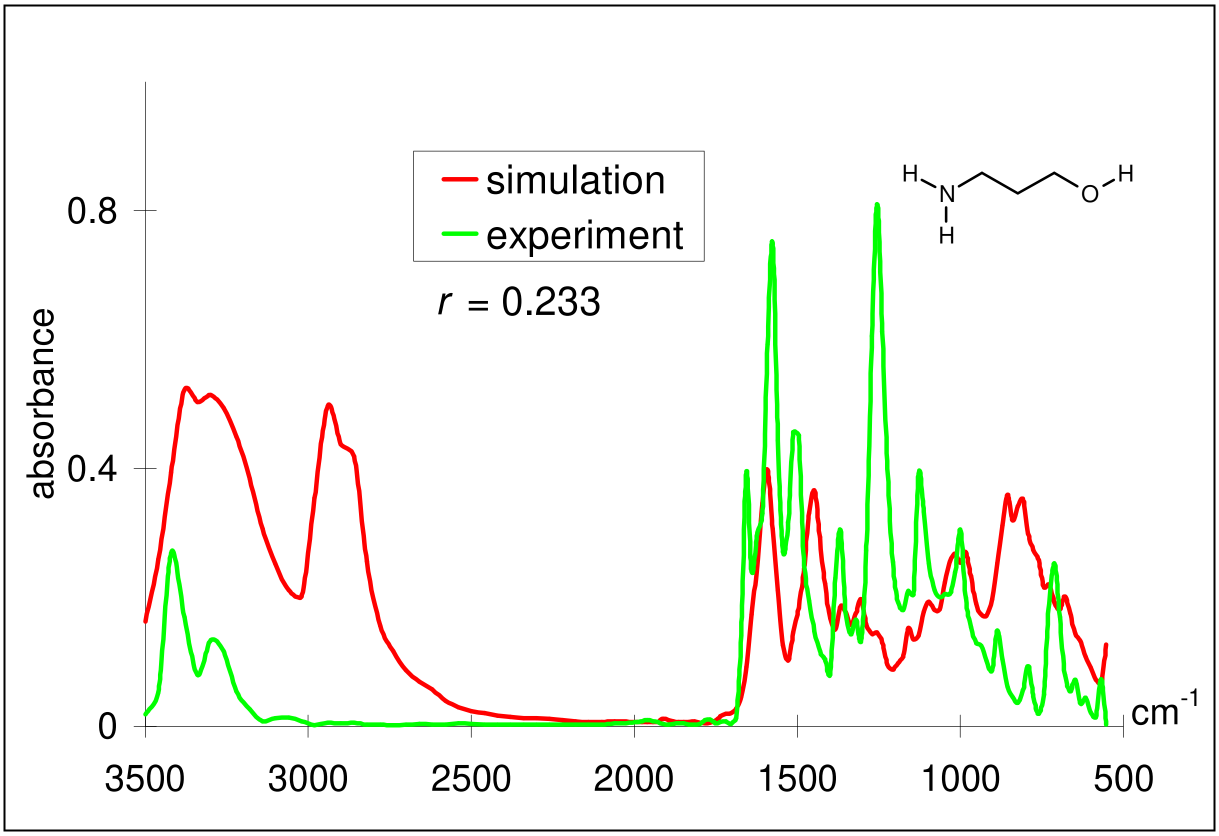


Abbildung 76: Die schlechteste Simulation des Datensatzes der primären Amine.

Abbildung 77 zeigt das simulierte Spektrum im Vergleich mit dem in Ref. [84] gefundenen Spektrum, daß für den Vergleich digitalisiert und dessen X-Achse von Microns in Wellenzahlen umgerechnet wurde. Deutlich ist die Übereinstimmung im Bereich zwischen 3500 und 1400 cm^{-1} zu sehen auch wenn die das experimentelle Spektrum aufgrund der fehlenden Basislinienkorrektur immer eine höhere Absorbanz aufweist. Die größte Abweichung ist eine starke schmale Bande bei 1100 cm^{-1}. Insgesamt sind die Abweichungen zwischen experimentellem und simuliertem Spektrum im Bereich unterhalb von 1400 cm^{-1} deutlicher als oberhalb. Ursache dürfte hier die geringe Größe von 3-Aminopropanol sein, die größere Abweichungen zwischen Test- und Trainingsstrukturen zur Folge haben muß.

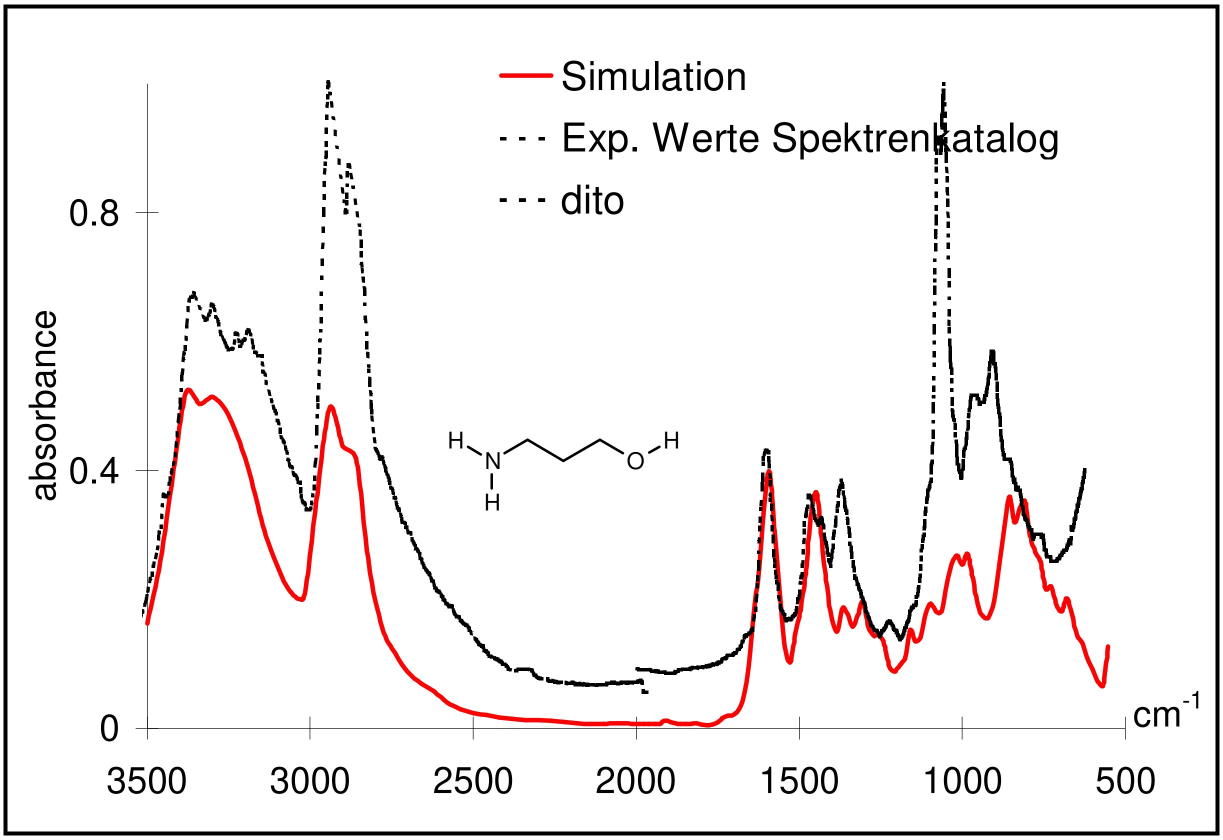


Abbildung 77: Simuliertes und experimentelles Spektrum von 3-Aminopropanol aus dem Spektrenkatalog (Ref. 84). Die Auflösung des experimentellen Spektrums ist schwankend, da es mit zwei verschiedenen Auflösungen und in Abhängigkeit von der Wellenlänge aufgezeichnet wurde, was bei abnehmenden Wellenzahlen zu einer größeren Auflösung führt.

Daß 3-Aminopropanol für eine gute Simulation zu klein ist, zeigen auch die umliegenden Trainingsmoleküle im trainierten CPG-Netz (vgl. Abbildung 78). Insgesamt 8 Moleküle befinden sich in der ersten Sphäre, sprich auf den direkt angrenzenden Neuronen des zur Simulation genutzten Neurons (6,7). Nur bei vier von diesen acht Molekülen gibt es überhaupt Sauerstoff und Stickstoffatome in einem Abstand von vier Bindungen, wie im Simulationsbeispiel. Nur zwei von diesen vier enthalten eine primäre Aminogruppe und die Hydroxyfunktion. Bei diesen zwei gehört jeweils eine der dazwischen liegenden Bindungen zu einem aromatischen Ring. Auf der anderen Seite führt die Simulation mit Diaminoderivaten anstelle von 3-Aminopropanol nicht zum gewünschten Ergebnis, da das Teilspektrum einer Aminogruppe nicht mit dem einer Hydroxygruppe vergleichbar ist.

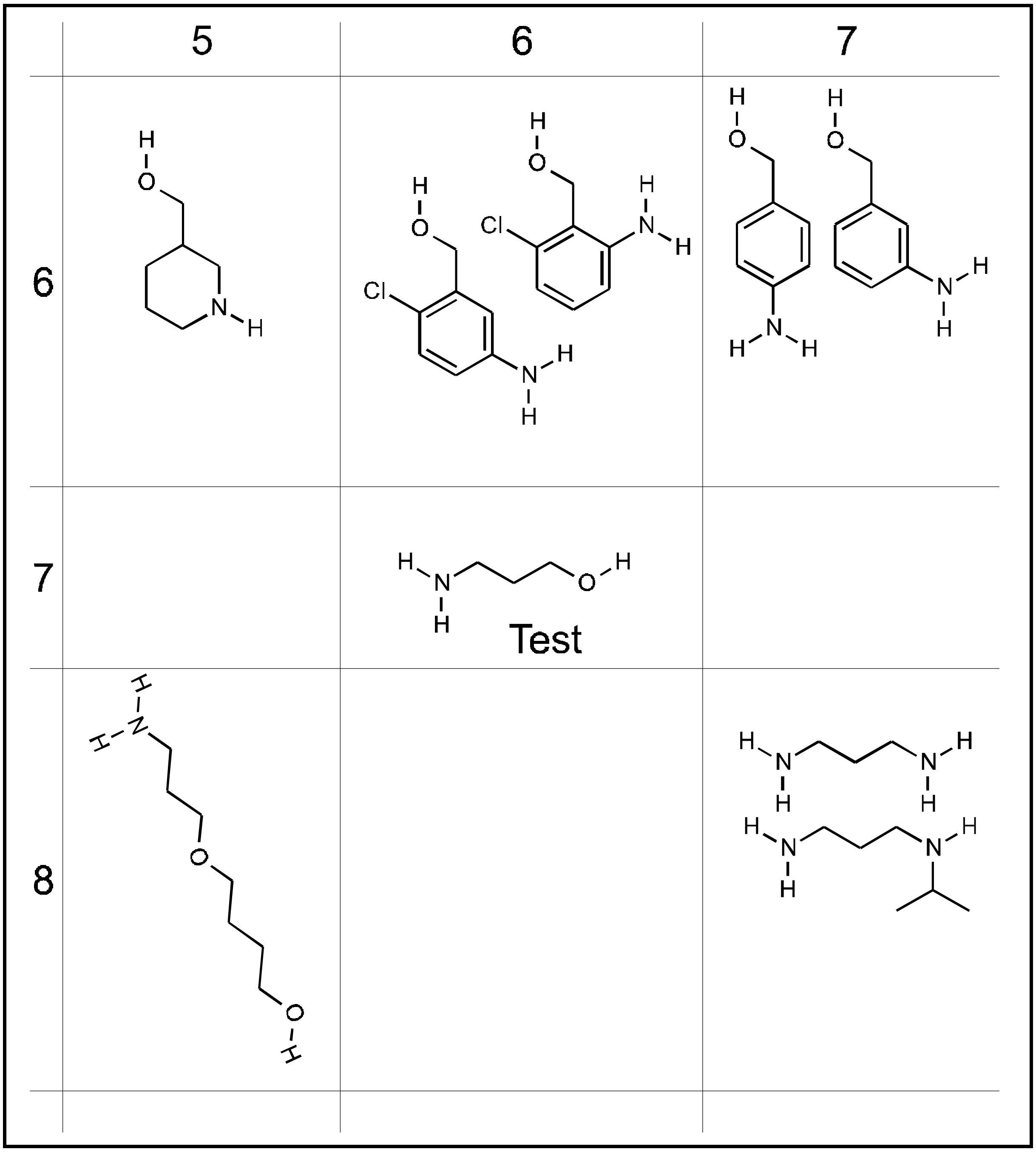


Abbildung 78: Die Lage von 3-Aminopropanol und der den umliegenden Neuronen assoziierten Trainingsmoleküle im CPG-Netz.

6.9.5 Fehler durch Lücken in der Datenbasis

Bei sechs bis sieben Verbindungen aus dem Datensatz der 77 primären Amine fehlten, wie Abbildung 79 zeigt, ähnliche Verbindungen in der Datenbank, so daß eine Simulation nicht erfolgversprechend möglich war. Das zeigt sich bei diesen Verbindungen auch am gewichteten Abstand zum ähnlichsten Neuron, AN_g (vgl. Gleichung 21), der hier überdurchschnittlich hoch ist. Der AN_g ist der auf den Betrag des 3D-MoRSE Codes des Moleküls normierte *rms*-Wert zwischen dem Strukturteil des ähnlichsten Neurons und dem 3D-MoRSE Code des Moleküls. Wesentlicher Schritt bei der Entwicklung des AN_g war die Normierung des *rms*-Wertes zwischen Struktur und Neuron auf den Betrag des 3D-MoRSE Codes. Da der Betrag des 3D-MoRSE Codes bei Verwendung von $A_i = q_{tot,i}$ wesentlich von der Polarität des Moleküls abhängt, würden sonst Unterschiede zwischen polaren Molekülen überbewertet bzw. Unterschiede zwischen unpolaren Molekülen unterbewertet werden.

Bei den sechs bis sieben primären Aminen, für die ähnliche Verbindungen zur Simulation in der Datenbank fehlten, lag der mittlere AN_g bei 3.13. Für die zehn besten Simulationen des Datensatzes aus 77 primären Aminen liegt der AN_g im Schnitt bei 1.56 und im Mittel für den Datensatz beträgt der AN_g 2.27.

$$AN_g = \frac{\sqrt{\frac{1}{n}\sum (M(s_i) - X_i)^2}}{\frac{1}{n}\sum M(s)} \tag{21}$$

AN_g	gewichteter Abstand des 3D-MoRSE Codes zum ähnlichsten Neuron
$M(s)$	3D-MoRSE Code der Verbindung
X	Eingabegewichte des ähnlichsten Neurons
n	Anzahl der 3D-MoRSE Codewerte bzw. der Eingabegewichte
i	Index

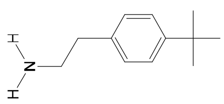
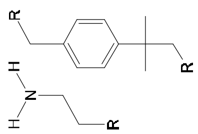
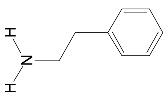
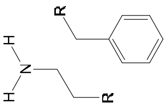
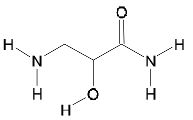
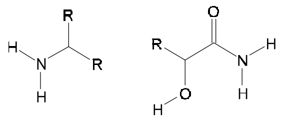
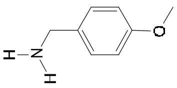
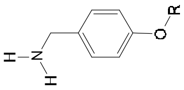
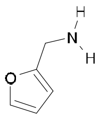
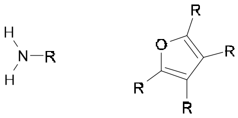
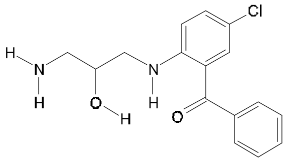
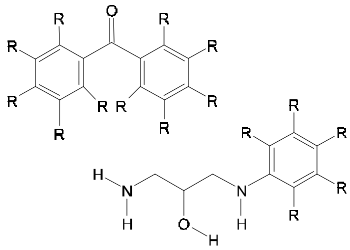
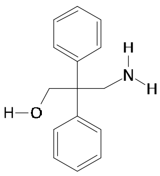
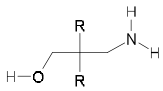
Verbindung	Substruktur/-en	Häufigkeit der Substruktur/-en in der Datenbank
		1
		3
		1
		1
		2
		1
		1

Abbildung 79: Die sieben Strukturen, der dreizehn schlechtesten Simulationen des Datensatzes aus den 77 primären Aminen der SpecInfo-Datenbank, die einen hohen AN_g-Wert aufweisen zusammen mit der Substruktur(-en) und den Ergebnissen der Substruktursuche nach ähnlichen Molekülen in der SpecInfo-Datenbank. R steht hierbei für einen beliebigen Rest.

Die auf der Bindungsliste basierte Substruktursuche fand mit den angegebenen Substrukturen in fünf der sieben Fälle nur das Ausgangsmolekül. Lediglich bei *C*-Furan-2-ylmethylamin und 2-Phenylethylamin wurde die Substruktursuche fündig. Bei *C*-Furan-2-ylmethylamin war die zweite gefundene Verbindung (5-Aminomethylfuran-2-yl)-methanol, die wegen der Hydroxygruppe weder bei der Simulation eine Rolle spielte, noch ein ähnliches Spektrum haben kann. So beträgt der Korrelationskoeffizient zwischen den experimentellen Infrarotspektren von (5-Aminomethylfuran-2-yl)-methanol und *C*-Furan-2-ylmethylamin 0.625 und ist damit nur wenig besser als der von der Simulation erzielte Wert. Damit steht fest, daß zu den fünf Verbindungen bei den die Substruktursuche nur das Ausgangsmolekül fand und zu *C*-Furan-2-ylmethylamin in der SpecInfo-Datenbank kein Molekül mit mutmaßlich ähnlichen Infrarotspektren existiert.

Bei 2-Phenylethylamin sieht es etwas anders aus. Neben 2-(3-Phenylmethoxy-4-methoxyphenyl)-ethylamin findet die Substruktursuche auch 3-Phenylpropylamin, das 2-Phenylethylamin strukturell ähnlich ist und auch ein ähnliches IR-Spektrum hat. Damit stellt sich die Frage, warum die Simulation des IR-Spektrums von 2-Phenylethylamin mißlang. Eine Analyse des zur Simulation verwendeten Netzes liefert eine schnelle, wenn auch überraschende Antwort. Statt des Spektrums von 3-Phenylpropylamin wurde vom neuronalen Netz das Spektrum von 2-(4-(1,1-Dimethylethyl)-phenyl)-ethylamin zur Simulation des Infrarotspektrums von 2-Phenylethylamin genutzt, weil der *rms*-Wert zwischen den 3D-MoRSE Codes von 2-Phenylethylamin und 2-(4-(1,1-Dimethylethyl)-phenyl)-ethylamin nur 0.075 beträgt gegenüber 0.105 zwischen 2-Phenylethylamin und 3-Phenylpropylamin. Die *rms*-Werte geben wiederum einen Hinweis auf die Notwendigkeit zur Verbesserung der 3D-Strukturcodierung gerade auch für Aromaten, die reine Kohlenwasserstoffsubstituenten tragen. Allerdings war dies nicht unbedingt Ziel dieses Versuchs, denn es ging um die Simulation der NH-Banden zwischen 3500 und 3060 cm^{-1} und diese wurden auch in diesem Beispiel korrekt simuliert, wie Abbildung 80 zeigt.

Nutzt man den Radialcode, wird im übrigen auch das Infrarotspektrum von 2-(4-(1,1-Dimethylethyl)-phenyl)-ethylamin zur Simulation genutzt, was zu einem identischen Simulationsergebnis führt, wie sich bei einem Versuch mit der auf dem Radialcode basierten Simulationsmethode der TeleSpec-IR-Projekts herausstellte.[85] In beiden Fällen ist 3-Phenylpropylamin nicht einmal einem Neuron in der unmittelbaren Nachbarschaft des zur Simulation genutzten

Neurons assoziiert worden, dessen Werte in beiden Fällen mit denen von 2-(4-(1,1-Dimethylethyl)-phenyl)-ethylamin identisch sind.

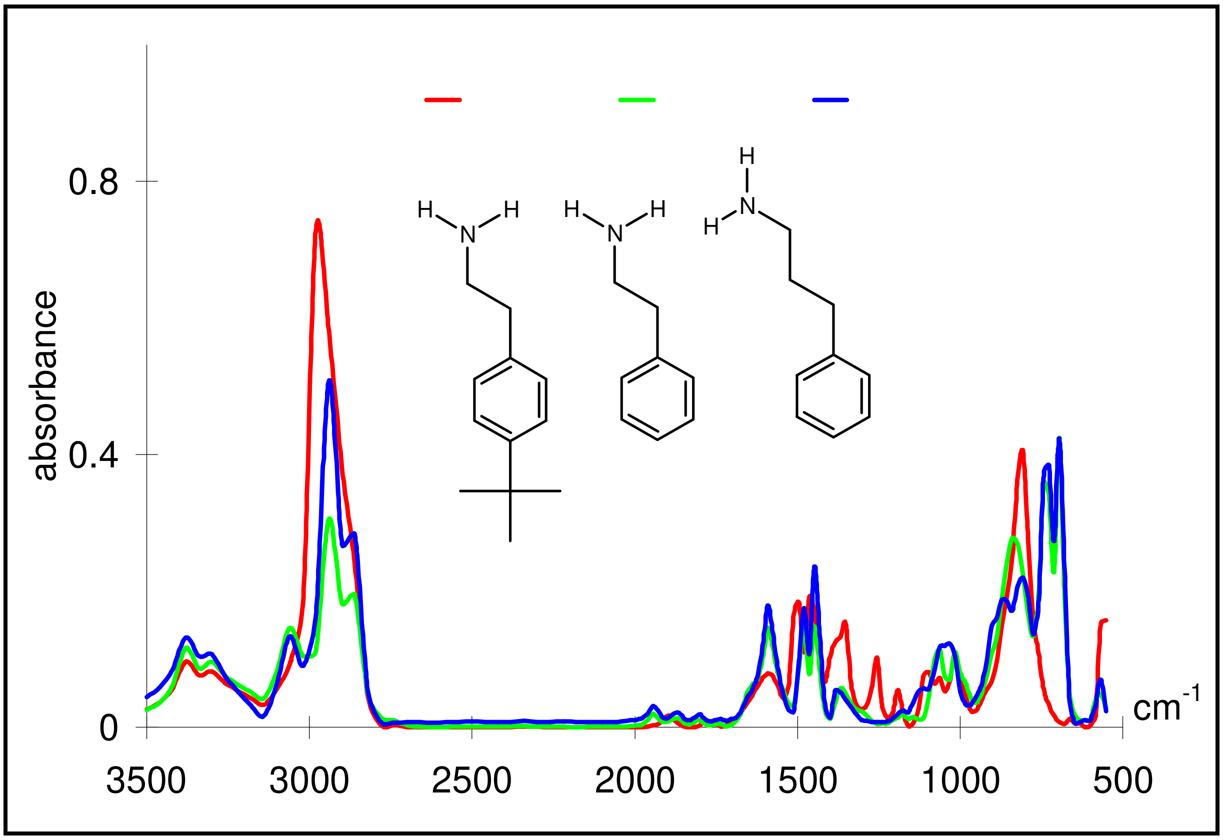


Abbildung 80: Die experimentellen Infrarotspektren von 2-(4-(1,1-Dimethylethyl)-phenyl)-ethylamin, 2-Phenylethylamin und 3-Phenylpropylamin. Deutlich wird die Übereinstimmung der NH-Banden bei 3300 cm^{-1} aber auch die Abweichung des Fingerprintbereichs von 2-(4-(1,1-Dimethylethyl)-phenyl)-ethylamin von den Spektren der beiden anderen Verbindungen.

6.9.6 Abweichungen durch andere Meßbedingungen

Die größten Abweichungen zwischen simuliertem und experimentellem Spektrum durch unterschiedliche Meßbedingungen zwischen Test- und Trainingsmolekül(-en) gibt es bei der Simulation von N^1-(8-Aminooctyl)-octan-1,8-diamin, wo nur ein Korrelationskoeffizient von 0.480 erreicht wird. Abbildung 81 zeigt das experimentelle in KBr gemessene Spektrum von N^1-(8-Aminooctyl)-octan-1,8-diamin und das zur Simulation genutzte experimentelle Spektrum von N^1-(3-Aminopropyl)-butan-1,4-diamin, das als Filmspektrum aufgenommen wurde. Deutlich zu sehen ist die viel breitere und intensivere Amino-Bande des in KBr gemessenen Spektrums von N^1-(8-Aminooctyl)-octan-1,8-diamin, die gegenüber dem Filmspektrum auch weniger strukturiert ist.

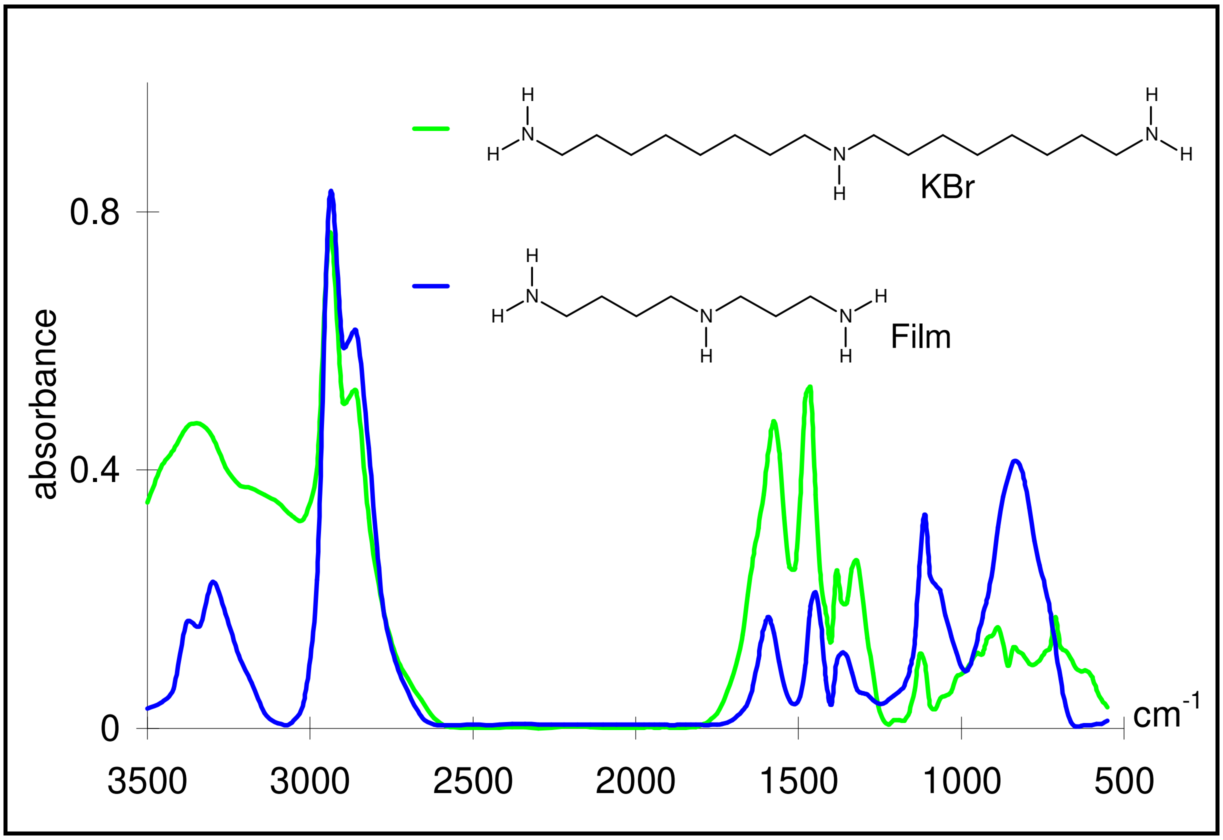


Abbildung 81: Die experimentellen Spektren von *N*[1]-(8-Aminooctyl)-octan-1,8-diamin und *N*[1]-(3-Aminopropyl)-butan-1,4-diamin aus der SpecInfo-Datenbank.

Abbildung 81 zeigt den Effekt des Übergangs vom KBr-Spektrum zum Film-Spektrum, der bei der Simulation des Spektrums von *N*[1]-(8-Amino-octyl)-octan-1,8-diamin zu einer großen Abweichung führte. So ist eine hohe und breite Aminobande, die schon oberhalb von 3500 cm^{-1} beginnt, für die in KBr gemessenen Infrarotspektren primärer Amine typisch. Typischerweise beträgt die Intensität Aminobande von in KBr gemessenen Infrarotspektren mehr als die Hälfte der Intensität der CH-Bande bei 3000 cm^{-1}, bei Filmspektren ist es signifikant weniger meistens unter 30 % der Intensität der CH-Bande. Ebenfalls fällt bei KBr-Spektren die Intensität der Banden zwischen 1600 und 1400 cm^{-1} meistens im Vergleich zur CH-Bande bei 3000 cm^{-1} höher aus.

Für die Simulation der Spektren von n-Pentylamin und n-Octylamin wird in beiden Fällen das Spektrum von n-Hexylamin aus dem jeweiligen Trainingsdatensatz zur Simulation genutzt. Dieses weicht aber, wie die folgende Abbildung 82 zeigt, erheblich von den Infrarotspektren der anderen n-Alkylamine der SpecInfo-Datenbank ab.

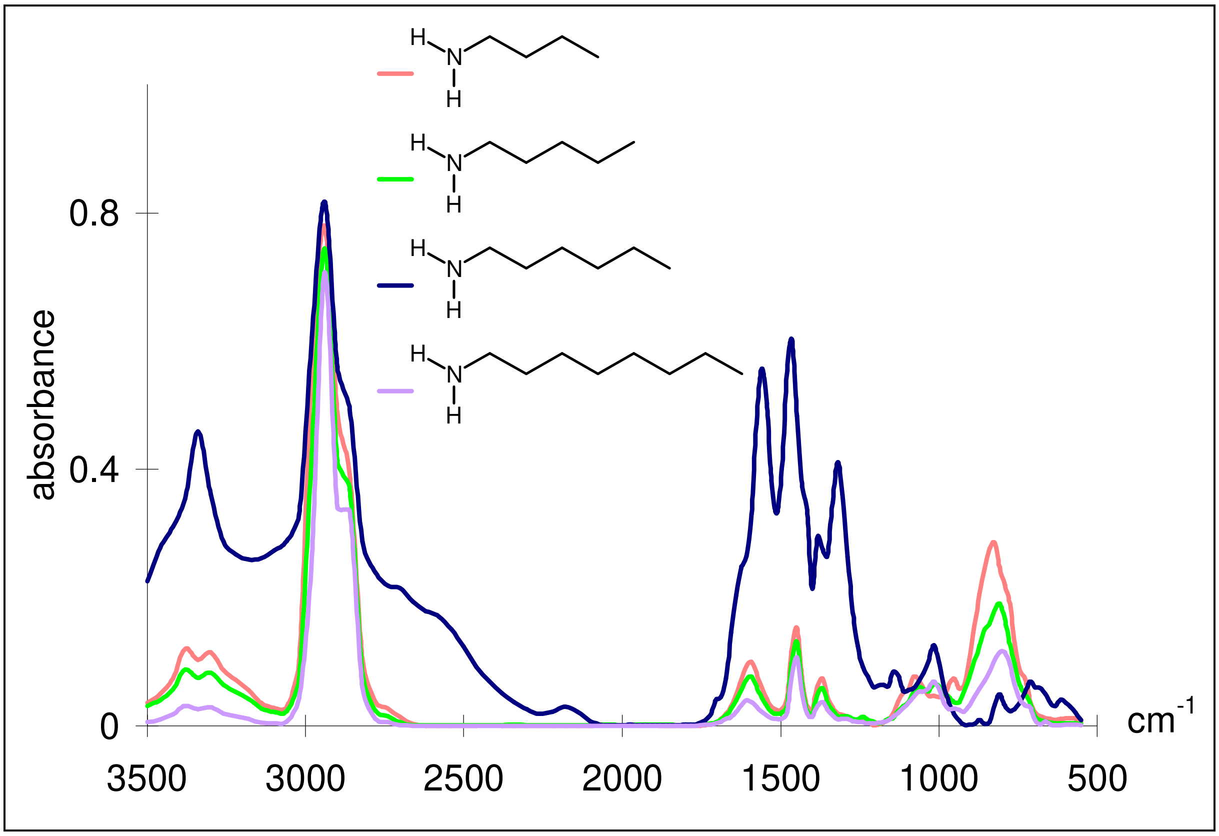


Abbildung 82: Spektren der vier n-Alkylamine der SpecInfo-Datenbank. Nur allzu deutlich fällt das abweichende Spektrum von n-Hexylamin (schwarz) in dieser homologen Reihe auf. Dieses wird jedoch aufgrund der ähnlichen 3D-Strukturen bei den benachbarten Molekülen in der homologen Reihe zur Simulation benutzt, was zu erheblichen Abweichungen zwischen den experimentellen und simulierten IR-Spektren bei diesen zwei Verbindungen führt.

Das Infrarotspektrum von n-Hexylamin wurde als Filmspektrum gemessen während die Spektren n-Butyl- und n-Octylamin in einer Kuevette vermessen wurden.[10] Für Pentylamin enthält die SpecInfo-Datenbank keine Angaben zum Medium, da das Spektrum aber perfekt in die homologe Reihe paßt, liegt der Verdacht nahe, daß es sich ebenfalls um ein Kuevettenspektrum handelt. Ferner könnte das Filmspektrum von n-Hexylamin durch Säurereste verunreinigt sein. Die Breite der Absorption bis hinunter zu 2100 cm^{-1} könnte auf eine derartige Verunreinigung hinweisen.

Ein ähnliches Phänomen findet sich bei der Simulation von 1,2-Diaminoethan. Auch hier ist die Aminobande im experimentellem Spektrum, das als Filmspektrum gemessen wurde, ungewöhnlich breit. Die Spektren der Trainingsmoleküle auf den umliegenden Neuronen sind aber zumeist als Küvettenspektren gemessen worden, wie zum Beispiel das Spektrum von N^1-(2-Aminoethyl)-ethan-1,2-diamin.

Die Küvettenspektren zeichnen sich durch eine deutlich schwächere NH-Bande bei 3300 cm^{-1} aus, während Filmspektren deutlich stärkere und breitere NH-Banden zeigen, wie die experimentellen Spektren von Hexylamin (Abbildung 82), 1,2-Diaminoethan und von N^1-(2-Aminoethyl)-propan-1,3-diamin (Abbildung 83) belegen. Das Meßmedium für das Spektrum von 1,2-Diaminopropan war leider nicht in der SpecInfo-Datenbank enthalten, jedoch erinnert das Spektrum sehr an ein Küvettenspektrum.

Abbildung 83 zeigt aber auch, daß 1,2-Diaminoethan für eine Simulation zu klein ist, da jedes hinzugefügte Atom eine deutliche Änderung im Infrarotspektrum zur Folge hat, wie die gemeinsamen Bande der Trainingsmoleküle bei 840 cm^{-1} zeigt, die im experimentellem Spektrum von 1,2-Diaminoethan fehlt.

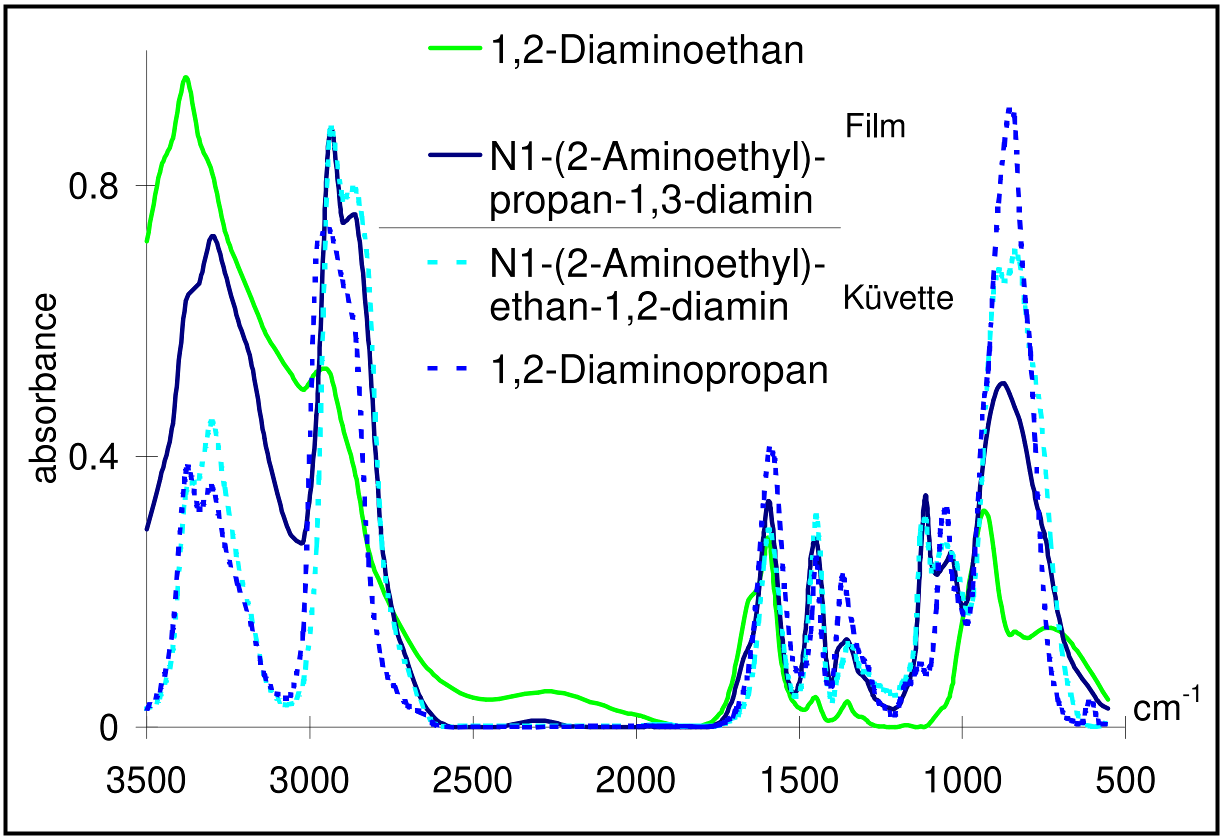


Abbildung 83: Die experimentellen Spektren von 1,2-Diaminoethan und von drei Trainingsmolekülen die benachbarten Neuronen assoziiert wurden.

Berechnet man die Infrarotspektren von 1,2-Diaminoethan und den Trainingsmolekülen 1,2-Diaminopropan und N^1-(2-Aminoethyl)-ethan-1,2-diamin, so ergibt sich mit PM3/HyperChem[83] für 1,2-Diaminoethan eine charakteristische Deformationsschwingung bei 920 cm^{-1}, in der das Molekül um die zentrale CC-Bindung rotiert. Bei den beiden Trainingsmolekülen finden sich an dieser Stelle zwei Schwingungen bei 1010 cm^{-1} bzw. bei 870 cm^{-1} mit ähnlicher Intensität und Bewegung der Atome, insbesondere die letztere Schwingung bei 870 cm^{-1} könnte der starken experimentellen Absorption 840 cm^{-1} entsprechen, die allen Trainingsmolekülen gemeinsam ist, aber im Spektrum von 1,2-Diaminoethan fehlt. Dies zeigt warum 1,2-Diaminoethan für die Simulation eigentlich zu klein ist, denn jegliches Hinzufügen von Atomen führt zur Verschiebung bzw. Aufspaltung der prominenten Bande bei 920 cm^{-1} (experimentell und berechnet) auf 840 cm^{-1} (870 cm^{-1} berechnet) bzw. auf 1010 cm^{-1}. Die folgende Tabelle illustriert dieses Problem noch einmal.

Molekül	Wellenz.	I	Bewegung der Atome
1,2-Diamino-ethan	920.57	8.314	Rotation um zentrale C-C - Bindung, Deformation N-H - Bindungen
	958.7	0.588	C-C Streckschwingung mit Deformation der Winkel an den Kohlenstoffatomen
	1018.02	1.213	Twistschwingung der C-C - Bindung
1,2-Diamino-propan	863.42	2.27	Rotation um zentrale C-C - Bindung
	905.48	0.418	C-C Streckschwingung der zentralen Bindung mit Deformation der Winkel an den Kohlenstoffatomen
	960.52	0.387	C-C und C-N Streckschwingung der zentralen C-C - Bindung und der C2-N - Bindung
	1006.78	2.95	C1-C2 Twistschwingung kombiniert mit eine N-H Streckschwingung
	1031.96	1.049	C2-C3 Streckschwingung
N1-(2-Aminoethyl)-ethan-1,2-diamin	866.55	9.575	Rotationsschwingung der C-C Bindungen und N-H Streckschwingung des sekundären Amins
	935.17	3.179	Rotationsschwingung der C-C Bindungen mit Deformation der N-H Bindungen
	980.97	3.024	C-N Twistschwingungen
	1009.01	10.06	C-C Streckschwinungen kombiniert mit N-H Streckschwingungen
	1024.21	0.923	C-C Twistschwingungen
	1039.26	1.597	C-C Streckschwingungen

Tabelle 9: Mit PM3[83] berechnete Schwingungen von 1,2-Diaminoethan und den Trainingsmolekülen 1,2-Diaminopropan und N^1-(2-Aminoethyl)-ethan-1,2-diamin im Bereich von 800 - 1050 cm^{-1}.

6.9.7 Vorbedingungen für gute Simulationen der Aminobanden

Die Voraussetzungen für gute Simulationen können allgemein wie folgt aus den bisher diskutierten Ergebnissen abgeleitet werden:

1. In der Datenbank muß mindestens eine Verbindung enthalten sein, deren Struktur mit der Struktur der Testverbindung in allen IR-spektroskopisch relevanten Strukturteilen ähnlich bzw. identisch ist. Anstelle einer Struktur kann auch zwischen verschiedenen eng verwandten Strukturen interpoliert werden.

2. Die Meßverfahren der Testverbindung und der zur Simulation benutzten Trainingsverbindungen müssen vergleichbare Ergebnisse liefern.

3. Beide vorstehenden Bedingungen müssen für die ähnlichste Verbindung bzw. ein Großteil der ähnlichen Verbindungen zutreffen.

Um die praktische Bedeutung dieser Voraussetzungen zu illustrieren und zu zeigen, wie sie das hier gewählte Simulationsverfahren umsetzt, werden nachfolgend einige der besten Simulationen dieses Versuchs analysiert.

6.9.7.1 Die beste Simulation des Aminobereichs

Mit einem identischen Korrelationskoeffizienten von 0.999 für den Bereich der Aminobanden zwischen 3500 und 3060 cm^{-1} zeigten die Simulationen von 4-Aminobutansäure und 6-Aminohexansäure die geringsten Abweichungen zwischen experimentellen und simuliertem Spektrum. Die identischen Korrelationskoeffizienten ergeben sich aus der wechselseitigen Nutzung des IR-Spektrums von 6-Aminohexansäure zur Simulation des Spektrums von 4-Aminobutansäure und umgekehrt. Bedauerlich im aktuellen Zusammenhang ist nur, daß im IR-Spektrum die typischen NH-Banden von der OH-Bande der Carbonsäure verdeckt werden, wie Abbildung 84 zeigt.

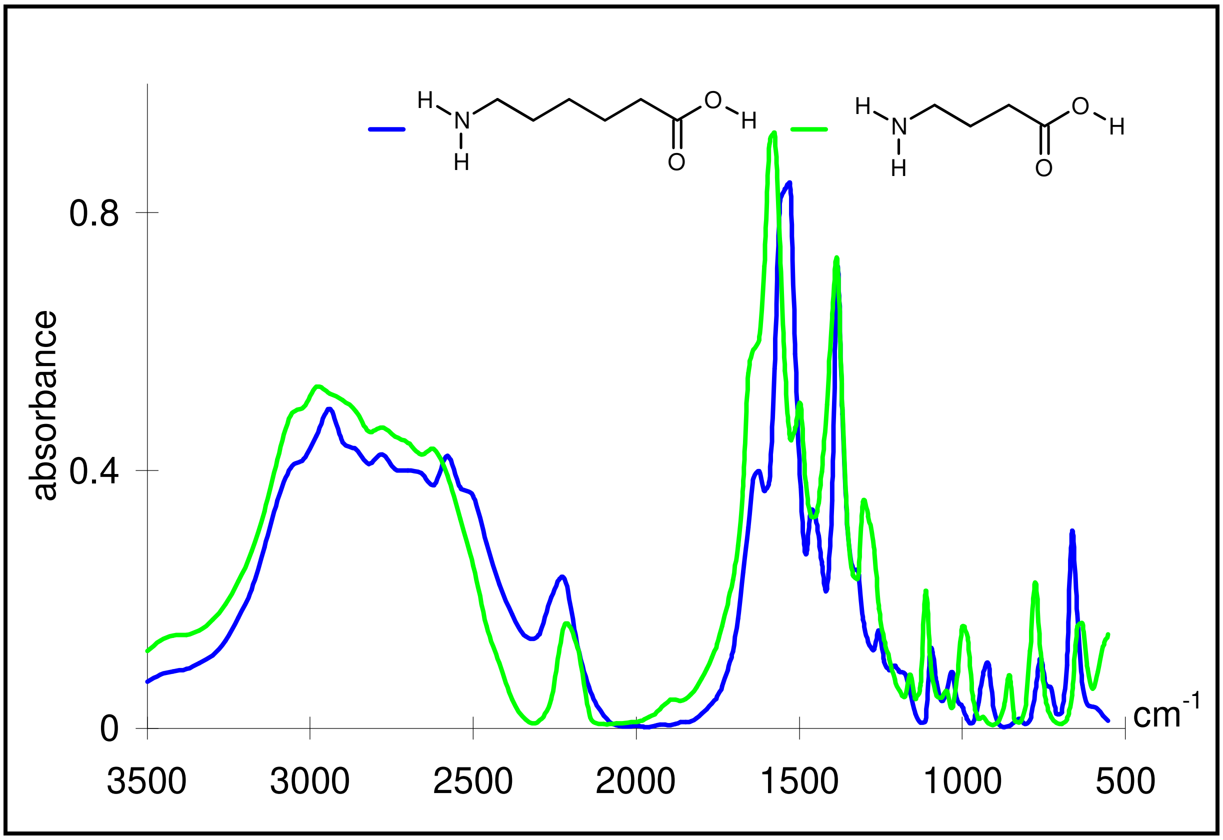


Abbildung 84: Experimentelle IR-Spektren von 6-Aminohexansäure und 4-Aminobutansäure, die wechselseitig zur Simulation des jeweils anderen Spektrums genutzt wurden und damit in der Simulation die höchsten Korrelationskoeffizienten für den Bereich NH-Banden ergaben.

6.9.8 Die beste Simulation des Gesamtspektrums und die drittbeste Simulation des Aminobereichs

Die Simulation für 1,5-Diamino-2,2-dimethylpentan mit einem Korrelationskoeffizienten von 0.984 zwischen experimentellem und simuliertem Infrarotspektrum für das gesamte Spektrum und einem Korrelationskoeffizienten von 0.997 für den Bereich der NH-Banden oberhalb von 3060 cm^{-1} ist die beste Simulation des Gesamtspektrums und die drittbeste Simulation des Aminobereichs des Datensatzes der 77 primären Amine. Abbildung 85 zeigt das simulierte und das experimentelle Spektrum von 1,6-Diamino-2,2,4-trimethylhexan, das vom Neuron (7,5) des neuronalen Counterpropagation-Netzes stammt. Die Werte des Neurons (7,5) sind mit den Werten des Trainingsmoleküls 1,6-Diamino-2,2,4-trimethylhexan identisch (Korrelationskoeffizient IR-Spektrum 1,6-Diamino-2,2,4-trimethyl-hexan - Ausgabegewichte Neuron (7,5) ist gleich 1.000 und der *rms*-Wert Eingabegewichte Neuron (7,5) - 3D-MoRSE-Code 1,6-Diamino-2,2,4-trimethylhexan ist kleiner $1*10^{-6}$). Insofern kann man sagen, daß das

Spektrum von 1,6-Diamino-2,2,4-trimethylhexan wurde zur Simulation des Spektrums von 1,5-Diamino-2,2-dimethylpentan benutzt.

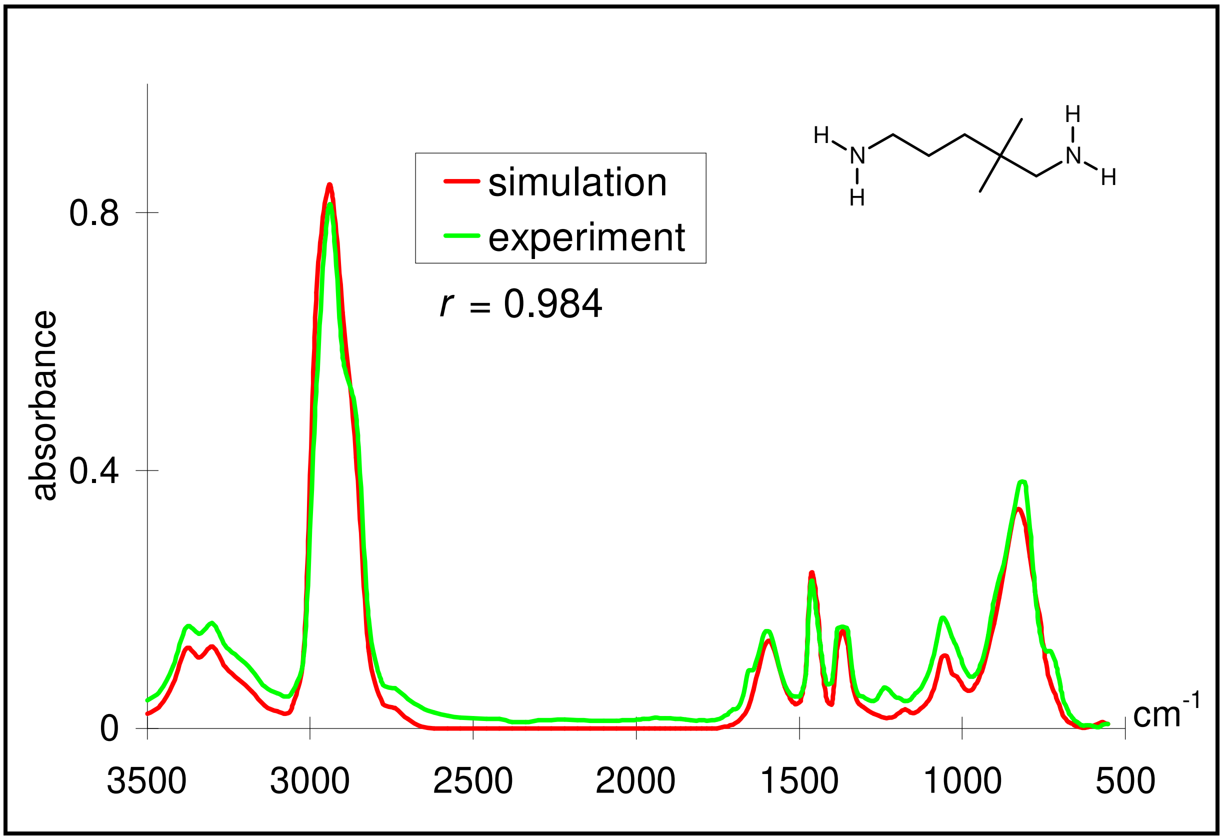


Abbildung 85: Das experimentelle und das simulierte Spektrum von 1,5-Diamino-2,2-dimethylpentan, eine der besten Simulationen des Aminodatensatzes.

Die fast perfekte Simulation des Infrarot-Spektrums von 1,5-Diamino-2,2-dimethylpentan ist insofern typisch für eine gute Simulation, als daß die funktionellen Gruppen der Test- und der Trainingsverbindung einschließlich der Nachbarschaft übereinstimmen: endständiges primäres Amin an einer aliphatischen Kette, primäres Amin mit einem quatärnären Kohlenstoff in β-Stellung. Der Unterschied in der Kettenlänge (die Testverbindung enthält eine n-Pentankette, die Trainingsverbindung eine n-Hexankette) wirkt sich bei gleicher Umgebung der funktionellen Gruppen nicht im Infrarot-Spektrum aus. Auch der zusätzliche Methylsubstituent in 4-Position des Trainingsmoleküls ist nicht IR-spektroskopisch wirksam. Dies ist insofern typisch, da es in beiden Verbindungen zwei weitere Methylgruppen gibt, und in direkter Nachbarschaft (hier vielleicht eine CC-Einfachbindung zum Substitutionsort) dieses zusätzlichen Methylsubstituenten keine polaren funktionellen Gruppen vorhanden sind.

Ein weiteres Beispiel für eine solche Simulation, basierend auf einer Verbindung, die sich nur durch eine CH_2-Einheit unterscheidet, ist die Simulation von 3-Ethoxypropanamin. Die Verbindung, deren Spektrum in diesem Fall zur Simulation genutzt wird, ist 3-Methoxypropanamin. Deutlich wird der Unterschied zwischen Ethoxy- und Methoxy-Gruppe im Spektrum nur in einer geringen Intensitätsabweichung der CH-Valenzschwingungsbande zwischen 3000 und 2800 cm^{-1} sowie in der Intensitätsverschiebung einer Bande bei 1368 cm^{-1}.

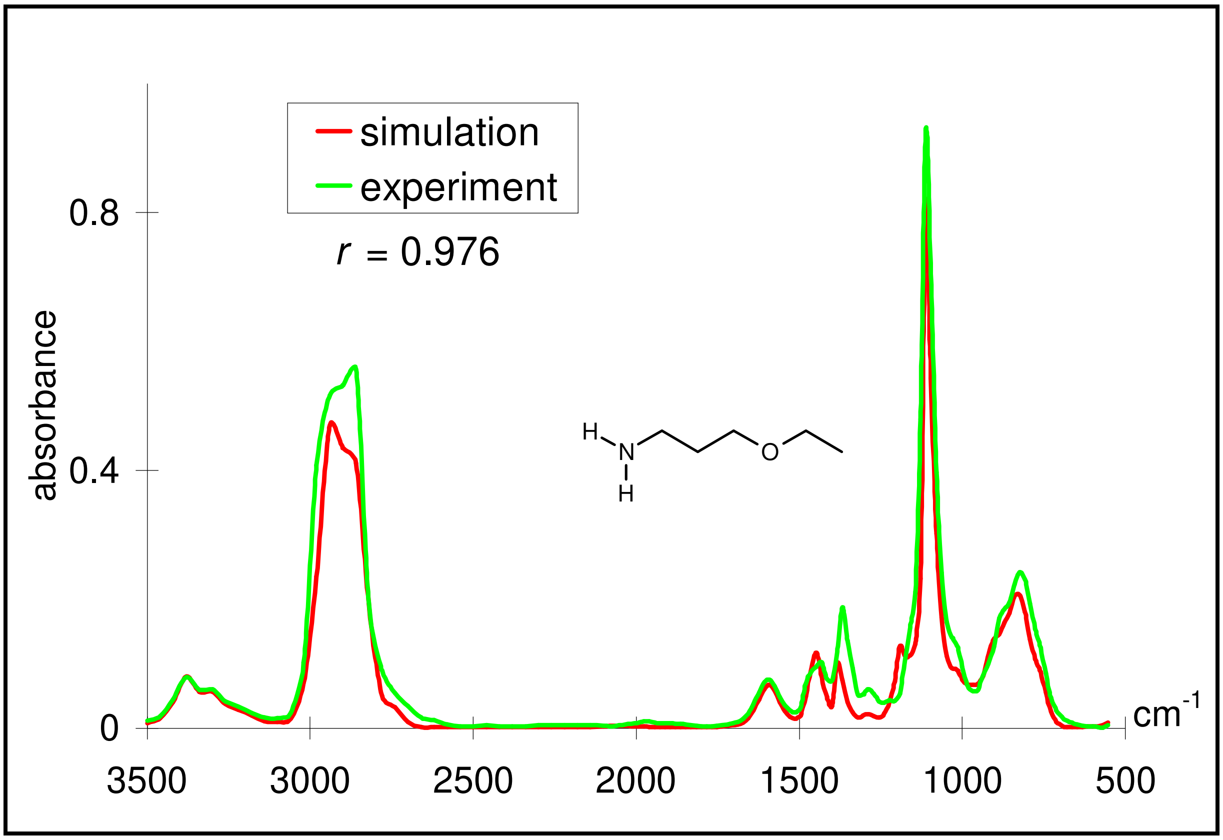


Abbildung 86: Experimentelles und simuliertes Infrarotspektrum von 3-Ethoxypropanamin. Das simulierte Spektrum ist mit dem Spektrum des Trainingsmoleküls 3-Methoxypropanamin identisch.

6.9.8.1 Simulation durch Interpolation

Neben der Fähigkeit das Molekül mit dem ähnlichsten IR-Spektrum zu finden, sollte die Methode der anfrageorientierten Simulation, aufgrund des verwendeten neuronalen Counterpropagation-Netzes, auch zu Interpolation zwischen den Spektren mehrerer Trainingsmoleküle fähig sein. Ein Beispiel für eine solche Simulation, in der das simulierte Spektrum interpoliert wurde, ist die Simulation von N^1-(2-Aminoethyl)-propan-1,3-diamin. Simuliertes und experimentelles IR-Spektrum sind in Abbildung 87 dargestellt, der Korrelationskoeffizient der Spektren beträgt 0.899. Der Korrelationskoeffizienten für den Bereich der Aminobanden (3500-

3060 cm^{-1}) ist mit 0.983 sogar noch höher, obwohl in diesem Bereich optisch die größten Intensitätsunterschiede im Spektrum bestehen, da aber die Linien in diesem Bereich immer dasselbe Intensitätsverhältnis bewahren ist der Korrelationskoeffizient hier hoch.

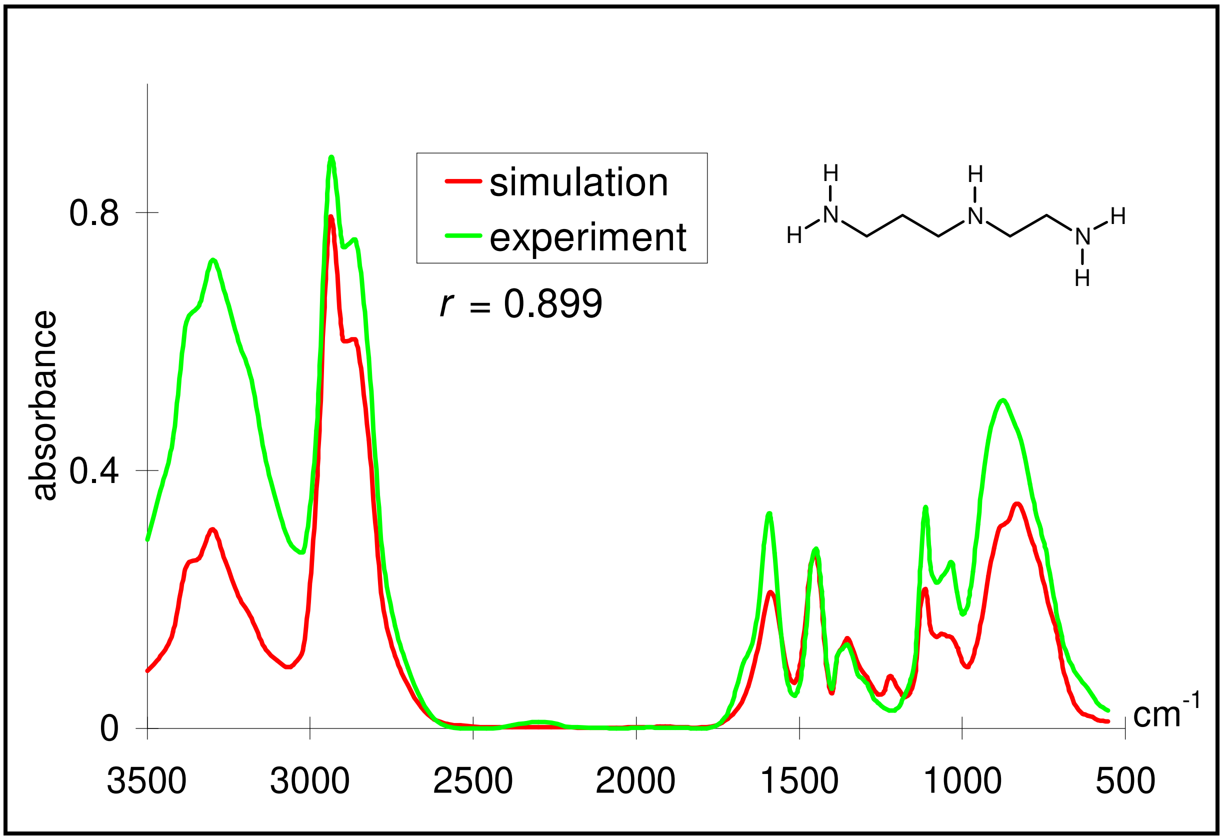


Abbildung 87: Interpoliertes Infrarotspektrum von N[1]-(2-Aminoethyl)-propan-1,3-diamin und das experimentelle Spektrum dieser Verbindung. Der Korrelationskoeffizient zwischen experimentellem und simuliertem Spektrum beträgt für das gesamte Spektrum 0.899 und für den Bereich der NH-Valenzschwingung zwischen 3500 und 3060 cm^{-1} 0.979.

Abbildung 88 zeigt die Lage des Testmoleküls in der rechten oberen Ecke, des zur Simulation benutzten CPG-Netzes. Unschwer sind die Übereinstimmungen der vier Moleküle im Bereich der Stickstoffsubstituenten zu erkennen. Mit Hilfe der unterschiedlichen Kohlenstoffkettenlängen gelang vermutlich die Simulation der Ethyl- und Propylkette vom N^1-(2-Aminoethyl)-propan-1,3-diamin.

Die Lage des Testmoleküls in der rechten oberen Ecke des planaren CPG-Netzes wirft aber auch ein Schlaglicht auf den Trainingsdatensatz. Denn die Konzentration der ähnlichsten Trainingsmoleküle in einer Ecke des planaren Netzes bedeutet nichts anderes, als daß das Testmolekül am Rande des Datenraums des Trainingsdatensatzes lag, denn das CPG-Netz erhält die Nachbarschaftsbeziehungen des Trainingsdatensatzes. Wenn das Testmolekül in der Mitte des Datenraums liegt, dürften die Möglichkeiten für eine Interpolation weiter steigen. Hier Methoden zu finden, die gewährleisten, daß entweder das Trainingsmolekül in der Mitte des Datensatzes liegt, oder im Falle das dem nicht so ist, davor gewarnt wird, ist eine der zukünftigen Aufgaben zur Verbesserung der Methode.

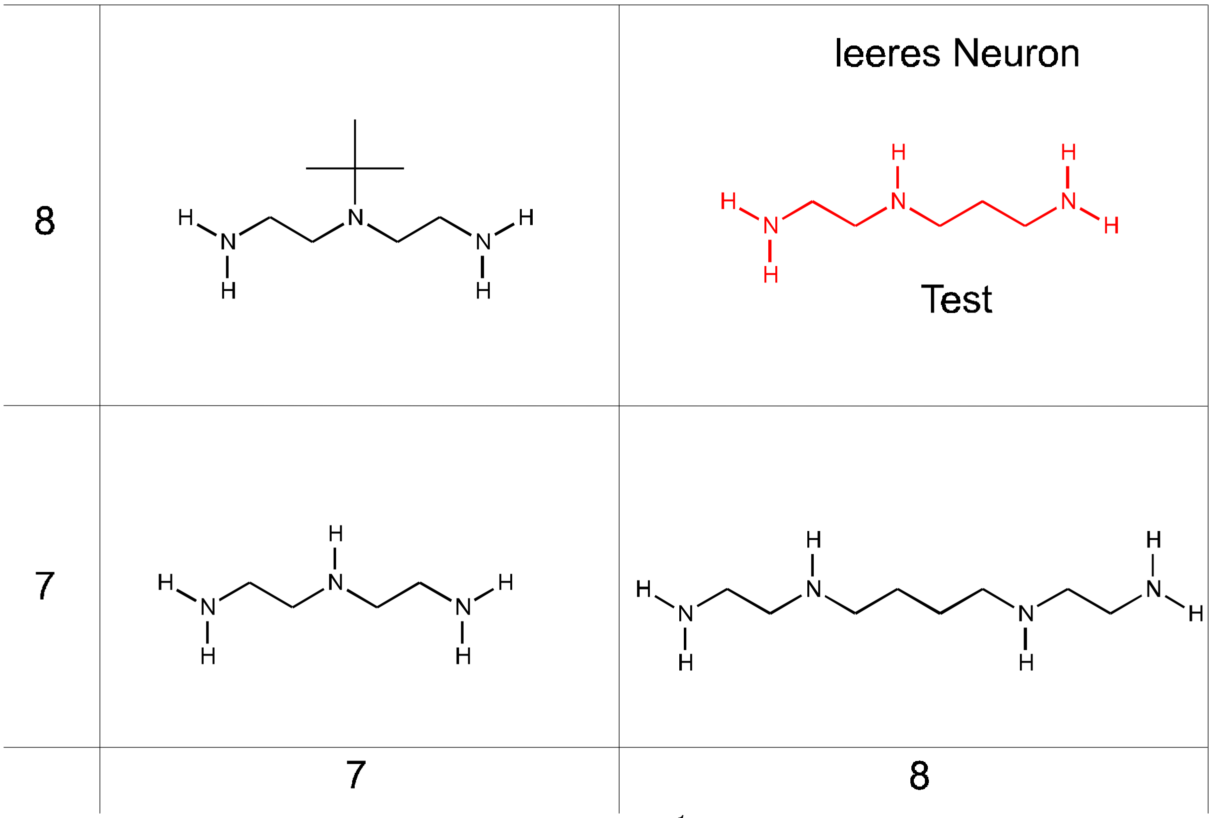


Abbildung 88: Die Lage des Testmoleküls *N*[1]-(2-Aminoethyl)-propan-1,3-diamin im CPG-Netz zusammen mit den Trainingsmolekülen, deren IR-Spektren bei der Interpolation des simulierten IR-Spektrums von *N*[1]-(2-Aminoethyl)-propan-1,3-diamin genutzt wurden.

6.10 Simulationen von Infrarotspektren zur Charakterisierung von Reaktionsprodukten

Das Ziel organisch-chemischer Forschung ist häufig die Synthese neuer Stoffe. Spektren erstmals synthetisierter Stoffe können aber nicht in Spektrendatenbanken enthalten sein und damit entfällt der übliche Weg zur Identifizierung von Stoffen durch Vergleich des gemessenen Spektrums mit dem Datenbankspektrum. Um die Natur des neuen Stoffes zu klären ist deshalb eine Interpretation seiner Spektren notwendig. Simulierte Infrarotspektren neuer Verbindungen könnten eine wesentliche Hilfe bei der Interpretation der Spektren dieser Verbindungen sein, zumal bei jeder geplanten Synthese die Strukturen des Syntheseziels und der vermutlichen Nebenprodukte bekannt sind. Bei einer erfolgreich durchgeführten Synthese einer neuen Verbindung kann diese durch den Vergleich des experimentellen Spektrums mit dem simulierten Infrarotspektrum des Syntheseziels identifiziert werden, wenn die Spektren beider Verbindungen übereinstimmen. Wie dies in der chemischen Praxis möglich ist, soll an zwei einfachen Reaktionen demonstriert werden, der Fries-Umlagerung von Essigsäurephenylester und der Oxidation von (4-Methylcyclohex-3-enyl)-methanol in Gegenwart von Methanol.

6.10.1 Fries-Umlagerung von Essigsäurephenylester

Bei der Fries-Umlagerung von Phenylestern werden diese gespalten und es entsteht ein Phenolderivat mit einer freien phenolischen OH-Gruppe, das am Benzolring mit dem Säurerest des Esters acyliert ist.[86] Wo die Acylierung erfolgt, hängt vom Substrat und den Reaktionsbedingungen ab. Ist der Ring des Phenylesters nicht weiter substituiert sind *ortho*- und *para*-Stellung bevorzugt, wobei zwischen diesen mit Hilfe der Reaktionsbedingungen weitgehend selektiert werden kann. Abbildung 89 zeigt ein Beispiel für eine Fries-Umlagerung, deren Hauptprodukt das in Abbildung 90 gezeigte IR-Spektrum aufweist.

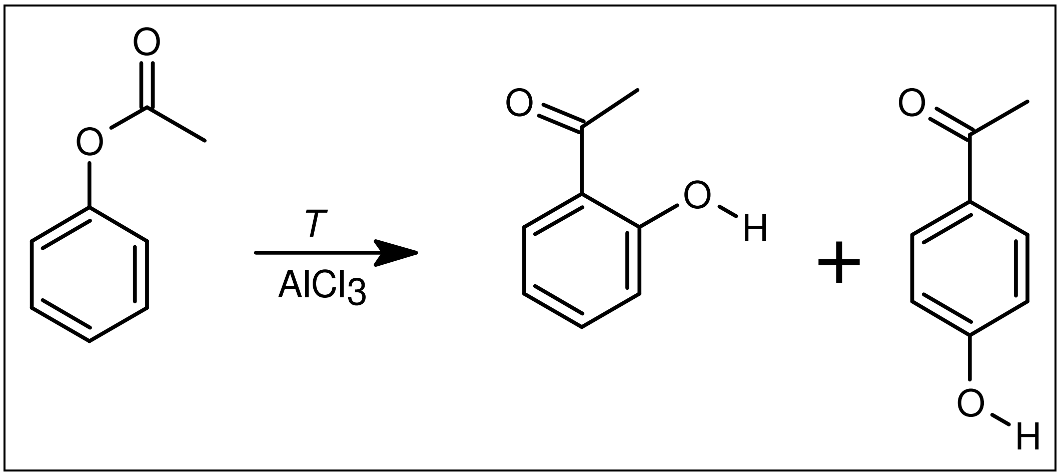


Abbildung 89: Reaktionsschema der Fries-Umlagerung mit den beiden möglichen Reaktionsprodukten der Umlagerung von Essigsäurephenylester.

Das IR-Spektrum des Hauptprodukts der obigen Reaktion ist durch sehr intensive Banden zwischen 1650 und 1000 cm^{-1} geprägt, die Banden unterhalb von 1000 cm^{-1} sind intensitätsschwach und die Banden bei 952, 824 cm^{-1} (mit einer Schulter bei 776 cm^{-1}) und 568 cm^{-1} passen in kein Interpretationsschema für Aromatensubstitution. So findet sich bei Pretsch et al. für unseren Fall eines konjugierten Carbonylsubstituenten nur die Bemerkung, daß die aromatischen CH-Deformationsschwingungen bei Anwesenheit stark konjugierter Substituenten wie Carbonyl, Nitrat oder Nitril nicht brauchbar seien.[87] Bei Hesse et al. findet sich obige Einschränkung nicht. Nach den dortigen Angaben ist für die *para*-di-Substitution eine scharfe intensive Bande zwischen 840 und 800 cm^{-1} charakteristisch. Das Kennzeichen für *ortho*-Disubstitution wäre hiernach eine scharfe Bande zwischen 760 und 740 cm^{-1}, die sich, wie auch die *para*-Substitutionsbande bei Anwesenheit stark elektronenziehender Substituenten zu höheren Wellenzahlen verschieben könne. Der Blick auf das nachfolgende Spektrum (Abbildung 90) zeigt keine intensive Bande in den genannten Bereichen und die Doppelbande bei 824 und 776 cm^{-1} würde, da ein elektronenziehender Substituent vorhanden ist, im Zweifelsfalle sowohl den Kriterien für *ortho*- als auch für *para*-Disubstitution genügen. Obendrein fehlt in dem experimentell gemessenen Spektrum die Carbonylbande, die üblicherweise bei 1700 cm^{-1} als intensivste Bande im Infrarot-Spektrum einer Carbonylverbindung zu sehen ist.

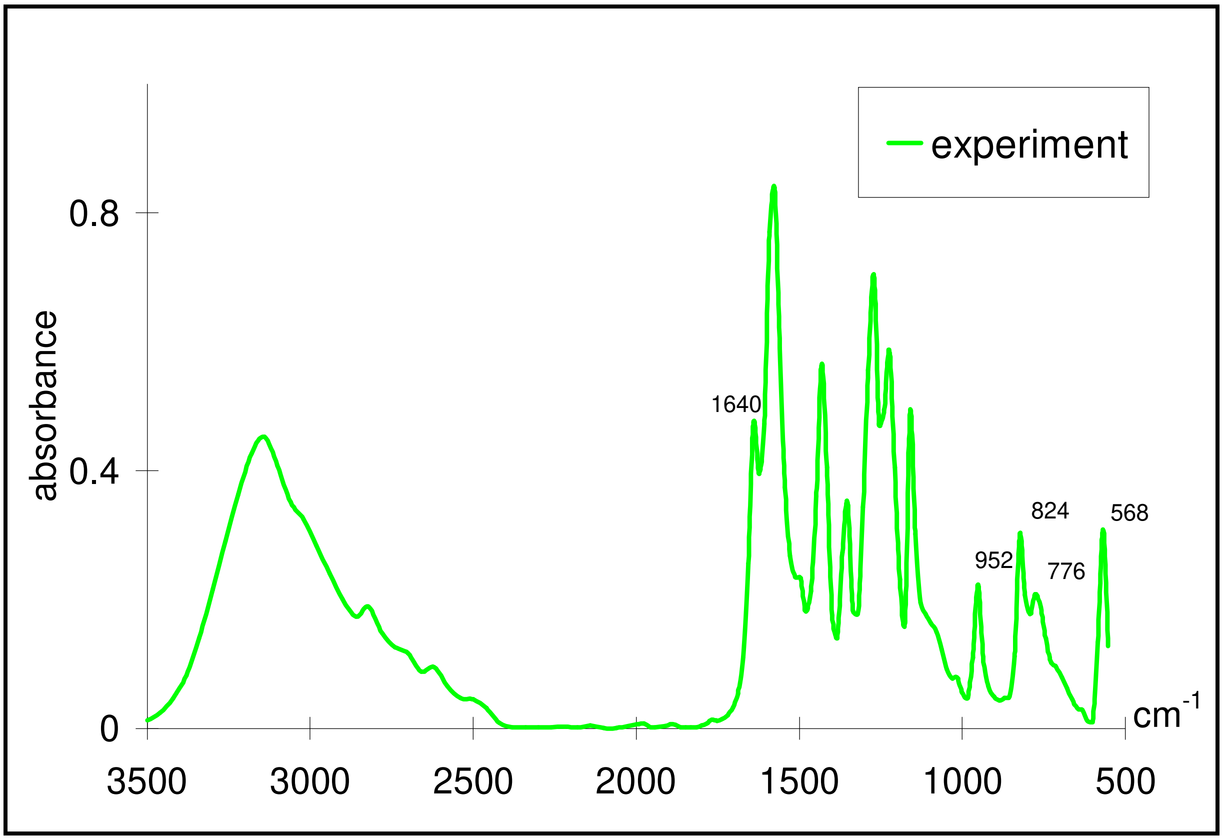


Abbildung 90: Das experimentelle Spektrum des Hauptprodukts der Fries-Umlagerung von Essigsäurephenylester.

Die folgenden zwei Abbildungen zeigen das experimentelle Spektrum im Vergleich mit den simulierten Spektren für das *ortho*- und das *para*-Produkt. Die Spektren wurden mit Hilfe der anfrageorientierten Methode unter Verwendung der Standardparameter simuliert. Das verwendete CPG-Netz hatte somit 8 x 8 Neuronen und war planar. Für den zur Codierung der Molekülstrukturen verwendeten 3D-MoRSE Code galten die folgenden Parameter: n=64, $A_i = q_{tot,i}$ und $s_{max} = 15.5$ Å^{-1}.

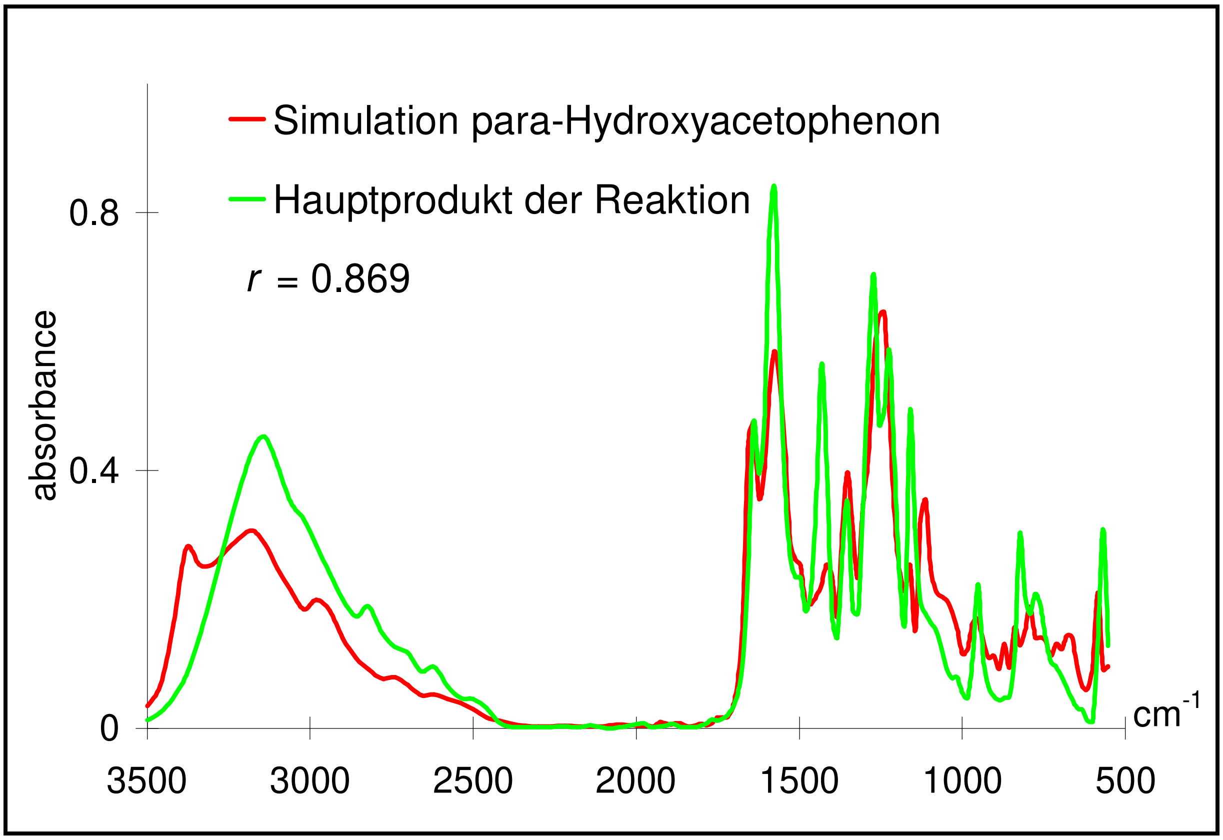


Abbildung 91: Experimentelles Spektrum des Hauptproduktes der Fries-Umlagerung sowie das simulierte Spektrum des möglichen Reaktionsproduktes *para*-Hydroxyactephenon. Der Korrelationskoeffizient zwischen beiden Spektren beträgt 0.869.

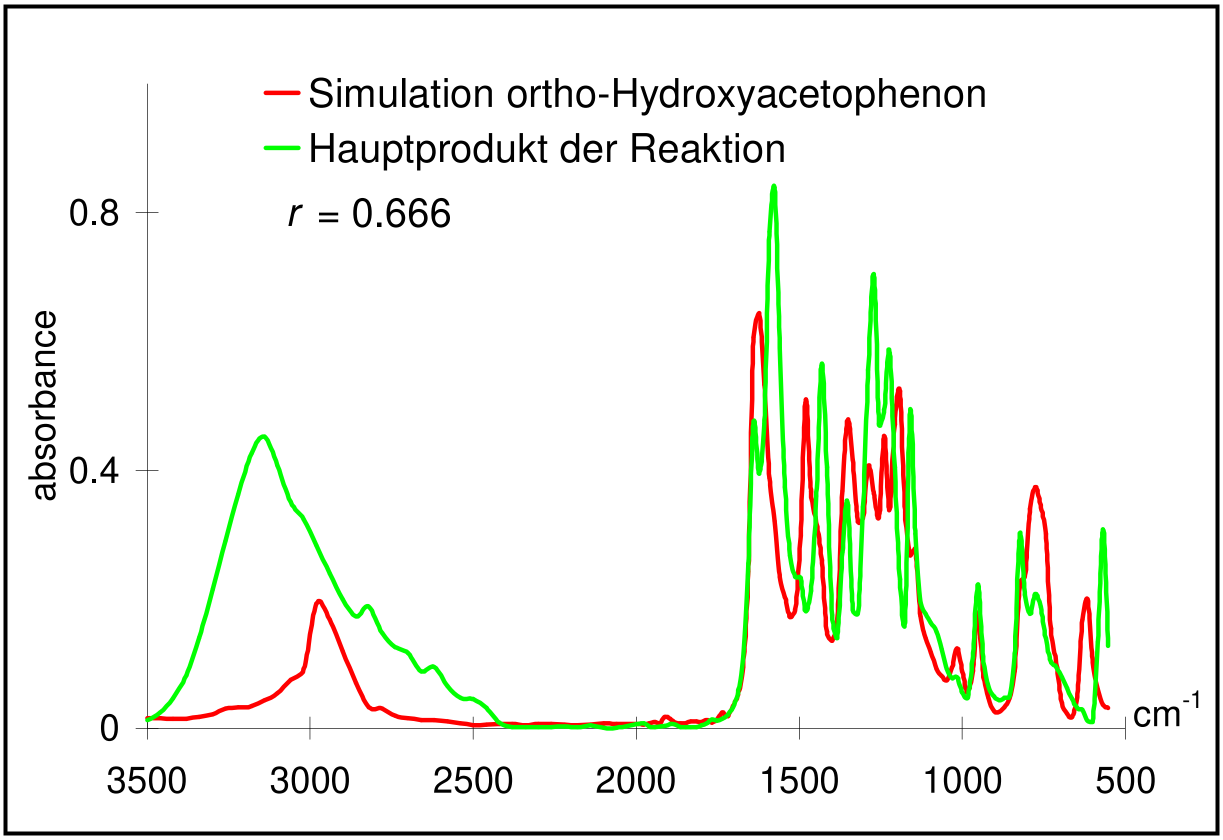


Abbildung 92: Experimentelles Spektrum des Hauptproduktes der Fries-Umlagerung sowie das simulierte Spektrum des möglichen Reaktionsproduktes ortho-Hydroxy-actephenon. Der Korrelationskoeffizient zwischen beiden Spektren beträgt 0.666.

Wie schon der Korrelationskoeffizient zwischen dem Infrarot-Spektrum des Hauptproduktes der Fries-Umlagerung und den simulierten Spektren der möglichen Reaktionsprodukte zeigt, machen die simulierten Infrarotspektren *para*-Hydroxyacetophenon als Reaktionsprodukt wahrscheinlicher. Noch eindeutiger ist die Sache beim Nebenprodukt *ortho*-Hydroxyacetophenon. Hier stimmen die Banden im Fingerprintbereich des simulierten Spektrums von *ortho*-Hydroxyacetophenon im wesentlichen mit den Banden des experimentellen Spektrums, des Nebenproduktes *ortho*-Hydroxyacetophenon, überein. Abweichungen beschränken sich hier im wesentlichen auf die Peakform, so sind die zwei experimentellen Peakspitzen bei 1432 und 1144 cm^{-1} im simulierten Spektrum nur als Schultern zu sehen. Demgegenüber weicht das simulierte Infrarotspektrum von *para*-Hydroxyacetophenon wesentlich deutlicher vom experimentellen Spektrum des Nebenproduktes ab. Dies zeigt sich durch drei zusätzliche Peaks zwischen 1600 und 1000 cm^{-1} im simulierten Spektrum von *para*-Hydroxyacetophenon und der, gegenüber dem experimentellen Spektrum, total veränderte

Verlauf der Absorption zwischen 1150 und 550 cm^{-1}. Dies kommt ebenfalls im Korrelationskoeffizienten zum Ausdruck. So beträgt der Korrelationskoeffizient zwischen dem experimentellem Spektrum und dem simulierten Spektrum von *ortho*-Hydroxyacetophenon 0.909, während der Korrelationskoeffizient zwischen dem experimentellem Spektrum und dem simulierten Spektrum von *para*-Hydroxyacetophenon 0.724 beträgt. Abbildung 93 zeigt die simulierten Spektren und das experimentelle Spektrum des Nebenproduktes.

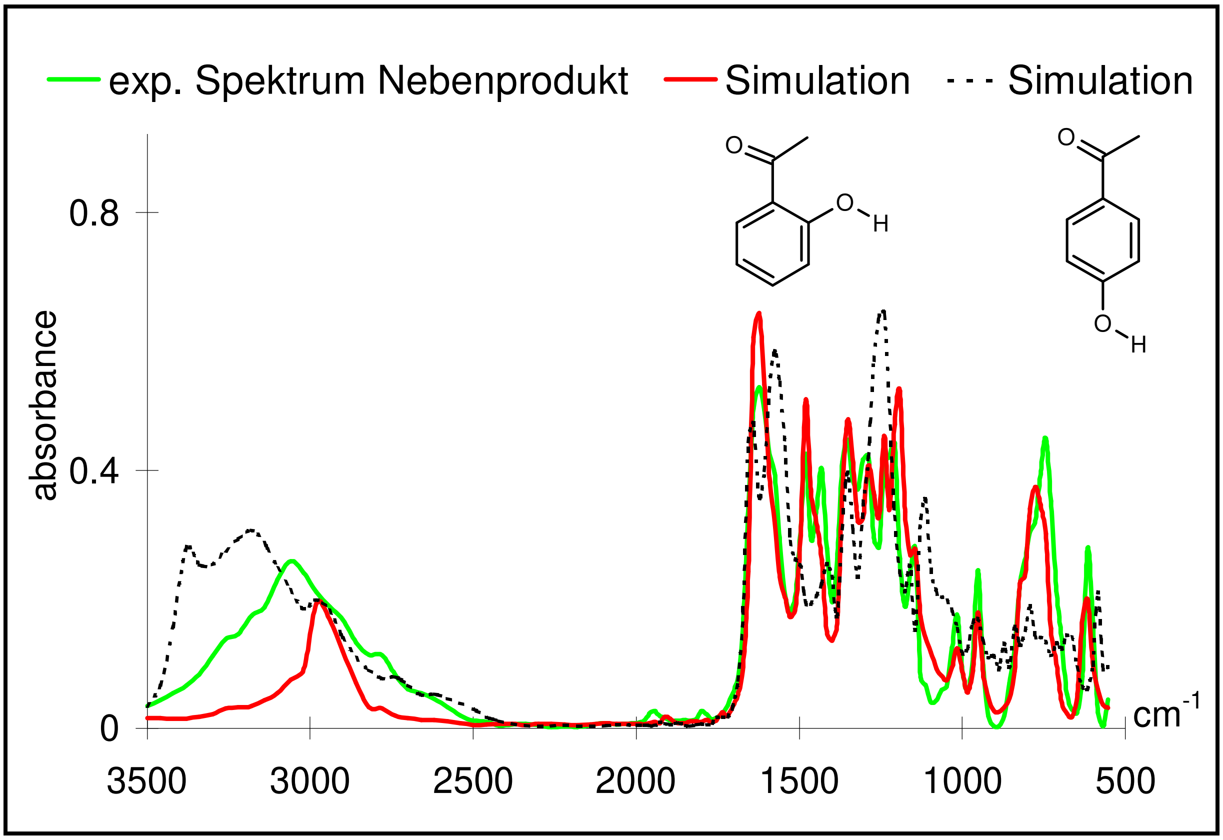


Abbildung 93: Das experimentelle Infrarotspektrum des Nebenprodukts (ortho-Hydroxyacetophenon) und die simulierten Spektren der beiden möglichen Umlagerungsprodukte *ortho*- und *para*-Hydroxyacetophenon. Die Korrelationskoeffizienten zwischen dem experimentellem Spektrum und den simulierten Spektren betragen für ortho-Hydroxy-acetophenon 0.909 und für para-Hydroxy-acetophenon 0.724.

6.10.2 Die Oxidation von (4-Methylcyclohex-3-enyl)-methanol

Ein Chemiker hat (4-Methylcyclohex-3-enyl)-methanol in methanolischer Lösung oxidiert, Ziel der Synthese war (4-Methylcyclohex-3-enyl)-carbonsäure. Da das nach dem Aufarbeiten erhaltene Reaktionsprodukt nicht sauer reagiert, versendet er einen Analyseauftrag mit folgender Reaktionsgleichung an das IR-Labor:

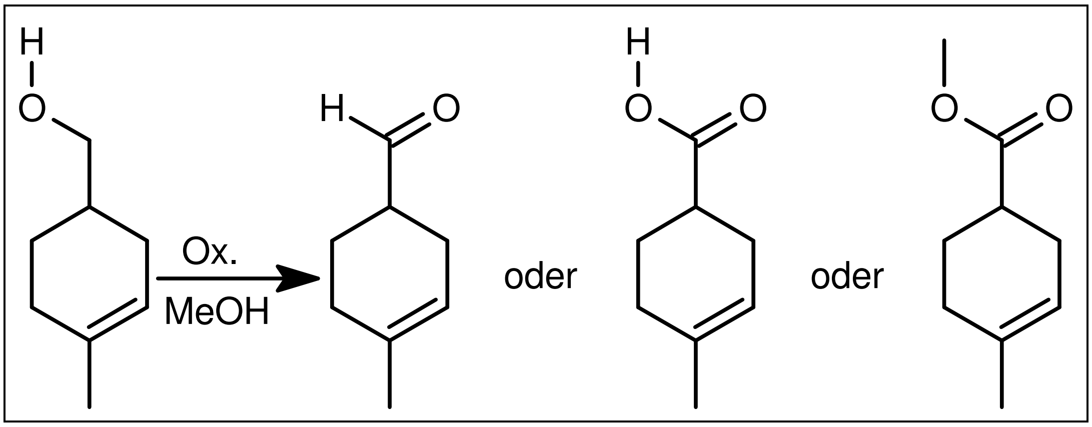


Abbildung 94: Mögliche Reaktionsprodukte der Oxidation von (4-Methylcyclohex-3-enyl)-methanol in methanolischer Lösung.

Das Infrarotspektrum des Reaktionsproduktes (Abbildung 95) zeigt keine charakteristisch verbreiterte OH-Bande zwischen 3500 und 2500 cm^{-1}, wie sie für Carbonsäuren typisch ist. Die scharfe und intensive Bande bei 1736 cm^{-1} kann als Carbonylbande interpretiert werden, allerdings ohne das hierdurch eine Unterscheidung zwischen dem Aldehyd und dem Methylester möglich wäre, zumal die für Ester typischen, sehr intensiven Banden der C-O - Bindung bei 1200 und 1000 cm^{-1} nicht eindeutig erkennbar sind. Die Bande der CC-Doppelbindung bei 1600 cm^{-1} ist nicht sichtbar.

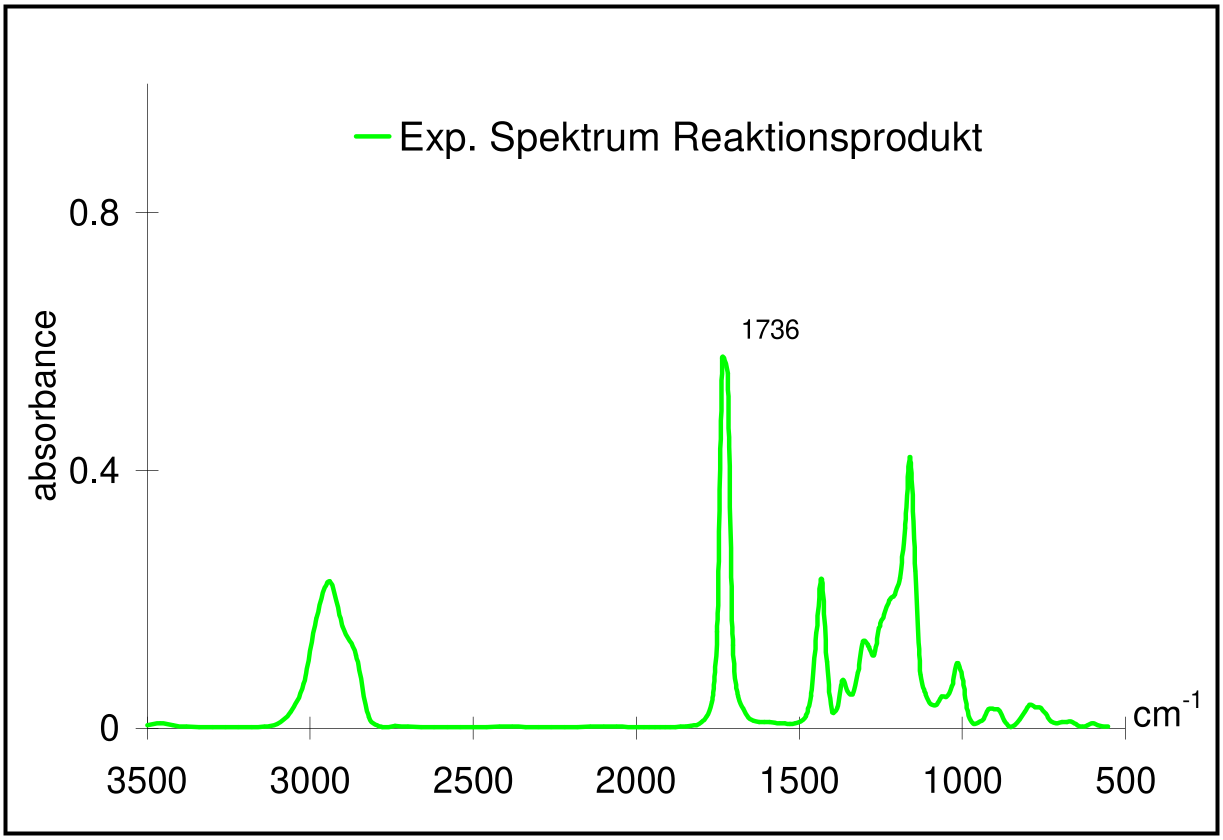


Abbildung 95: Experimentelles Infrarotspektrum des Reaktionsproduktes der Oxidation von (4-Methylcyclohex-3-enyl)-methanol.

Aus dem vorstehenden wird klar, daß die Säure vermutlich nicht isoliert wurde. Eine Unterscheidung anhand des Spektrums, ob die Oxidation auf der Stufe des Aldehyds stehen geblieben ist oder ob die Carbonsäure unter den Reaktionsbedingungen (methanolische Lösung) gleich zum Methylester weiter reagiert hat, ist anhand des experimentellen Infrarotspektrums nicht ohne weiters möglich. Der Vergleich des experimentellen Spektrums des Oxidationsproduktes mit den simulierten Spektren des Aldehyds und des Methylesters hilft weiter, wie Abbildung 96 zeigt.

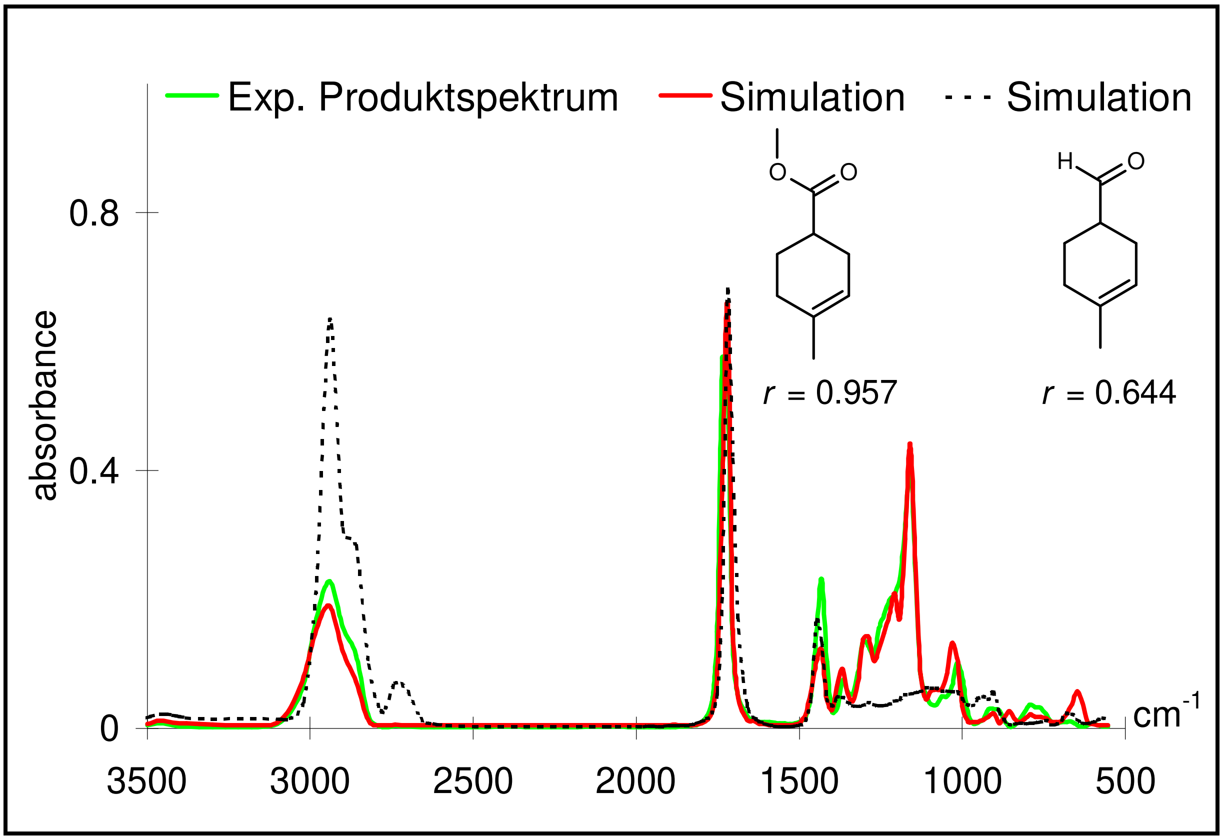


Abbildung 96: Das Produktspektrum der Oxidation von (4-Methylcyclohex-3-enyl)-methanol sowie die simulierten IR-Spektren der möglichen Reaktionsprodukte (4-Methylcyclohex-3-enyl)-methansäuremethylester und (4-Methylcyclohex-3-enyl)-methanal.

Der optische Vergleich zeigt die hohe Übereinstimmung des simulierten Spektrums von (4-Methylcyclohex-3-enyl)-methansäuremethylester mit dem experimentellen Spektrum des Reaktonsproduktes und identifiziert damit das Reaktionsprodukt als (4-Methylcyclohex-3-enyl)-methansäuremethylester. Das simulierte Spektrum von (4-Methylcyclohex-3-enyl)-methanal weicht im Bereich der CH-Valenzschwingungsbanden und im Fingerprintbereich sehr stark vom experimentellen Spektrum ab, womit (4-Methylcyclohex-3-enyl)-methanal als Reaktionsprodukt ausgeschlossen werden kann. Das Ergebnis des optischen Vergleichs spiegelt sich auch in den Korrelationskoeffizienten wieder. So beträgt der Korrelationskoeffizient zwischen dem Spektrum des Reaktionsproduktes und dem simulierten Spektrum von (4-Methylcyclohex-3-enyl)-methansäuremethylester 0.957, bzw. 0.644 zwischen dem Spektrum des Reaktionsproduktes und dem simulierten Spektrum von (4-Methylcyclohex-3-enyl)-methanal.

Die Spektren wurden mit Hilfe der anfrageorientierten Methode simuliert. Das verwendete CPG-Netz hatte 10 x 10 Neuronen und war toroidal. Für den zur Codierung der Molekülstrukturen verwendeten 3D-MoRSE Code galten die folgenden Parameter: n=64, $A_i = q_{tot,i}$ und s_{max} = 15.5 Å^{-1}. Alle anderen Parameter für die anfrageorientierte Methode entsprachen den Standardwerten.

Das experimentelle Spektrum des Reaktionsproduktes dieses fiktiven Beispiels stammt wie alle experimentellen Spektren dieser Arbeit aus der SpecInfo-Datenbank. Die Simulationen waren weitestgehend korrekt, wie ein Vergleich der oben gezeigten simulierten Spektren mit den experimentellen Spektren aus der Datenbank zeigt.

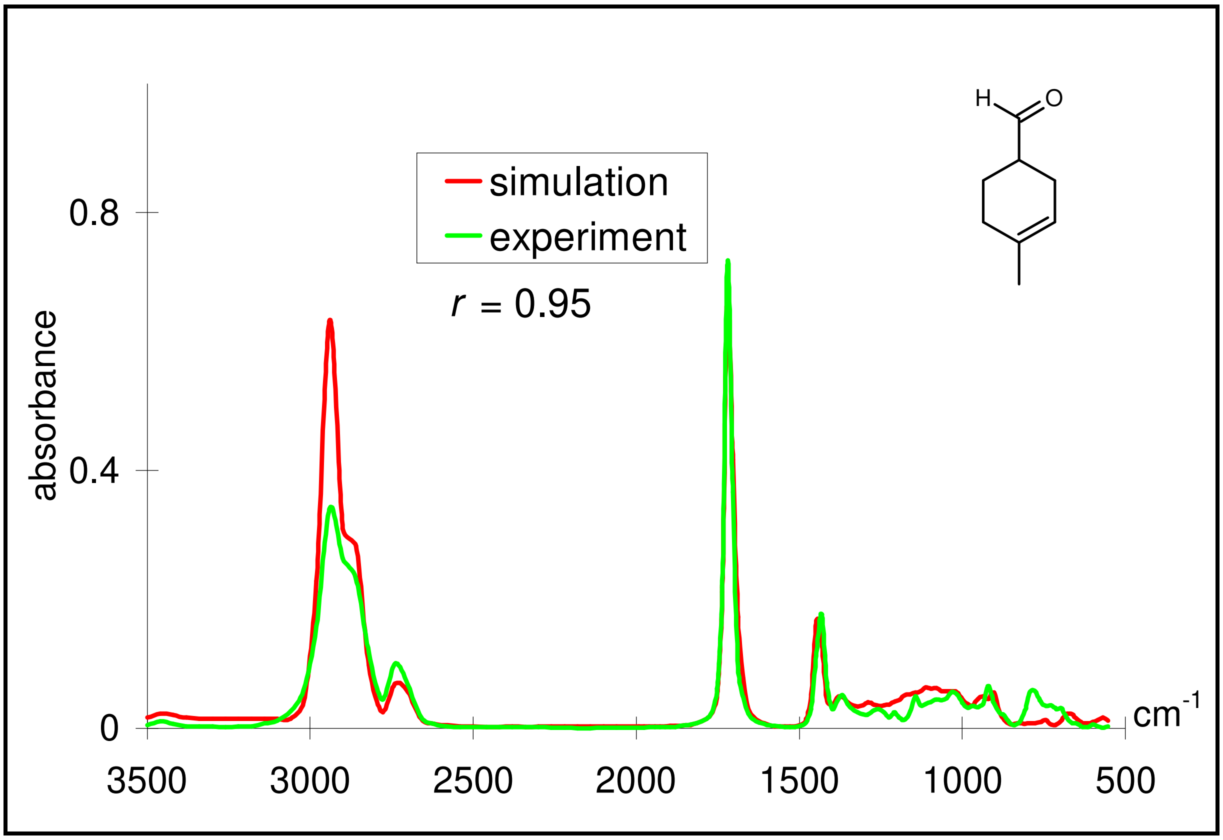


Abbildung 97: Experimentelles und simuliertes IR-Spektrum von (4-Methylcyclohex-3-enyl)-methanal mit einem Korrelationskoeffizient von 0.950.

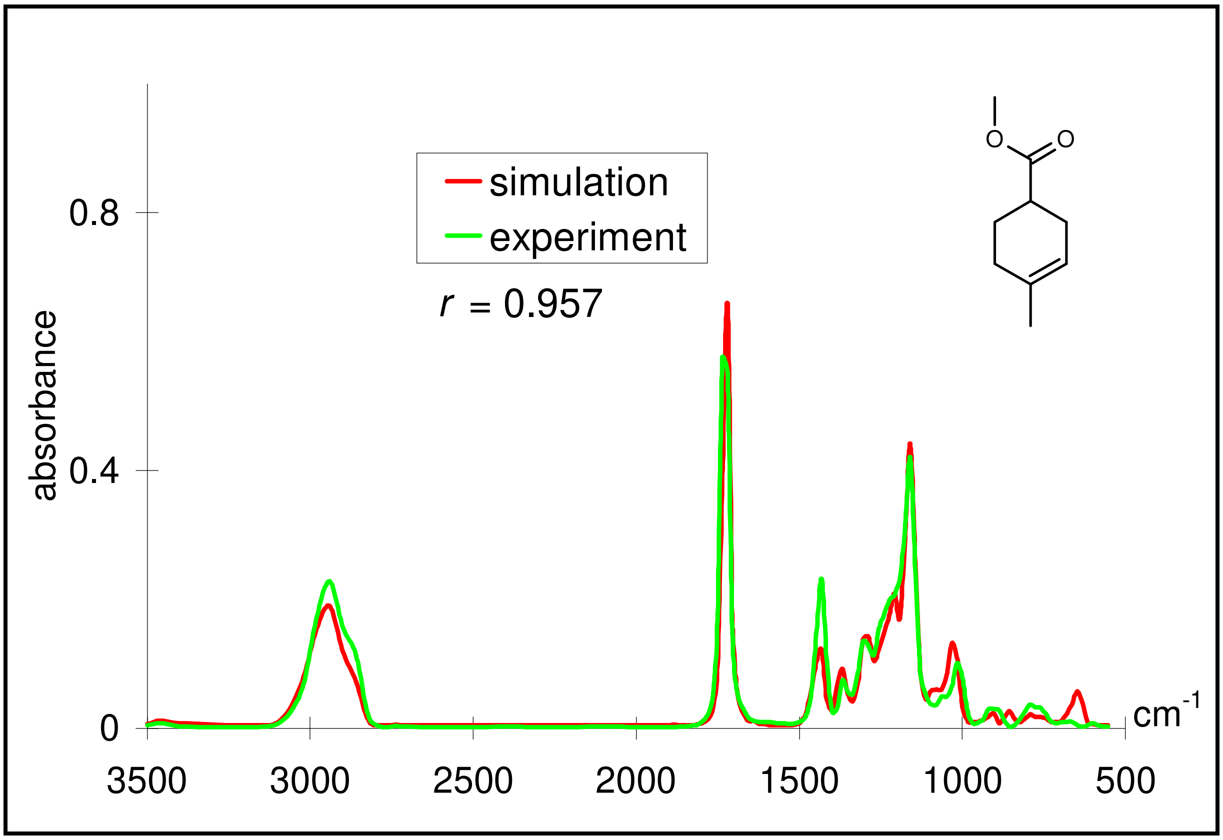


Abbildung 98: Experimentelles und simuliertes IR-Spektrum von (4-Methylcyclohex-3-enyl)-methansäuremethylester mit einem Korrelationskoeffizient von 0.957.

Die Notwendigkeit von Spektrensimulationen in der täglichen Praxis unterstreicht die Tatsache, daß das experimentelle Spektrum des mutmaßlichen Reaktionsproduktes von (4-Methylcyclohex-3-enyl)-methansäure nicht in der SpecInfo-Datenbank enthalten war. Die Simulation des IR-Spektrums von (4-Methylcyclohex-3-enyl)-methansäure war jedoch möglich. Abbildung 99 zeigt das simulierte Spektrum.

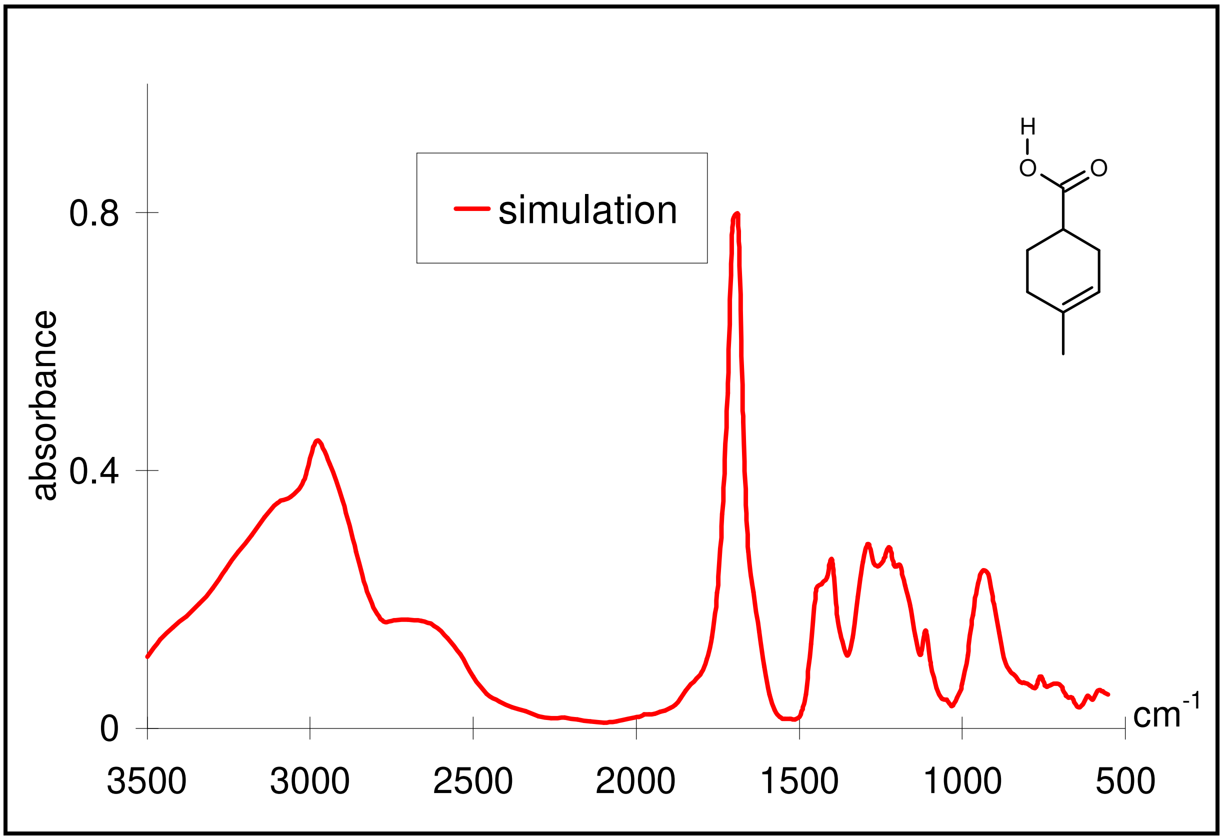


Abbildung 99: Das simulierte Spektrum von (4-Methyl-cyclohex-3-enyl)-methansäure. Ein experimentelles Spektrum dieser Substanz ist nicht in der SpecInfo-Datenbank[52] enthalten.

Das simulierte Spektrum von (4-Methylcyclohex-3-enyl)-methansäure zeigt die für Carbonsäuren typische verbreiterte OH - Bande zwischen 3500 und 2500 cm^{-1} sowie die Carbonylbande. Damit sind die ersten Kriterien für eine korrekte Simulation erfüllt. Ob aber der Fingerprintbereich unterhalb von 1500 cm^{-1} korrekt simuliert wurde, kann ohne ein Vergleichsspektrum nicht beurteilt werden. Die Angaben von Absorptionsbanden durch Manjarrez et al. von Banden in CCl_4 bei 2900, 1710, 1440, 1380, 1140, 1075 und 1025 cm^{-1} genügen für ein Spektrenvergleich nicht, da beispielsweise Angaben zur Intesität und Gestalt der Banden sowie zur Auflösung des Spektrums fehlen.[88] Ein Vergleich dieser experimentellen Peaklagen mit den Peaklagen des simulierten Spektrums zeigt jedoch eine Übereinstimmung bis auf die beiden letzten Banden bei 1075 und 1025 cm^{-1}, wobei zumindest die letzte Bande auch auf CCl_4 zurückgehen könnte, dessen stärkere Banden zwischen 1600 und 1480, 1280 und 1200, 1020 und 960, 860 und 680, 620 und 610 und 480 und 420 cm^{-1} das Infrarotspektrum der untersuchten Substanz stark beeinflussen.[89] Eine Beurteilung der Simulation ist hiermit aber aufgrund der fehlenden Angaben zum Spektrum (z.B. Peakformen, Intensitäten und Schichtdicke) nur sehr begrenzt möglich. Es wäre deshalb wichtig, wenn aus der Simulation selbst eine Abschätzung der Zuverlässigkeit der Simulation möglich wäre, zumal, wie eingangs bereits erwähnt, das zur Simulation benutzte CPG-Netz immer ein Ergebnis liefern wird, welches aber nicht mit dem experimentellem Spektrum übereinstimmen muß. Die bisher gefundenen Möglichkeiten zur Vorhersage der Zuverlässigkeit der Simulation werden im folgenden Kapitel diskutiert.

6.11 Versuche zur Vorhersage der Simulationsqualität

Die Vorhersage der Simulationsqualität muß ohne zusätzliches äußeres Wissen, allein aus Informationen, die vor der Durchführung einer anfrageorientierten Simulation vorhanden sind oder während des Simulationsvorganges entstehen, erfolgen. Damit sind die zur Verfügung stehenden Informationen beschränkt auf die Anfragestruktur, den Trainingsdatensatz, das trainierte Netz und die Abfrage des trainierten Netzes.

Einfache Versuche, die Simulationsqualität mit dem *rms*-Wert zwischen dem Strukturcode der Anfragestruktur und dem Strukturteil des Neurons zu korrelieren, scheiterten, ebenso wie der Versuch die Simulationsqualität mit dem mittleren *rms*-Wert zwischen dem Strukturcode der Anfragestruktur und den Strukturcodes des Trainingsdatensatzes zu korrelieren. Ob der Ansatz des betragsgewichteten *rms*-Wertes, wie er im Kapitel über die Simulation der Infrarotspektren primärer Amine (Kapitel 6.9.5) vorgestellt wurde, weiterführt, kann noch nicht abschließend beurteilt werden. Wie Abbildung 100 zeigt, warnt der betragsgewichtete *rms*-Wert nur in sieben Fällen der 77 Simulation des Datensatzes der primären Amine nicht vor einem schlechten Simulationsergebnis (Simulationen oberhalb der gestrichelten helle Linie in Abbildung 100). Am deutlichsten versagt der betragsgewichtete *rms*-Wert bei der Simulation von 3-Aminopropanol ($r = 0.233$, $AN_g = 2.5$), für dessen Simulation die Methode der anfrageorientierten Simulation aber prinzipiell nicht geeignet ist (vgl. Kapitel 6.9.4.1), bzw. dessen Datenbankspektrum vermutlich auch nicht korrekt ist. Drei der fehlenden Warnungen haben ihre Ursache in anderen Meßbedingungen (vgl. die Simulationen von 1-Aminopentan, 1-Aminooctan und 1,2-Diaminoethan in Kapitel 6.9.6). Bei einer der fehlenden Warnungen vor Abweichungen handelt es sich um die Simulation von 2-Phenylethylamin, hier könnte der verwendete 3D-MoRSE die Ursache sein (vgl. Kapitel 6.9.5). Bei beiden anderen fehlenden Warnungen ist der Grund wegen fehlender Information zur Meßmethode nicht zu erkennen.

Der betragsgewichtete *rms*-Wert kann nur zur Vorhersage der Simulationsqualität genutzt werden, wenn man davon ausgeht, daß wenigstens in etwa für alle Simulationen eine hypothetische Korrelation (durchgezogene Linie) zwischen betragsgewichtetem *rms*-Wert und Korrelationskoeffizient gelten soll. D.h. alle Simulationen wenigstens in dem durch die gestrichelten Linien begrenzten Bereich liegen sollten (vgl. Abbildung 100).

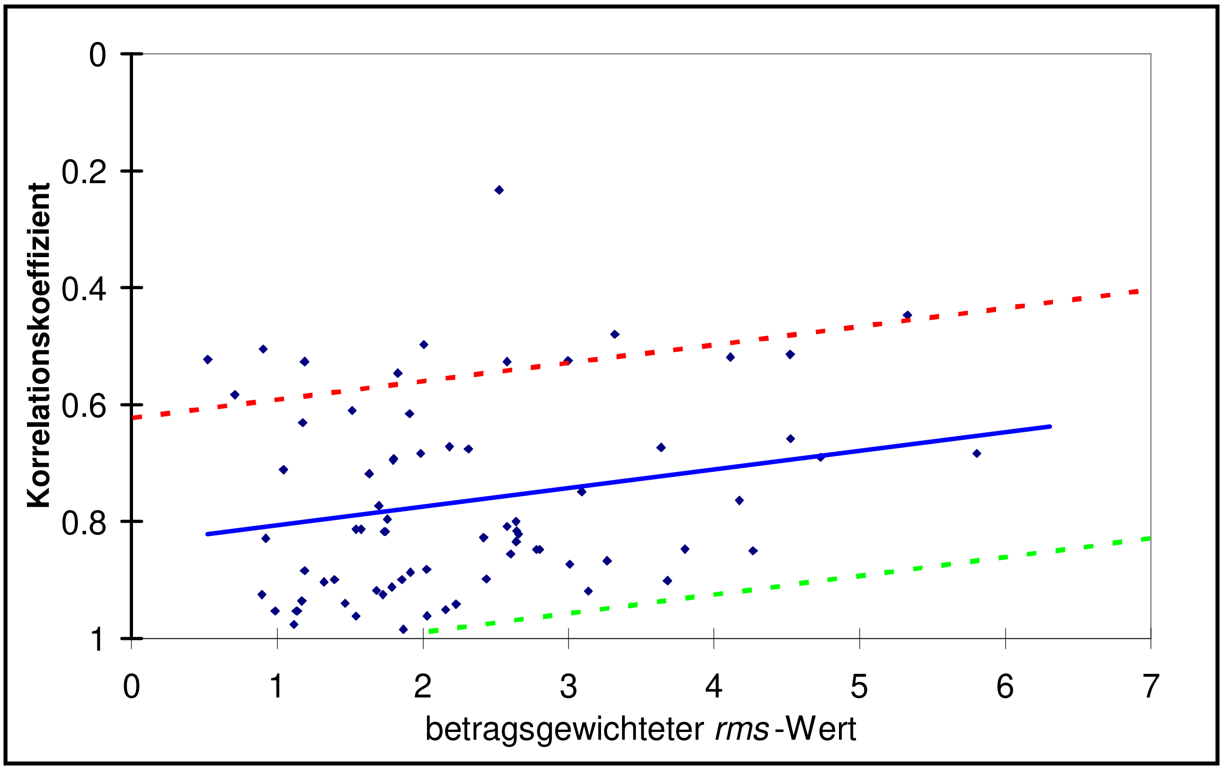


Abbildung 100: Betragsgewichteter *rms*-Wert aufgetragen gegen den Korrelationskoeffizienten zwischen experimentellem und simuliertem IR-Spektrum für die jeweilige Simulation. Die durchgezogene Linie zeigt eine hypothetische Korrelation zwischen dem betragsgewichtetem *rms*-Wert und dem Korrelationskoeffizient an, die gestrichelten Linien zeigen deren Grenzen.

Eins zeigt Abbildung 100: Mit dem betragsgewichtetem *rms*-Wert (AN_g vgl. Gleichung 21) ist die Lösung des Problems der Vorhersage der Simulationsqualität noch nicht gefunden, aber vielleicht kann hiermit in Zukunft ein Beitrag geliefert werden. Denn wie so manche der vorstehenden Simulationen und die Analyse der fehlenden Warnungen zeigte, ist die Ursache für Abweichungen zwischen experimentellem und simuliertem Spektrum die Verwendung einer anderen Meßmethode für die Anfragestruktur als für die Trainingsmoleküle, deren Spektren zur Vorhersage genutzt wurden. Zweitens repräsentiert der 3D-MoRSE Code die Struktur nicht immer optimal, wie auch drittens der Korrelationskoeffizient beim Vergleich der Infrarotspektren nicht immer optimale Ergebnisse liefert, so reagiert er bei bandenreichen Spektren auf die Differenzen bei einzelnen mittleren oder schwacher Banden zu gering.[41]

Solange aber die vorstehenden drei Ursachen für schlechte Simulationen bzw. für eine schlechte Korrelation von Gütemaß und Qualität der vorhergesagten Spektren nicht beseitigt sind, wird die Entwicklung eines funktionierenden quantitativen Gütemaßes für die Vorhersage der Simulationsqualität im Rahmen einer anfrageorientierten Simulation schwierig bis unmöglich bleiben.

Um so wichtiger bleibt die Kontrolle der anfrageorientierten Simulation durch den Benutzer selbst, indem man die Strukturen und Spektren der Trainingsmoleküle auf dem und um das zur Simulation benutzte Neuron untersucht, deren Strukturen mit der Anfragestruktur vergleicht und ebenso für die Spektren die Übereinstimmung der Meßmethode der Trainingsspektren mit der Meßmethode des experimentellen Spektrums überprüft. Beispielhaft ist dies für die Strukturen im Rahmen des vom DFN-Verein geförderten TeleSpec-Projektes realisiert.[90] Die in Abbildung 101 und Abbildung 102 abgebildeten Web-Simulationsseiten zeigen neben der Anfragestruktur auf der linken Seite und dem simulierten Infrarotspektrum oben rechts, die Draufsicht (Mitte oben) auf das zur Simulation genutzte CPG-Netz inklusive der Belegung der Neuronen mit Trainingsmolekülen. Das zur Simulation genutzte Neuron wird durch das helle, am Bildschirm goldene Symbol indiziert. Der Nutzer ist nun in der Lage, sich durch einen Mausklick auf ein Neuron, die dem Neuron assoziierten Trainingsmoleküle anzeigen zu lassen (rechte untere Ecke), sowie den Strukturcode des Neurons und das in dem Neuron gespeicherte Infrarot-Spektrum mit den Werten des zur Simulation benutzten Neurons zu vergleichen.

Bei der in Abbildung 101 gezeigten Simulation des Infrarotspektrums von Cyclobutanon wurde Neuron (9,2) des trainierten CPG-Netzes zur Simulation genutzt. Aufgrund der Tatsache, daß dem Nachbarneuron (10,2) das Spiro-Lacton 2,7-Dioxa-spiro[4,4]-1,6-dion assoziiert wurde, kann man schließen, daß das Simulationsergebnis für Cyclobutanon wahrscheinlich nicht optimal ist. Das Beispiel wurde gewählt, weil hier kein sinnvolles Simulationsergebnis möglich ist, da es keine Verbindung gibt deren Infrarotspektrum dem von Cyclobutanon ähnlich ist und die eine ähnliche Struktur wie Cyclobutanon hat, was auch anhand der benachbarten Moleküle klar zu erkennen ist.

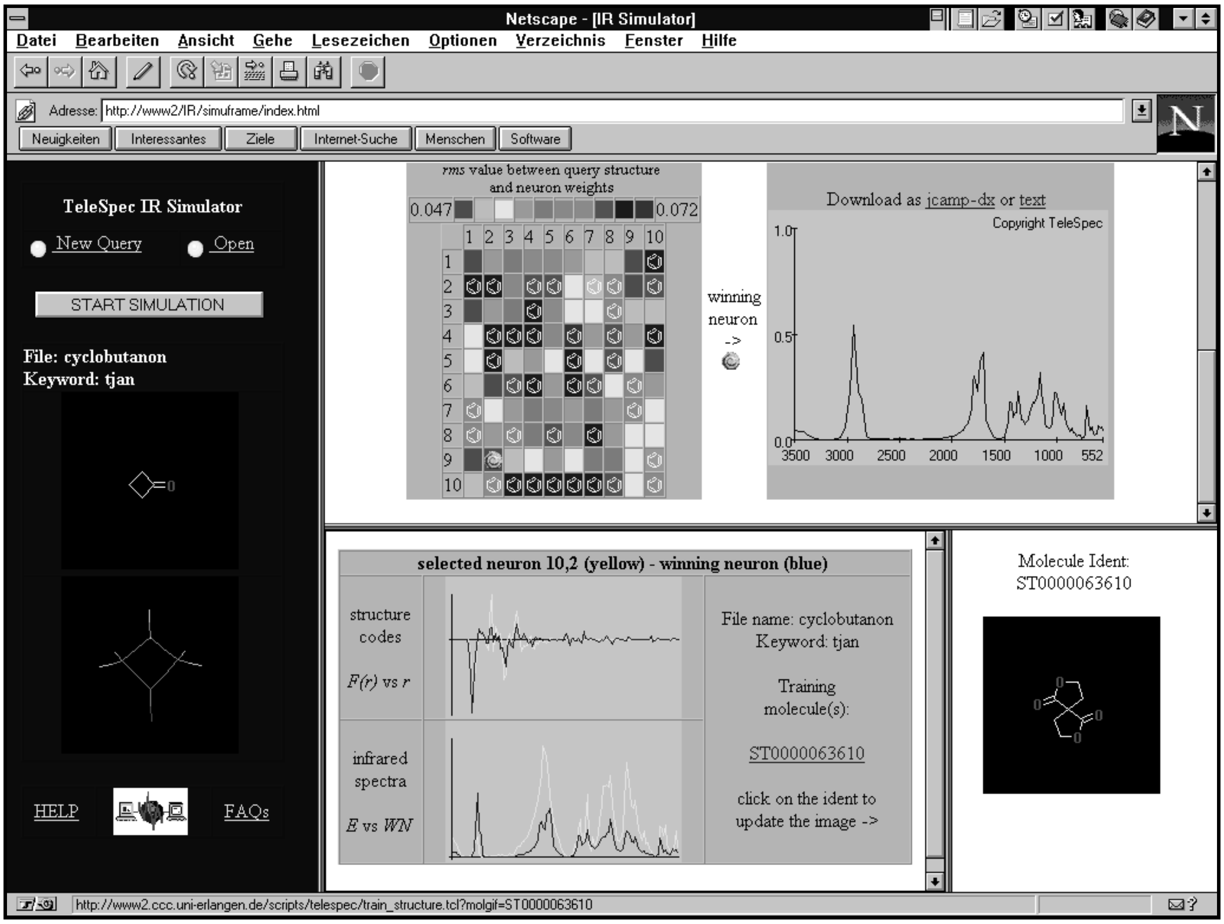


Abbildung 101: Bildschirm mit der TeleSpec Infrarotspektren - Simulationsseite und der Simulation des IR-Spektrums von Cyclobutanon. Zur Simulation wurde das Neuron (9,2) genutzt.

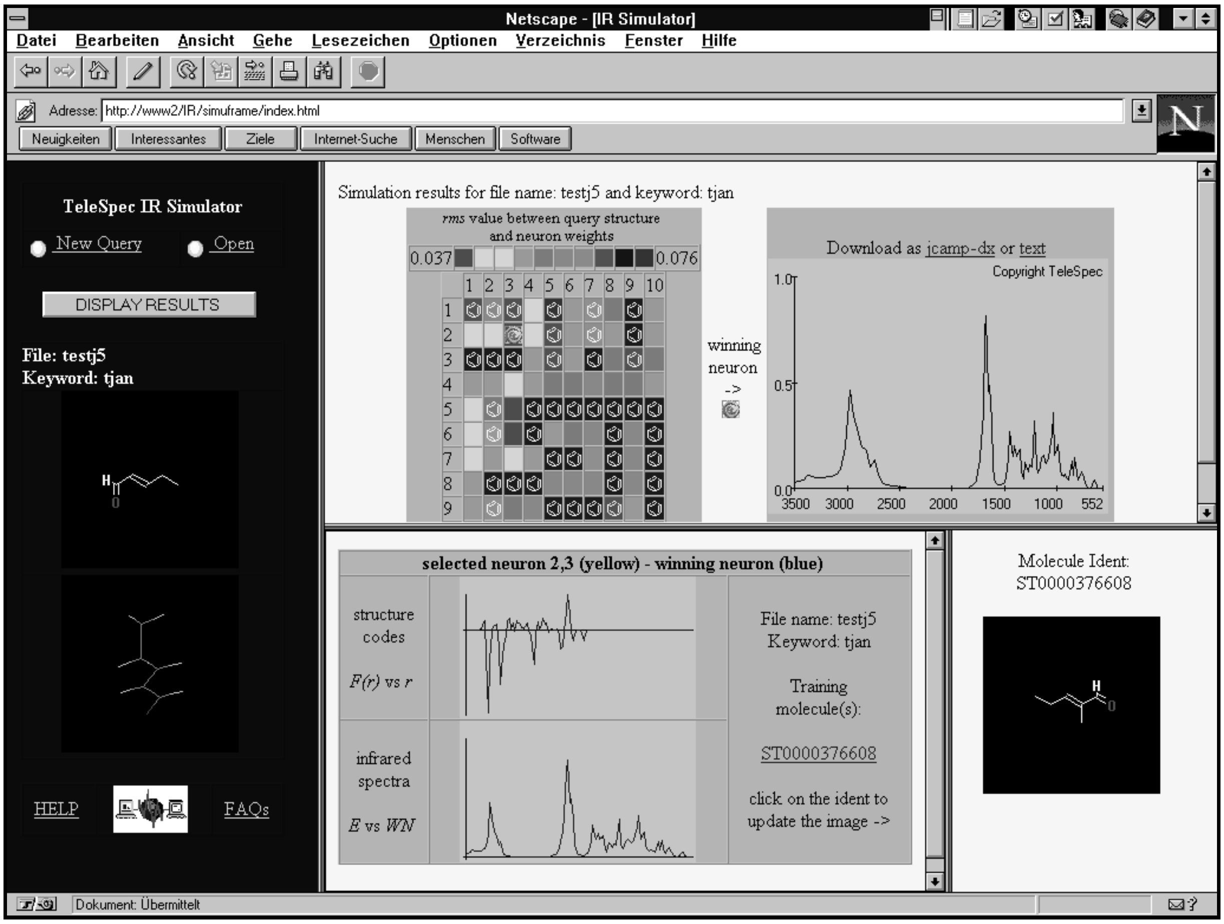


Abbildung 102: Bildschirm mit der TeleSpec Infrarotspektren - Simulationsseite und der Simulation des IR-Spektrums von Pent-2-en-al. Zur Simulation wurde das Neuron (2,3) genutzt, dem als Trainingsmolekül 2-Methylpent-2-enal assoziiert wurde.

Die deutlich größere strukturelle Verwandtschaft zwischen der Anfragestruktur Pent-2-enal und dem Trainingsmolekül 2-Methylpent-2-enal (vgl. Abbildung 102) als beim Simulationsbeispiel von Cyclobutanon, läßt bessere Chancen für eine gute Simulation erwarten.

7 QSAR - die Klassifizierung von Dopamin D1/D2 Agonisten

Das Ziel der Aufstellung quantitativer Struktur - Wirkungsbeziehungen (QSAR engl. Quantitative Structure Activity Relationships) für die biologische Wirkung von Stoffen ist die Vorhersage der biologischen Wirkung bzw. Aktivität von neuen Verbindungen. Kann man ferner Vorhersagen über die Aufnahme und Verteilung eines Stoffes im Körper treffen, ist eine Vorentscheidung möglich, ob eine Verbindung ein potentieller neuer Wirkstoff sein könnte oder nicht, ohne daß die Verbindung synthetisiert werden muß. Quantitative Struktur - Wirkungsbeziehungen erlauben zu dem einen Einblick in die komplexen Beziehungen von Struktur und biologischer Aktivität.

7.1 Die Vorbedingung

Die Vorbedingung aller Untersuchungen im QSAR-Bereich ist, daß der Zusammenhang zwischen der untersuchten Eigenschaft bzw. Wirkung und der Struktur derselbe ist. Dies ist bei vielen Medikamenten trotz gleichem makroskopischem Effekt keineswegs selbstverständlich. Meßbare Auswirkungen von Arzneimitteln, wie antibiotische Wirkung, Entzündungshemmung Fiebersenkung usw. können, aber müssen nicht auf demselben chemisch-biologischen Mechanismus beruhen. Ein chemisch-biologischer Mechanismus der Wirkung eines Arzneimittels, d.h. definierte Wirkungen des Arzneimittels und eventuell wirksamer Metaboliten an definierten Wirkorten, im besten Fall in bestimmten Bindungstaschen mit definierter Position von exakt charakterisierten Proteinen, soll Wirkprinzip heißen. Wie wichtig, aber auch wie schwer, die genaue Definition eines Wirkprinzips ist, sollen die folgenden Beispiele zeigen. Die Zahl der Antibiotika ist hoch, damit würden sich Antibiotika eigentlich für statistische Untersuchungen zwischen Struktur und Wirkung anbieten. Doch das als zweites entdeckte und wohl bekanntestes Antibiotikum Penicillin zeigt auch gleich die Problematik. Penicillin wirkt nur bei gram positiven Bakterien durch Hemmung eines zum Aufbau der Bakterienwand benötigten Enzyms. Seine Wirkung nun mit anderen Antibiotika vergleichen zu wollen ist nur dann sinnvoll, wenn man diese in bezug auf die Hemmung dieses Proteins vergleicht. Damit ist aber der Vergleich mit anderen Antibiotika-Klassen, wie z.B. den Tetracyclinen, deren Wirkung auf der Blockade der bakteriellen ribosomalen Proteinsynthese[91] beruht, ausgeschlossen.

Selbst bei so einfachen Verbindungen, die gleichzeitig strukturell so ähnlich sind, wie Salicylsäure und Acetylsalicylsäure und die auch ähnliche Wirkungen haben, kann sich das Wirkprinzip unterscheiden. Salicylsäure hemmt die Cyclooxygenase COX durch kompetitive Verdrängung des natürlichen Substrats Arachidonsäure. Acetylsalicylsäure hingegen acetyliert die Aminosäure Serin-530 der Cyclooxygenase selektiv und verschließt damit die Bindungstasche für die Arachidonsäure dauerhaft. Nun gibt es aber eine Cyclooxygenase-Mutante bei der Serin-530 durch Alanin ersetzt ist. Diese Mutante ist enzymatisch voll aktiv, da das Serin-530 nicht zum aktiven Zentrum gehört. Diese wird durch Salicylsäure gehemmt (kompetitive Verdrändung), aber nicht durch Acetylsalicylsäure, da die Möglichkeit zur Acetylierung fehlt.[91]

7.2 Klassifizierung von Dopamin D1/D2 Agonisten

Dopamin-Rezeptoren und ihr komplexes Wechselwirkungssystem sind für Störungen anfällig. Schwerste neurologische Krankheiten wie das Parkinson-Syndrom (Schüttellähmung) und die Schizophrenie beruhen auf Störungen des Dopamin-Rezeptorsystems. Inzwischen wurden mit Hilfe der Molekularbiologie mindestens 18 verschiedene Dopaminrezeptoren im menschlichen Genom nachgewiesen.[92] Allerdings können Dopaminrezeptoren bezüglich der Gabe von hemmenden bzw. aktivierenden Substanzen, Antagonisten und Agonisten nach wie vor in die Klassen D1 und D2 eingeteilt werden, da bisher selektive Antagonisten bzw. Agonisten fehlen, die innerhalb dieser Klassen unterscheiden.[i] Die Klasse der Dopamin-D1-Rezeptoren besteht nach Niznik et al. (92) aus Dopamin-D_1- und -D_5-Rezeptoren, während die Klasse der Dopamin-D2-Rezeptoren aus den D_2(lang und kurz)-, D_3- sowie den D_4-Rezeptoren besteht.

Während es bei der Schizophrenie-Behandlung im wesentlichen um eine Blockade der Dopaminrezeptoren und insbesondere der D2-Rezeptoren geht, ist das Parkinson-Syndrom im wesentlichen auf einen Mangel an Dopamin im Gehirn zurückzuführen, weshalb Dopamin-Agonisten zur Behandlung eingesetzt werden und zwar insbesondere Dopamin-D1-Agonisten. Dopamin-D2-Agonisten werden nur zum Teil zur Behandlung des Parkinson-Syndroms eingesetzt. Andere Krankheiten wie Bluthochdruck, Überdruck im Auge (Glaukom oder Grüner Star) sind meistens die Therapieziele bei der Gabe von Dopamin-D2-Agonisten. Bedenkt man, daß beim gesunden Menschen alle Dopamin-Rezeptoren von Dopamin aktiviert werden, bei der Behandlung von Krankheiten aber gerade die selektive Stimulation bzw. Hemmung der zwei Dopamin-Rezeptorklassen notwendig ist, wird die Herausforderung einer Klassifikation

[i] Gilt mit Ausnahme des zur Dopamin-D2-Rezeptorklasse gehörenden Dopamin-D3-Rezeptors. Für diesen Rezeptor gibt es einigermaßen selektiv bindende Wirkstoffen.

von Dopaminrezeptor-Agonisten in D1- und D2-Agonisten deutlich. So führt z.B. die Gabe von Dopamin[i] bei Parkinson-Kranken zum Rückgang der Lähmungserscheinungen, bei Überdosierung aber auch zu Halluzinationen und Verwirrung.

7.2.1 Die Strukturen und Daten für die Klassifizierung von Dopaminrezeptor-Agonisten

Die Strukturen und ihre Klassifikation für diese Untersuchung stammen aus der MDDR-3D Datenbank von MDL.[93] Bei der Suche in dieser Datenbank wurden 22 Dopamin-D1-Agonisten und 67 Dopamin-D2-Agonisten gefunden. Viele dieser Verbindungen enthalten das Dopamin, **D1,** als Substruktur. Aber es gibt auch zahlreiche Beispiele in denen die Substruktur des Dopamins maskiert wurde oder ganz fehlt (siehe z.B. **D4**).

---

[i] In Form der Aminosäure L-Dopa, die die Blut-Hirn-Schranke mit Hilfe des Aminosäuretransport-Apparates der Hirnzellen überwindet.

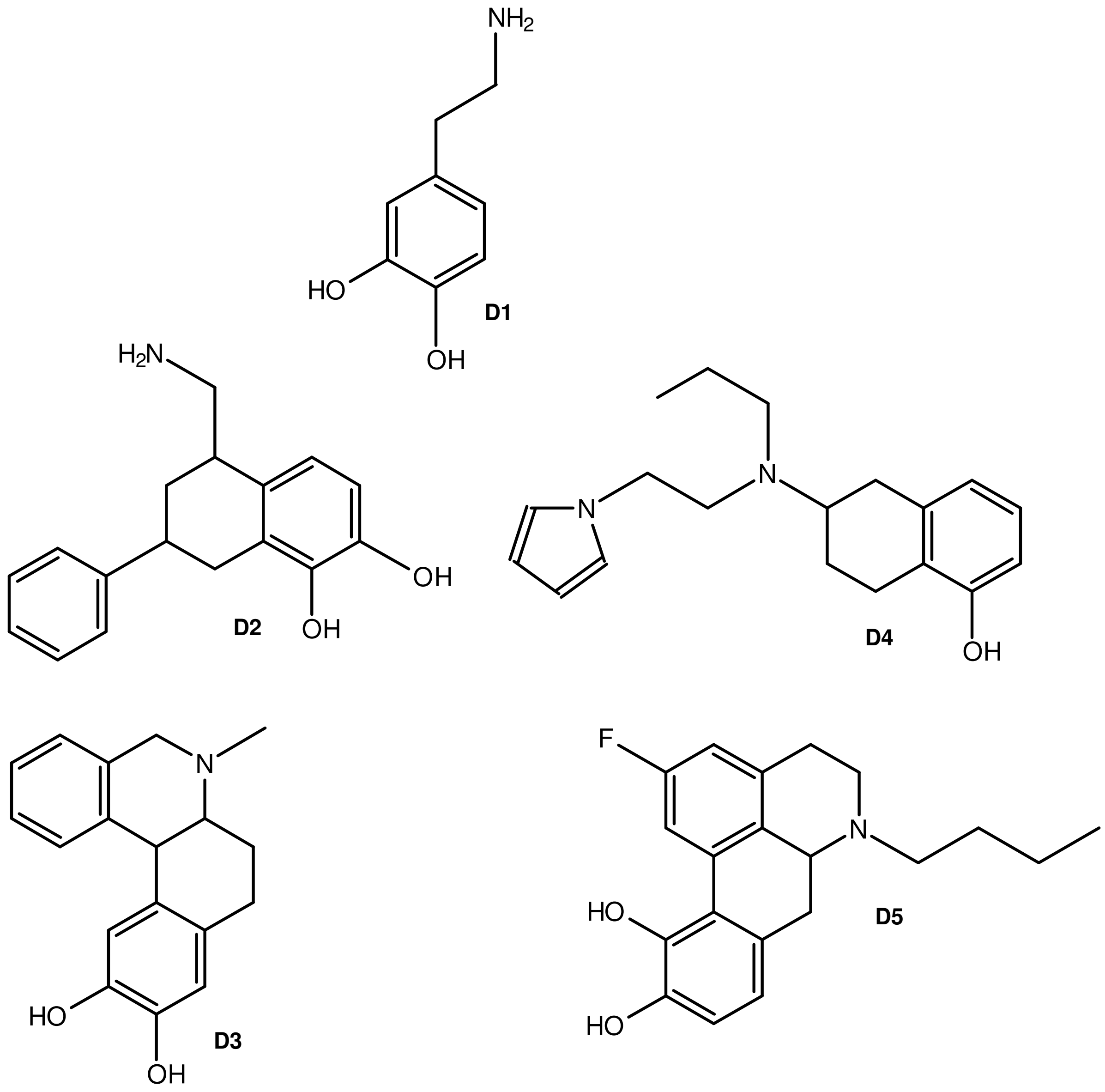


Abbildung 103: Dopamin (D1) und je zwei Dopamin-D1-Agonisten (D2-D3) und Dopamin-D2-Agonisten (D4-D5).

7.2.2 Die Codierung

Zur Codierung der 89 Dopamin-Agonisten wurden die folgenden Schritte durchgeführt:

Generierung der 3D-Struktur mit dem Programm CORINA[58]

- Berechnung der partiellen Atomladungen nach der PEOE-Methode von Gasteiger et al..[63-65]

- Codierung der Dopamin-Agonisten mit dem 3D-MoRSE unter Verwendung der folgenden Parameter n=32, s_{max} = 31 Å, $A_i = q_{tot,i}$.

- D1-Aktivität wurde mit 0 codiert, D2-Aktivität mit 1.

7.2.3 Das CPG-Netz zur Klassifizierung

Zur Klassifizierung wurde ein planares neuronales Counterpropagation-Netz mit 10 x 10 Neuronen verwendet. Das Training des Netzes erfolgte unbeaufsichtigt. Ziel war die räumliche Trennung der Dopamin-Agonisten in D1- und D2-Agonisten auf der Netzwerkoberfläche und die damit verbundene Klassifizierung der Agonisten, nicht jedoch eine Vorhersage der D1- oder D2-Aktivität für andere Dopamin-Agonisten. Aus diesem Grund konnte auch auf eine Aufteilung des Datensatzes in einen Trainings- und Testdatensatz verzichtet werden.

7.2.4 Ergebnisse der Klassifizierung von Dopamin-D1/D2-Agonisten

Die numerische Auswertung des trainierten CPG-Netzes ergibt für 88 Verbindungen eine korrekte Klassifikation, nur ein Dopamin-D2-Agonist wird fälschlicherweise als D1-Agonist klassifiziert. Wie die weitere Diskussion zeigen wird, ist die Fehlklassifikation nicht schwerwiegend.

Interessant ist die Verteilung der Werte der Ausgabeschicht im trainierten CPG-Netz im Zusammenhang mit der Assoziation der D1- und D2-Agonisten an die einzelnen Neuronen im Rahmen des Erinnerungstestes. Abbildung 104 zeigt, wie die Verteilung der Werte innerhalb einer Schicht von Gewichten eines CPG-Netzes, als Karte dargestellt werden kann. Abbildung 105 zeigt die Karte der Ausgabeschicht des trainierten CPG-Netzes sowie die Zuordnung der einzelnen Dopamin-D1/D2-Agonisten zu den Neuronen des Netzes.

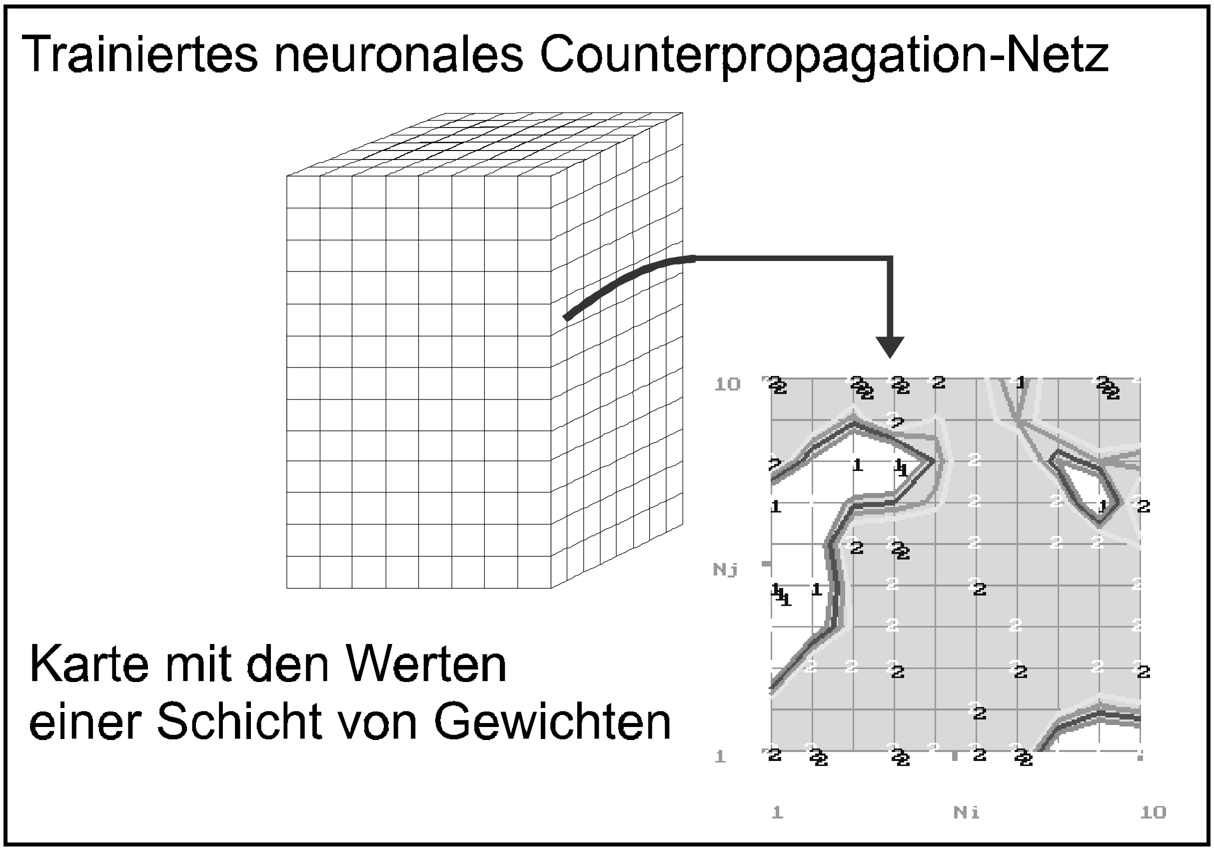


Abbildung 104: Die Karte der Schicht eines CPG-Netzes

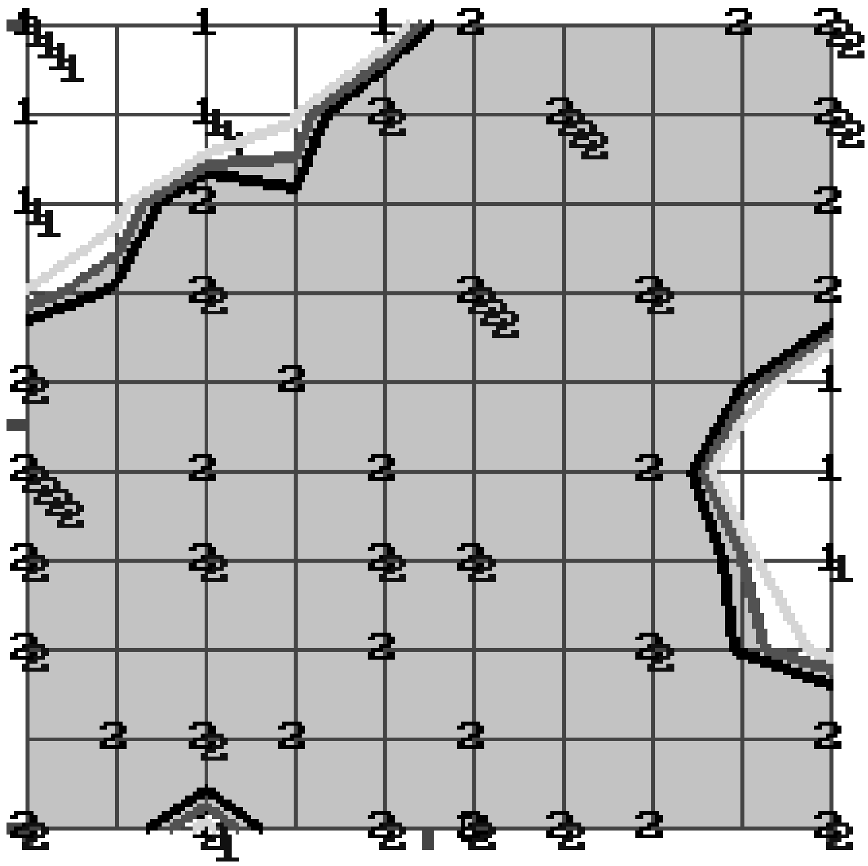
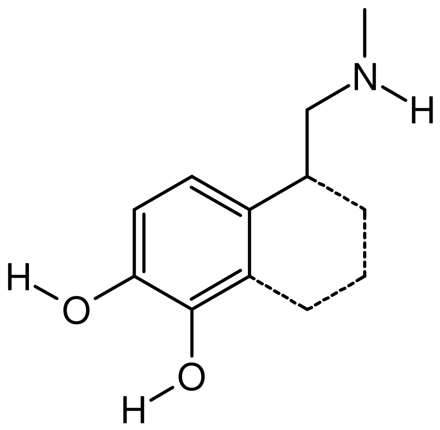


Abbildung 105: Karte der Ausgabeschicht des trainierten CPG-Netzes zur Klassifikation von Dopamin-D1/D2-Agonisten in Verbindung mit der Zuordnung der Verbindungen zu den einzelnen Neuronen. Dabei steht eine 1 für einen D1-Agonist und die 2 für einen D2-Agonisten. Die Einfärbung der Neuronen erfolgte gemäß ihrem Gewicht in der Ausgabeschicht. War dies höher als 0.5 und damit näher an 1 als dem Wert für D2-Agonisten, wurde das Neuron grau eingefärbt; unter 0.5 blieb das Neuron weiß. Die Linien grenzen die Bereiche voneinander ab.

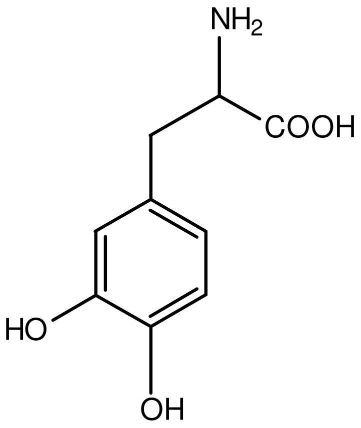
Abbildung 105 zeigt ganz klar einen zusammenhängenden Bereich (hellgraue Fläche) für die 67 Dopamin-D2-Agonisten des Datensatzes sowie zwei Bereiche für Dopamin-D1-Agonisten. Analysiert man diese Bereiche, so finden sich alle 16 Dopamin-D1-Agonisten mit der in Schema 9 dargestellten Substruktur in dem größten Cluster von D1-Agonisten in der linken oberen Ecke des CPG-Netzes.



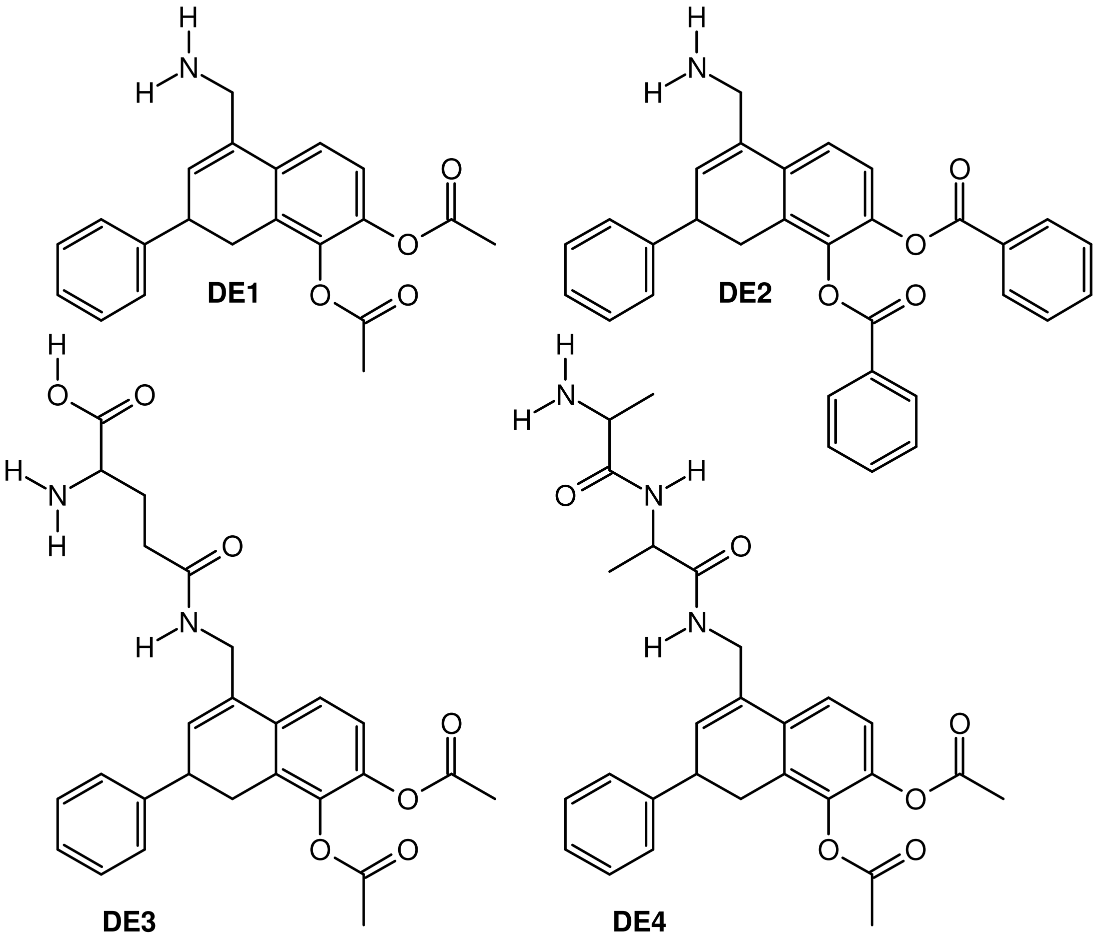
Schema 9: Die Substruktur der 16 Dopamin-D1-Agonisten aus dem großen Cluster in der linken oberen Ecke.

Der zweite Cluster an Dopamin-D1-Agonisten enthält die in Schema 11 dargestellten Verbindungen. Ihnen ist, im Gegensatz zu allen anderen Dopamin-D1-Agonisten des Datensatzes, eine Veresterung der normalerweise freien phenolischen OH-Gruppen der Dopaminsubstruktur gemeinsam, die entweder mit Essigsäure oder mit Benzoesäure erfolgte. Damit zeigt das Netz ganz deutlich die grundverschiedene Struktur dieser vier Dopamin-D1-Agonisten. Auch das Wirkprinzip der vier Dopamin-D1-Agonisten wird sich vermutlich von dem der anderen Dopamin-D1-Agonisten unterscheiden. So darf vermutet werden, insbesondere aufgrund der Aminosäureketten von DE3 und DE4, daß hier ähnlich wie bei dem ältesten Parkinson-Therapeutikum der Aminosäure L-Dopa, erst in der Hirnzelle nach partiellen enzymatischem Abbau, der Wirkstoff freigesetzt wird, und die hier abgebildeten Agonisten nur sogenannte pro-Drugs sind, aus denen erst im Körper der eigentliche Wirkstoff entsteht. Ein wichtiges Problem beim Design von Wirkstoffen, die an Rezeptoren in den Nervenzellen des Gehirns binden sollen, ist die Überwindung der Blut-Hirn-Schranke in Form einer Membran. Diese verhindert die Diffusion von polaren Stoffen, wie beispielsweise von Dopamin, in die Nervenzellen des Gehirns. Zu ihrer Überwindung gibt es im Prinzip zwei Möglichkeiten entweder der Wirkstoff oder seine Vorstufe ist hinreichend unpolar oder man nutzt einen der aktiven Transportmechanismen der Membran. Ein Essigsäure- und Benzoesäureester ist sicherlich unpolarer als eine freie OH-Gruppe insofern macht die Veresterung der zwei Hydroxygruppen der Dopaminsubstruktur in den vier Dopamin-D1-Agonisten DE1 - DE4 Sinn, um die Diffusion der Agonisten in die Nervenzellen zu ermöglichen. Die in DE3 und DE4 vorhandenen Aminosäureketten könnten noch einen weiteren Sinn haben, die Nutzung des aktiven Aminosäuretransportmechanismus der Nervenzellen. Die Nutzung dieses Transporters wurde mit dem ältesten Parkinsontherapeutikum der Aminosäure L-Dopa bereits eingeführt.

Der anschließende enzymatische Abbau setzt dann aus L-Dopa in der Hirnzelle das fehlende Dopamin frei.[94] Da auch die Hydrolyse von Estern bei der Verwendung von Wirkstoffvorstufen als Medikament ein bekanntes Prinzip ist [95], wird vermutlich bei den vier Dopamin-D1-Agonisten DE1 - 4 der eigentliche Wirkstoff erst nach der Überwindung der Blut-Hirn-Schranke in der Hirnzelle durch eine enzymatische Hydrolyse der Esterbindungen frei gesetzt. Bei den Dopamin-D1-Agonisten DE3 und DE4 dürften zudem die Aminosäureseitenketten wie bei L-Dopa enzymatisch abgebaut werden.



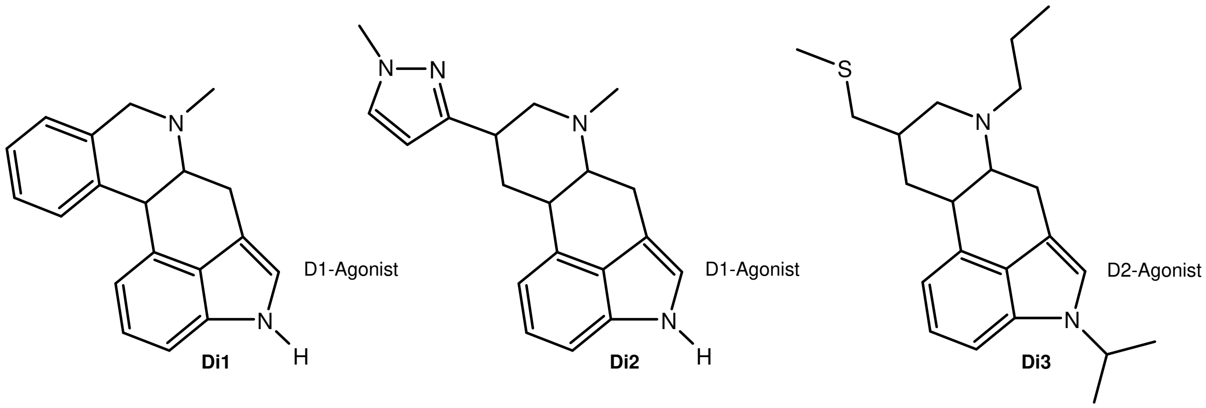
Schema 10: L-Dopa



Schema 11: Dopamin-D1-Agonisten aus dem Cluster am rechten Rand des CPG-Netzes. Auffällig ist die Veresterung der phenolischen OH-Gruppen der Dopaminsubstruktur.

Die abweichende Klassifikation des Dopamin-D2-Agonisten, **Di3**, der durch die gemeinsame Assoziation dieses Agonisten zusammen mit zwei D1-Agonisten an das Neuron (3,1)entsteht, wird bei einem Blick auf die Strukturen dieser Verbindungen sofort deutlich (Schema 12). Alle drei Agonisten sind Ergolinderivate ohne Sauerstoff am Benzolring der Dopaminsubstruktur, zudem ist der Stickstoff der Dopaminsubstruktur in allen drei Derivaten vollständig alkyliert. Dies ist für Dopamin-D1-Agonisten sehr ungewöhnlich, alle anderen Dopamin-D1-Agonisten weisen phenolische OH-Gruppen auf. Das Fehlen phenolischer OH-Gruppen ist hingegen für D2-Agonisten durchaus normal, so verfügen 40 der 67 D2-Agonisten des Datensatzes über keine phenolischen OH-Gruppen und 13 der 40 noch nicht einmal über ein Sauerstoffatom, während bei den D1-Agonisten nur 6 über keine phenolischen OH-Gruppen verfügen, was nur bei **Di1** und **Di2** nicht auf Maskierung der OH-Gruppe durch Veresterung beruht.

Hinzuzufügen ist, daß **Di1** und **Di3** in bezug auf ihre Wirkung nicht gerade typische Vertreter ihrer Klasse sind. So wirkt **Di3** vor allem als Entzündungshemmer und **Di1** zeigt Halluzinogene Wirkungen, was bei der nahen strukturellen Verwandtschaft mit LSD nicht verwundert. Bei **Di2** stellt sich die Frage, ob nicht die beiden Stickstoffatome des Pyrazol-Ringes als Ersatz für die phenolischen OH-Gruppen des Dopamins in Frage kommen, zumal sich der Stickstoff des tertiären Amins in derselben Entfernung zu den Stickstoffatomen des Pyrazol-Ringes befindet wie die Aminogruppe des Dopamins zu den zwei Hydroxygruppen desselben.



Schema 12: Die Dopamin-Agonisten von Neuron (3,1). Di3 ist hierbei die einzige Fehlklassifikation des Netzes.

7.2.5 Fazit der Klassifikation von Dopamin-Agonisten

Bei diesem Versuch zeigte sich, daß die Kombination von 3D-MoRSE Code und einem CPG-Netz in der Lage ist, den Datensatz aus 22 Dopamin-D1-Agonisten und 67 Dopamin-D2-Agonisten sinnvoll zu ordnen. Insbesondere die Verteilung der Dopamin-D1-Agonisten auf die

Neuronen des Netzes gab Aufschluß über die Zusammenhänge zwischen den einzelnen Verbindungen. Hierbei ist insbesondere die Abtrennung der Dopamin-Esterderivate zu erwähnen, die einen deutlichen Hinweis auf einen anderen Wirkmechanismus gab (Pro-Drug-Konzept). Bei der einzigen Fehlklassifikation im Rahmen dieses Versuch zeigte sich, daß die gemeinsame Ergolin-Substruktur von drei Dopamin-Agonisten erkannt und diese dementsprechend einem Neuron zugeordnet wurden. Dies ist angesichts der gemeinsamen Substruktur und der nicht gerade typspezifischen Wirkungen der drei Agonisten wahrscheinlich gerechtfertigt.

8 Zusammenfassung

Die 3D-Struktur eines Moleküls legt die Moleküleigenschaften fest. Der hier vorgestellte 3D-MoRSE Code erlaubt die Codierung der 3D-Struktur von Verbindungen in einem Vektor konstanter Länge. Die meisten Korrelationsverfahren setzen einen Vektor konstanter Länge voraus, insofern schafft der 3D-MoRSE damit die Voraussetzung für Korrelationen zwischen der 3D-Struktur und anderen Eigenschaften einer Verbindung (Spektren, biologische Aktivität). An der Simulation von Infrarotspektren und der Klassifizierung biologisch aktiver Substanzen wird gezeigt, wie der 3D-MoRSE Code genutzt werden kann.

Der 3D-MoRSE Code beschreibt die Beugung eines Elektronenstrahls an einem imaginären Molekularstrahl. Die 3D-Struktur der Moleküle wird dabei als starr angesehen und der Streuquerschnitt der Atome durch eine geeignete Eigenschaft beschrieben bzw. ersetzt. Damit beruht der 3D-MoRSE Code auf der Wierlschen Gleichung (10) für die Berechnung von theoretischen Beugungsmustern, in der die Ordnungszahl des Atoms zur Beschreibung des Streuquerschnittes genutzt wird. Im 3D-MoRSE Code wird anstelle der Ordnungszahl Z_i die Atomeigenschaft A_i genutzt. Je nach Anforderung kann als Atomeigenschaft A_i beispielsweise die partielle Atomladung $q_{tot,i}$, die Atompolarisierbarkeit α_i oder das Produkt aus Atommasse und der partiellen Atomladung genutzt werden. Die allgemeine Formel für den 3D-MoRSE Code gibt Gleichung (11) an:

$$I_M(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin sr_{ij}}{sr_{ij}} \tag{11}$$

Für die Codierung wird (11) diskretisiert, indem ein Bereich für den Beugungswinkelwert s festgelegt wird, für den $I_M(s)$ in äquidistanten Abständen berechnet wird. Dabei konnte festgelegt werden, wie sich die Anzahl der berechneten Werte und der gewählte Bereich auf die Beschreibung von Abständen im Molekül auswirken. Für den praktischen Einsatz der Codierung war ferner die Entwicklung eines Verfahrens zur Skalierung der einzelnen Werte des 3D-MoRSE Codes von Bedeutung, das eine datensatzunabhängige Skalierung der Werte auf denselben Wertebereich ermöglicht. Ohne Skalierung würden die Werte des 3D-MoRSE Codes mit $1/s$ gegen Null gehen.

Wie sich zeigte, ist der so erhaltene 3D-MoRSE Code zur Beschreibung der 3D-Struktur eines Moleküls gut geeignet. Bei Raumtemperatur und in Abhängigkeit von der Umgebung liegen die Moleküle konformativ flexibler Verbindungen jedoch in vielen verschiedenen Konformationen und damit unterschiedlichen 3D-Strukturen vor. Die Schwankungsbreite des Abstandes zweier Atome wird dabei in erster Näherung mit seiner mittleren Länge korreliert sein. Prinzipiell gibt es damit zwei Möglichkeiten zum Umgang mit konformativ flexiblen Verbindungen: erstens die verschiedenen Konformationen werden separat codiert oder zweitens im Code wird ein Mechanismus zur Repräsentation der konformativen Flexibilität eines Abstandes implementiert. Der 3D-MoRSE Code ist von der Konformation eines Moleküls abhängig. Die konformative Flexibilität von Molekülen wird dadurch berücksichtigt, daß die Bedeutung von räumlichen Atomabständen für die Werte des 3D-MoRSE Codes mit der Länge des Abstandes mit $1/r_{ij}$ abnimmt (vg. Gleichung 11). Damit wirken sich lange, mutmaßlich flexible Abstände weniger auf den Code aus als kurze, aber wenn sich zwei durch viele Bindungen getrennte Molekülteile aufgrund der Konformation nahekommen, wird dies deutlich im Code sichtbar sein.

Die Korrelation von 3D-Struktur und Infrarotspektrum, auf der Basis des 3D-MoRSE Codes, erlaubt die Vorhersage von Infrarotspektren auf der Basis der Molekülstruktur. Als Korrelationsverfahren werden im Rahmen dieser Arbeit neuronale Counterpropagation Netze genutzt. Diese neuronalen Netze können nach dem Training auf der Basis des 3D-Strukturcodes einer Verbindung das Infrarotspektrum simulieren. Die Kenntnis der experimentellen 3D-Struktur der Verbindung ist hierbei nicht Voraussetzung für die Vorhersage des Infrarotspektrums, da eine Konformation mit niedriger Energie, wie sie mit Hilfe des 3D-Strukturgenerators CORINA erhältlich ist, als 3D-Struktur genutzt werden kann. Für den 3D-Strukturgenerator CORINA ist die Bindungsliste eines Moleküls ausreichend, auf deren Basis dieser innerhalb kürzester Zeit eine 3D-Struktur generiert.

Die Simulation von Infrarotspektren eröffnet die Möglichkeit, die Infrarotspektroskopie als eine vielseitige und schnelle spektroskopische Methode mit einem großem Potential zur Erkennung von Substanzen wieder vermehrt in der Strukturaufklärung einzusetzen, indem das experimentelle Spektrum der unbekannten Substanz mit dem simulierten Spektrum des Strukturvorschlages verglichen wird. Die prinzipielle Nutzbarkeit des Verfahrens wird anhand eines Datensatzes von 871 mono-, di- und trisubstituierten Benzolderivaten gezeigt. Ergebnisse für einzelne Benzolderivate zeigten aber auch, daß die Verwendung eines fest definierten Satzes an Molekülen zu Problemen an den Rändern des Definitionsbereiches führt. Diese waren unter anderem Anlaß zur Entwicklung der anfrageorientierten Simulation, bei der auf der Basis des

3D-Strukturcodes der Anfragestruktur ein Trainingsdatensatz mit 50 ähnlichen Molekülen aus der Datenbank extrahiert wird. Mit dem Trainingsdatensatz wird dann ein neuronales Counter-propagation Netz für die Anfrage trainiert. Aus der Abfrage des trainierten Netzes wird das simulierte Infrarotspektrum für die Anfragestruktur erhalten.

Mit Hilfe der anfrageorientierten Simulation können Infrarotspektren für beliebige Substanzen simuliert werden, wenn ausreichend ähnliche Verbindungen in der Datenbank enthalten sind. Die Simulation der Infrarotspektren von 77 primären Aminen mit vielen verschiedenen weiteren Substituenten zeigt, daß die Simulationsmethode auch dann geeignet ist, wenn die Verbindungen intermolekulare Wasserstoffbrückenbindungen ausbilden können. Dies gelingt, weil aus experimentellen Daten gelernt wird.

Die Klassifizierung von Dopamin-D1/D2-Agonisten, zeigt wie diese biologisch aktiven Stoffe anhand ihres 3D-MoRSE Codes geordnet werden können. Die erhaltene Ordnung, bestehend aus einem großen Cluster von Dopamin-D2-Agonisten und einem größeren und zwei kleineren Clustern für Dopamin-D1-Agonisten, erlaubt sogar Rückschlüsse auf die Beziehung von Struktur und Wirkung bei den Dopamin-D1-Agonisten. So enthält der eine der kleineren Cluster für Dopamin-D1-Agonisten alle Verbindungen bei denen die zwei Hydroxygruppen der Dopamin-Substruktur verestert sind. Dies läßt im Zusammenhang mit Erkenntnissen aus der historischen Entwicklung von Dopamin-D1-Agonisten vermuten, daß diese Verbindungen sogenannte Pro-Drugs sind, die erst im Körper zum Wirkstoff reagieren, im Fall der Esterderivate durch enzymatische Esterhydrolyse.

Die Beispiele zeigen, daß der 3D-MoRSE Code eine sinnvolle Codierung der 3D-Struktur von Molekülen erlaubt. Die so unterschiedlichen Anwendungen des 3D-MoRSE Codes für die Korrelation von 3D-Struktur und Infrarotspektrum einer Substanz bzw. der 3D-Struktur von Dopamin-Agonisten mit ihrer biologischen Wirkungen lassen hoffen, daß der 3D-MoRSE Code für viele Aufgaben aus dem Bereich quantitativer Struktur - Eigenschafts- bzw. Wirkungsbeziehungen (QSPR/QSAR) eingesetzt werden kann.

9 Ausblick

Die vorgestellten Arbeiten gliederten sich schwerpunktmäßig in die Entwicklung einer 3D-Strukturcodierung, die Entwicklung einer Methode zur Simulation von Infrarotspektren sowie der Untersuchung quantitativer Struktur-Wirkungsbeziehungen auf der Basis des 3D-MoRSE Codes. Im Rahmen dieses Kapitels soll versucht werden, einen Blick über die hier vorgestellten Arbeiten hinaus, auf kommende Entwicklungen zu werfen.

9.1 Die Zukunft der 3D-Strukturcodierung

3D-Strukturcodierungen sind ein neues wichtiges Werkzeug zur Verarbeitung chemischer Informationen, denn Moleküle sind dreidimensional, ihre Wechselwirkungen beruhen auf ihrer dreidimsionalen Struktur. Die Anstrengungen, diese dem Chemiker verfügbar zu machen, zeigen Erfolge. 3D-Strukturgeneratoren, entwickelt zur Konvertierung ganzer Datenbanken mit Bindungslisten von Verbindungen in die korrespondierenden 3D-Strukturen, verlassen dieses Aufgabengebiet und werden als Teil von Programmpaketen für alle Chemiker verfügbar.[96] Gleichzeitig erleichtern graphische Oberflächen und immer höhere Rechenleistungen die Nutzung quantenmechanischer Verfahren zur Optimierung von 3D-Strukturmodellen, auch unter Berücksichtigung einer wässrigen Umgebung.[97] Zudem wachsen die Datenbanken mit experimentellen 3D-Strukturen. All dies erleichtert den Zugang zu 3D-Strukturen erheblich.

Die schnell wachsende Zahl verfügbarer 3D-Strukturen wird die Informations- und Datenverarbeitung in der Chemie erheblich beeinflussen. Da zum einen die Eigenschaften einer Verbindung von der dreidimensionalen Struktur des einzelnen Moleküls bestimmt werden, bilden die 3D-Strukturen eine wesentlich bessere Grundlage für QSAR/QSPR-Untersuchungen. Zum anderen stellt die wachsende Anzahl von 3D-Strukturen auch neue Anforderungen an die Handhabung. Bisher beruhte die Nutzung von 3D-Strukturen auf Vergleichen und der direkten Ableitung räumlicher Eigenschaften, wie etwa innermolekularer Distanzen. Vergleiche von 3D-Strukturen wurden bisher vor allem bei Datenbanksuchen nach ähnlichen Strukturen und gleichen Substrukturen vorgenommen, wobei diese Suchen je nach Sorgfalt des Verfahrens beliebig kompliziert sind, da es sich beim Vergleich der Atompositionen um ein NP vollständiges Problem handelt, das zudem bei konformativ flexiblen Molekülen für alle möglichen Kombinationen der Konformationen dieser Moleküle zu lösen wäre.

Die Aufgabe von 3D-Strukturcodierungen ist es, dieses kombinatorische Problem zu umgehen. 3D-Strukturcodierungen ermöglichen die Darstellung der 3D-Struktur in einem Vektor fester Länge, dessen Dimensionen eine definierte Bedeutung haben. Damit reduziert sich der Aufwand für den Vergleich von zwei 3D-Strukturen auf den Vergleich zweier Vektoren. Dies ist um Faktoren schneller als jeder Vergleich vollständiger 3D-Strukturen. Inwieweit mit Hilfe von 3D-Strukturcodes ein Vergleich der 3D-Strukturen konformativ flexibler Verbindungen und damit letztlich ihres Konformationsraums möglich ist, hängt davon ab, wie stark der Einfluß der Konformation auf den 3D-Strukturcode ist und wie drastisch sich die einzelnen Konformationsänderungen auf die 3D-Struktur auswirken. Bei einem Molekül, das aus einem großen starren Mittelstück und zwei kurzen flexiblen Seitenketten besteht, wird die Beschreibung des Konformationsraums durch einen 3D-Strukturcode für viele Anwendungen ausreichen, zumal sich die physikalischen Eigenschaften vieler Konformationen mit niedriger Energie kaum unterscheiden dürften. Ein einfaches Beispiel für eine solche Verbindung wäre 1,4-Diethylbenzol. Wirkt sich jedoch die konformative Flexibilität durch große Unterschiede in den 3D-Strukturen der Konformationen auf die interessierende Eigenschaft aus oder ist dies zu erwarten und wäre es zu beobachten, so sollte der verwendete 3D-Strukturcode auch deutlich von der Konformationen abhängen.

Jede Codierung der 3D-Struktur bedeutet jedoch gleichzeitig einen Verlust an Information. Damit stellt sich die Frage nach der Aussagekraft von 3D-Strukturcodierungen, die zumeist auf mathematischen Transformationen der Distanzmatrix beruhen. Prinzipiell ist es möglich, die Distanzmatrix verlustfrei zu transformieren, solange nach der Transformation die Anzahl der Werte mindestens genau so groß ist, wie die Hälfte der Werte in der Distanzmatrix. Somit muß der Code $x(x-1)$ Werte zur Codierung von x Atomen in einer Struktur enthalten.[i] Damit wären aber bereits für Propan, C_3H_8, 110 Codewerte erforderlich (mit $x = 11$, $11 * 10 = 110$), um die Distanzmatrix verlustfrei codieren zu können. Ein Code mit rund 100 Werten mag vielleicht noch sinnvoll sein, aber ein Strukturcode muß für alle Verbindungen die gleiche Länge haben, für Butan wären aber bereits 182 Werte notwendig und für Cholesterin, $C_{27}H_{46}O$, 5402 Werte. Da aber eine Strukturcodierung mit 100 Werten pro Molekül sicher noch sinnvoll ist, mit 5000 Werten aber nicht mehr, wird man für alle Moleküle, außer den einfachsten, einen Informationsverlust in Kauf nehmen müssen.

[i] Die Distanzmatrix ist eine symmetrische, quadratische Matrix. Die halbe Matrix enthält, abzüglich der Nullen auf der Diagonale, $x(x-1)/2$ Werte, da für jeden Wert noch das Atompaar gespeichert werden muß, ergeben sich $x(x-1)$ - Werte.

Trotz des Informationsverlustes kann ein 3D-Strukturcode sinnvoll sein, da er gegenüber einem 3D-Hashcode mehr Informationen enthält und zudem die gesamte Information des Moleküls in ihn eingeht. Gelingt es bei der Codierung zwischen relevanter und unwichtiger oder unsicherer Information zu trennen, kann ein 3D-Strukturcode die wesentlichen Informationen sehr kompakt und leicht vergleichbar darstellen. Damit dies gelingt, müssen die folgenden Bedingungen erfüllt sein:

1. Vorherige Auswahl und Beschränkung der Information über die Atome auf das wesentliche, beispielsweise wie im Fall der Infrarotspektroskopie auf die partielle Atomladung. Atome, die in ihrer chemischen, physikalischen und oder biologischen Rolle ähnlich sind, sollten auch ähnlich codiert werden.

2. Beschränkung der Distanzinformation auf das wesentliche. Wichtige Abstände sollten im Code betont werden. Zum Beispiel ist für das Infrarotspektrum die Distanz von Atomen mit großen partiellen Ladungen wichtiger, als die neutraler oder fast neutraler Atome.

3. Berücksichtigung der konformativen Flexibilität von Molekülen, wenn dies erforderlich ist. Die Schwankungsbreite eines intramolekularen Abstandes zwischen den einzelnen Konformationen wird in erster Näherung mit seiner mittleren Länge korreliert sein. Ein 3D-Strukturcode sollte in der Lage sein, dies wiederzugeben.

Der 3D-MoRSE Code erfüllt diese Bedingungen weitgehend. Insbesondere die Berücksichtigung von Atomparametern, wie der partiellen Atomladung, erleichtert die Erfüllung der Bedingungen 1 und 2 im besonderen Maße und hilft auch bei der Erfüllung der Bedingung 3, da Atomeigenschaften über die Bindungsordnung der beteiligten Atome mit der konformativen Flexibilität im Zusammenhang stehen. Ferner nimmt die Bedeutung langer und damit mutmaßlich konformativ flexibler Distanzen im 3D-MoRSE Code mit dem Faktor $1/sr_{ij}$ ab (vgl. Gleichung 11). Insofern eignet sich der 3D-MoRSE Code gut zur Codierung von Molekülen für eine Vielfalt von Aufgaben.

Einen Nachteil hat der 3D-MoRSE Code aber auch, er ist schwer zu interpretieren. Zwar arbeitet M. Hemmer[98] an einem Algorithmus zur allgemeinen Decodierung von 3D-Strukturcodes, aber dies ist noch kein Ersatz für eine direkte Interpretation einzelner Codewerte. Der von Steinhauer et. al. eingeführte Radialcode[29] oder die verwandten Distanzhistogramme[27,28], die einen direkten Rückschluß vom Codewert auf die intramolekulare Distanz erlauben, sind hier im Vorteil. Beide Codierungen lassen sich durch die im folgenden vorgestellten Maßnahmen noch weiter verbessern und um Möglichkeiten zur Berücksichtigung der konformativen Flexibilität von Abständen erweitern.

Eine einfache Verbesserung des Radialcodes, die analog zum 3D-MoRSE-Code die Möglichkeit zur Informationsdifferenzierung und Beschränkung gibt, ist der Ersatz der Ordnungszahl im Radialcode durch einen Atomparameter. Dies wurde erstmals im Rahmen des TeleSpek-Projektes bei der Codierung von 3D-Strukturen zur Simulation von Infrarotspektren genutzt.[90]

Eine andere Möglichkeit zur Verbesserung des Radialcodes bzw. der Distanzhistogramme liegt in der Gewichtung und Definition der Distanzintervalle. Bisher waren alle Distanzintervalle gleich lang und es wurden nur Distanzen bis zu einer maximalen Distanz *Dmax* berücksichtigt.

Das nur Distanzen bis *Dmax* berücksichtigt werden, wäre einfach durch die Öffnung des letzten Intervalls für alle intramolekularen Distanzen zwischen *Dmax* und unendlich zu korrigieren, sofern man davon ausgeht, daß allen Distanzen größer oder gleich *Dmax* dieselbe Bedeutung zukommt. Letzteres wird ab einem bestimmten Wert für *Dmax* immer der Fall sein. Gleiches dürfte für ein intelligent definiertes erstes Intervall gelten. Abstände unterhalb des kürzesten bekannten Bindungsabstandes *Dmin* sind nicht sinnvoll. Ein Intervall von minus unendlich bis *Dmin*, würde alle offensichtlich falschen Distanzen sammeln. Die Belegung dieses Intervalls würde einen offensichtlicher Fehler sofort anzeigen, gleich aus welcher Quelle der Fehler stammt, sei es nun der Datenbasis, der Software oder der Hardware.[i]

Im Bereich zwischen *Dmin* und *Dmax* sind die Codierungsintervalle nun entsprechend den Anforderungen der Benutzer zu definieren. Sind beispielsweise funktionelle Gruppen von Interesse, so ist eine genaue Unterscheidung der Bindungsabstände wichtig, da sich die Bindungslängen innerhalb der funktionellen Gruppen spezifisch unterscheiden. So beträgt die Länge der CC-Doppelbindung durchschnittlich 134 pm, im Benzol liegt die CC-Bindungslänge bei 139 pm und die durchschnittliche Länge einer CC-Einfachbindung wird mit 154 pm angegeben.[99] Dabei hängt die genaue Bindungslänge von der chemischen Umgebung der Bindung ab, so ist beispielsweise die CH-Bindung in Cyclopropan mit 108.9 pm zwei pm kürzer als in Cy-

[i] Die Hardware ist eine nicht zu vernachlässigende Fehlerquelle, denn gerade bei der Behandlung von Gleitkommazahlen und damit aller realer Zahlen finden sich immer wieder Fehler in dem Microcode der Prozessoren, wie beispielsweise der legendäre Pentium Bug.

clobutan mit 110.9 pm[100]. Deshalb sollte der Abstand zwischen zwei Werten in diesem Bereich bei 1-2 pm liegen, wenn funktionelle Gruppen für die Codierung wichtig sind.

Den kürzesten Atomabstand in einem neutralen Molekül hat mit 74 pm sicherlich molekularer Wasserstoff, gefolgt von 92 pm bzw. 96 pm für Fluorwasserstoff und OH-Einfachbindungen. Da die 3D-Struktur des Wasserstoffmoleküls für Korrelationsuntersuchungen uninteressant ist, kann *Dmin* auf 90 pm entsprechend 0.9 Å festgelegt werden.

Dmax sollte so festgelegt werden, daß für eine Codierung deren Schwerpunkt auf den funktionellen Gruppen der Moleküle liegt, die größten Abstände zwischen zwei Atomen der größten vorkommenden funktionellen Gruppe noch erfaßt werden können. Ist dies beispielsweise eine Carboxylgruppe mit einer primären Aminogruppe in α-Position (vgl. α-Aminosäuren), so beträgt der Abstand zwischen dem Carbonylsauerstoff und den Wasserstoffen am Stickstoff je nach Konformation 260 - 390 pm. Damit wäre ein *Dmax* von 400 pm bzw. 4.0 Å für diese Aufgabe vermutlich eine gute Wahl.

Soll die Codierung aber beispielsweise für die Suche nach einem Pharmakophor dienen, sollte *Dmax* sicherlich den größten Abstand im Pharmakophor einschließen und falls dieser nicht bekannt ist, etwas größer sein als der größte Abstand zwischen zwei funktionellen Gruppen bekannter Wirkstoffe, die die gewünschte Wirkung zeigen. Hierbei ist es aber durchaus zu hinterfragen, welche Bedeutung dem exakten Abstand zukommt, da dieser bei einer flexiblen Konformation zwischen dem freien Wirkstoff und der aktiven Konformation am Rezeptor recht unterschiedlich sein kann.

9.1.1 Berücksichtigung der Flexibilität intramolekularer Abstände im Rahmen einer 3D-Strukturcodierung

Aus dem Vorstehenden heraus wird die folgende Modifikation des Radialcodes vorgeschlagen um der Tatsache Rechnung zu tragen, daß die Schwankungsbreite des Abstandes zweier Atome in erster Näherung mit seiner Größe korreliert sein wird:

$$R(r) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} A_i A_j * e^{-B(1-fr/Dmax)(r-r_{ij})} \qquad Dmin < r < Dmax \qquad (22)$$

R	Radialcode mit Berücksichtigung konformativer Flexibilität
$A_{i,j}$	passende Atomeigenschaft

Dmax maximaler zu berücksichtigender Abstand

Dmin minimaler zu berücksichtigender Abstand

B Unsicherheitsfaktor für Abstände auch Temperatur- oder Glättungsparameter

f Skalierungsfunktion für den *B*-Parameter in Abhängigkeit von *r*

r Distanz

r_{ij} Abstand der Kernpositionen der Atome *i* und *j*

Da eine 3D-Strukturcodierung einen Vektor fester Länge erfordert, müssen definierte Werte für *r* festgelegt werden. Entsprechend dem vorstehenden sollte der kleinste Wert für die Distanz *Dmin* 90 pm betragen. Zur Codierung der einzelnen Bindungsabstände sollten im folgenden die Abstände der einzelnen Werte nicht weiter als 1 pm auseinanderliegen, wenn eine genaue Unterscheidung von funktionellen Gruppen erforderlich ist, entsprechend mehr, wenn Bindungen nur grob klassifiziert werden sollen. Bei größeren Abständen sind die genauen Werte aufgrund der konformativen Flexibilität weniger interessant. So schwankt der kürzeste Abstand zwischen zwei Wasserstoffatomen in Ethan je nach Konformation zwischen 231 und 252 pm. Deshalb sollten hier beispielsweise Werte mit Abständen von 10-20 pm ausreichen. Da ferner *B* mit zunehmenden *r* sinkt, bietet es sich an, die Abstände von *r* oberhalb eines Grenzwertes auf einen Prozentsatz des Wertes von *r* festzulegen.

Aus dem vorstehenden wird folgende Funktion für die Abstände der einzelnen Codewerte Δr vorgeschlagen:

$$\Delta r = \Delta min + p_f r + \frac{p_s r}{1 + \exp(-(r - r_w) / f_s)} \tag{23}$$

r Distanz mit $Dmin < r < Dmax$ für Gleichung (23)

Δr Abstand der Werte r_x und r_{x+1}

Δmin minimaler Abstand für die ersten Werte

p_f erster prozentualer Faktor für die langsame Steigerung der Werte von Δr im Anfangsbereich

p_s zweiter prozentualer Faktor für die schnelle Steigerung der Werte von Δr bei höheren Distanzen

r_w Wendepunkt beim Übergang von ersten zum zweiten prozentualen Faktor

f_s sigmoidaler Faktor bestimmt die Steilheit des Übergangs vom ersten zum zweiten prozentualen Faktor

Die folgenden Abbildungen zeigen das Verhalten von *r* für 128 Werte. Deutlich ist das exponentielle Wachstum von *r* zu sehen, das auf einer Wachstumsrate von maximal $\Delta r = p_f + p_s$ für die letzten Werte beruht.

Für die Abbildung 106 - Abbildung 110 gilt (alle Angaben in pm):

- $Dmin = 90$
- $\Delta min = 0.5$
- $Dmax = 5000$
- Die Tiefstellung der Indizes von p_f, p_s, p_t, f_s und r_w unterbleibt aus technischen Gründen.

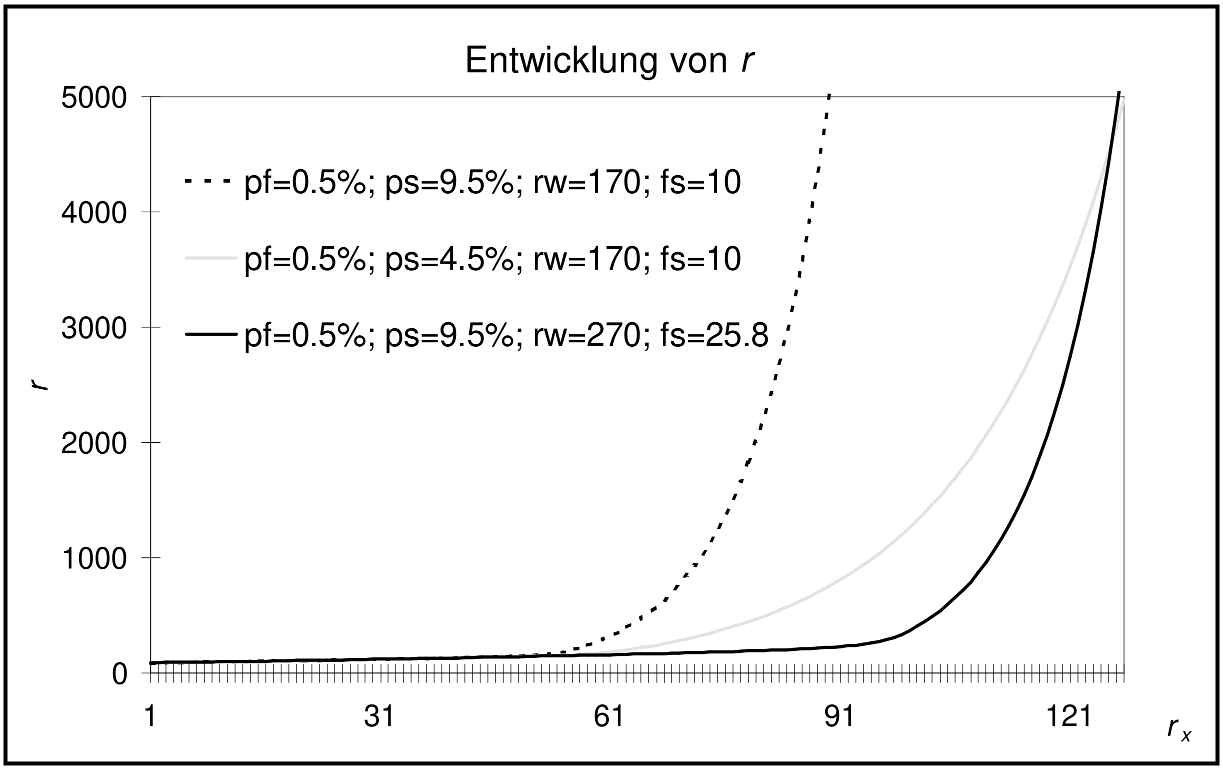


Abbildung 106: Die Entwicklung von *r* in Abhängigkeit der für *Δr* gewählten Funktion. Deutlich zu sehen ist das exponentielle Wachstum von *r.* Auf Darstellung von *r* > 5000 pm wurde verzichtet, da *Dmax* 5000 pm betragen soll.

Die erste Kurve mit $r_w = 170$ und $p_s = 9.5\%$ übersteigt deutlich den Wert von *Dmax* mit 5000 pm. Dies mahnt, daß der erste und der letzte Codewert dazu genutzt werden sollten um Distanzen kleiner als *Dmin* bzw. größer als *Dmax* zu summieren, da sie Hinweise auf Fehler im Codierungsprozeß bzw. in der Datenbank liefern können.

Abbildung 106 zeigt deutlich den Einfluß sowohl des Wendepunktes für den Übergang zwischen Wachstumsraten, r_w, als auch der Summe von p_f+p_s auf den Verlauf der Wachstumskurven. So kann mittels r_w der Beginn der exponentiellen Wachstumsphase verschoben werden, während die Summe der Wachstumsraten die Steilheit des exponentiellen Wachstums beeinflußt. Abbildung 107 zeigt diesen Effekt noch einmal deutlicher.

Zu beachten ist auch das lineare Wachstum von *r* bis ca. 140 pm, das auf *Δmin* beruht. Erst oberhalb dieses Wertes ist exponentielles Wachstum zu erkennen, wobei die Geschwindigkeit und die Stärke des Wachstums von p_f, r_w und f_s abhängt. So findet sich mit p_s = 9.5%, r_w = 170 pm und f_s = 10 pm zwischen dem 50. und 60. Codewert ein scharfer Übergang zwischen linearem und exponentiellen Wachstum statt, während dieser sich mit p_s = 9.5%, r_w = 270 pm und f_s = 25.8 pm zwischen dem 50 und dem 110 Codewert erstreckt.

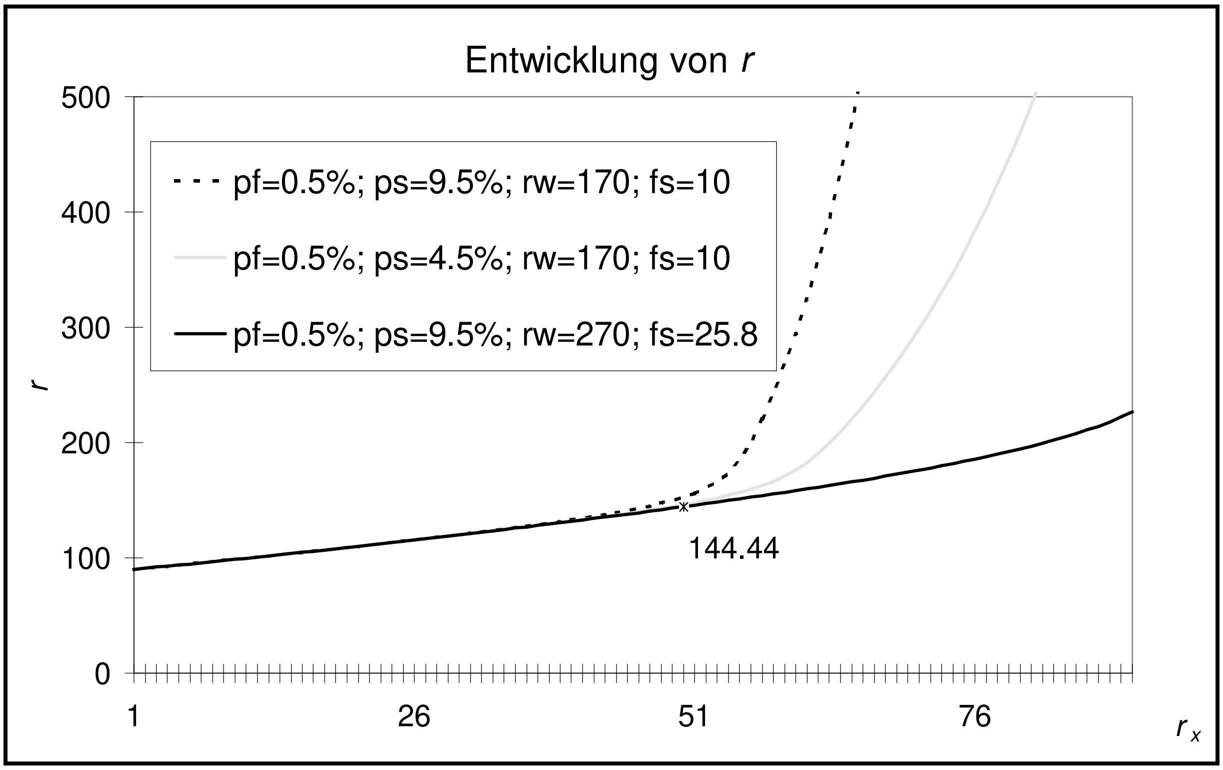


Abbildung 107: Die Werte von *r* für die ersten 90 Radialcodewerte. Deutlich zu sehen ist, daß bis etwa 144 pm das lineare Wachstum in Abhängigkeit von *Δmin* immer überwiegt.

Viel deutlicher als die zwei vorangegangenen Abbildungen zeigt Abbildung 108 den sigmoiden Übergang von der einen zur anderen Wachstumsrate, wenn man das prozentuale Wachstum der Radialcodewerte aufträgt.

Zu erkennen ist auch, daß nach dem Übergang zwischen Wachstumsraten sich diese additiv verhalten. Die Wachstumsraten bei r_{128} betragen 0.998, 0.615 bzw. 0.982.

Abbildung 108 gibt auch einen Eindruck davon wie sich der Übergang zwischen den zwei Wachstumsraten mittels der Werte von r_w und f_s beeinflussen läßt. So verschiebt r_w den Mittelpunkt des Übergangs zwischen den Wachstumsraten parallel zur x-Achse, während f_s die Steigung des Übergangs beeinflußt (allerdings in Abhängigkeit von r).

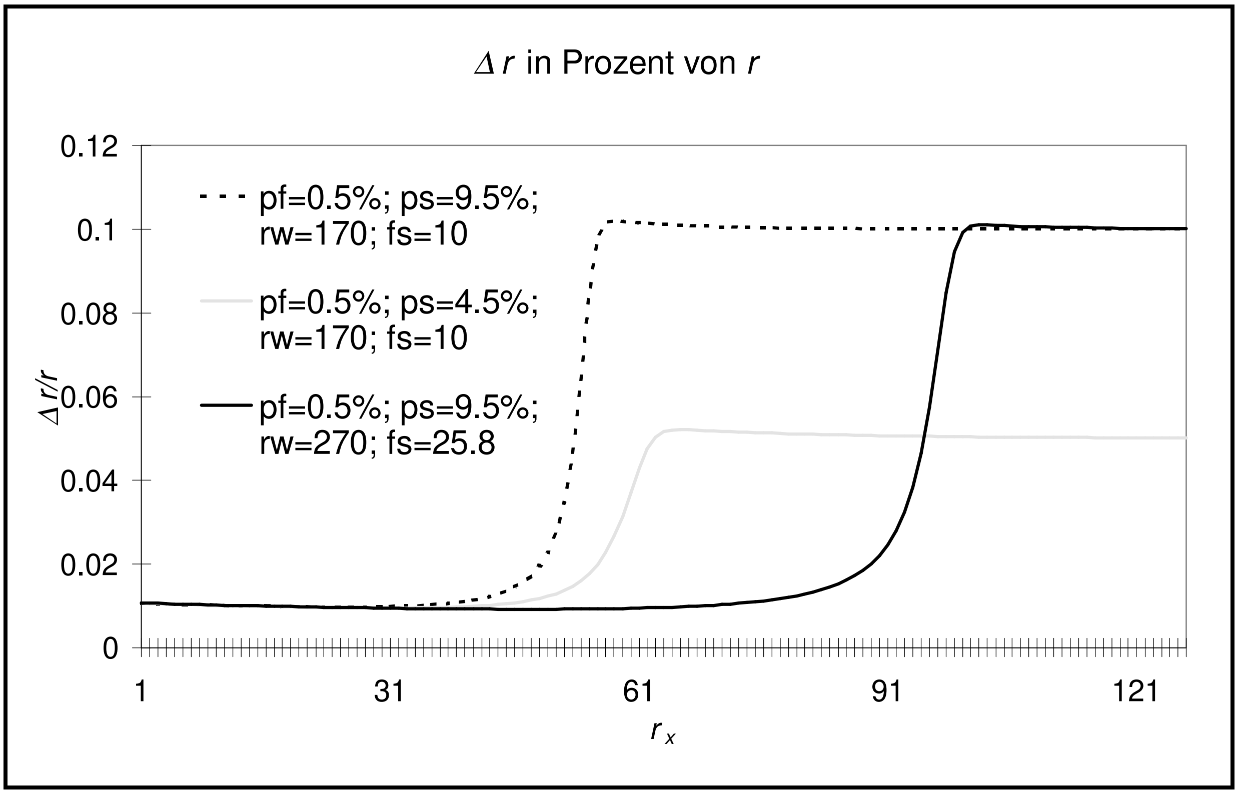


Abbildung 108: Die prozentuale Wachstumsrate der Radialcodewerte in Abhängigkeit von p_s, r_w und f_s an drei Beispielen. Die Schärfe des oberen Übergangs ist eine Ursache des linearen Wachstumsfaktors Δmin in Höhe von 0.5 pm.

Gleichung (23) gibt einen Hinweis darauf, wie das Wachstum von r beeinflußt werden kann. Modifikationen an Gleichung (23) bieten einen weiten Spielraum zur Wahl des Wachstumsverlaufs von r. Entscheident ist der verwendete sigmoide Funktionsteil, wie er auch als Transfer-Funktion in neuronalen Backpropagation Netzen genutzt wird,[101] erlaubt er doch den stetigen Übergang zwischen den Wachstumsraten. Die Gleichungen (24) und (25) sollen zeigen, wie (Gleichung 23) für drei verschiedene exponentielle Wachstumsraten erweitert, bzw. auf drei lineare Wachstumsraten reduziert werden kann. Selbstverständlich sind darüber hinaus auch Kombinationen aus linearen und exponentiellen Wachstumsraten möglich und evtl. sinnvoll.

$$\Delta r = \Delta min + p_f\, r + \frac{p_s r}{1 + \exp(-(r - r_{w1,2}) / f_{s1,2})} + \frac{p_t r}{1 + \exp(-(r - r_{w1,3}) / f_{s2,3})} \qquad (24)$$

$$\Delta r = \Delta min + \frac{\Delta w1}{1 + \exp(-(r - r_{w0,1}) / f_{s0,1})} + \frac{\Delta w2}{1 + \exp(-(r - r_{w1,2}) / f_{s1,2})} \qquad (25)$$

p_t dritte prozentuale Wachstumsrate

$f_{sx,y}$ Steilheitsfaktor für den x sigmoiden Übergang

$r_{wex,y}$ Wendepunkt des sigmoiden Übergangs

$\Delta w1$, $\Delta w2$ lineare Wachstumsraten

Abbildung 109 zeigt den Verlauf von r und der Wachstumsrate $\Delta r/r$ bei der Verwendung von Gleichung (24) mit zwei sigmoidalen Übergängen zwischen den Wachstumsraten.

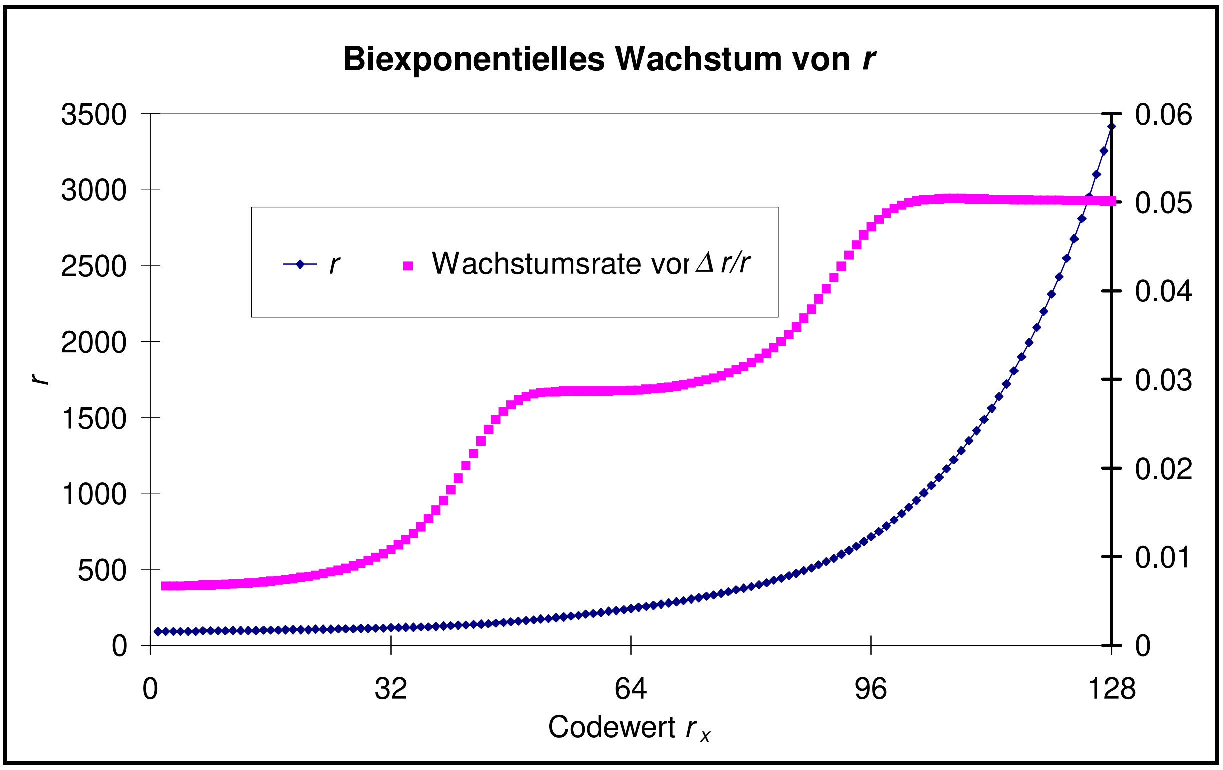


Abbildung 109: Biexponentielles Wachstum von *r* mit zwei sigmoidalen Übergängen zwischen zwei prozentualen Wachstumsraten. Die hier verwendeten Parameter waren: p_f=0, p_s=p_f =0.25%, $r_{w1,2}$=125 pm, $r_{w2,3}$=500 pm, $f_{s1,2}$=10, $f_{s2,3}$=100.

Aus Abbildung 110 mit drei linearen Wachstumsraten von *r*, zwischen denen sigmoidale Übergänge verwendet wurden, ist klar zu erkennen, das jetzt zwar *r* linear wächst, aber die Wachstumsrate für hohe Werte von *r* abnimmt. Ein Effekt, der unabdingbar mit dem linearen Wachstum von *r* verbunden, aber nicht erwünscht ist. Denn im allgemeinen sollten die möglichen Auswirkungen konformativer Flexibilität auf die Distanz zweier Atomkerne linear von deren Distanz abhängen, wenn nicht gar exponentiell mit dieser wachsen, keinesfalls aber unabhängig von der Distanz stets den gleichen Wert betragen.

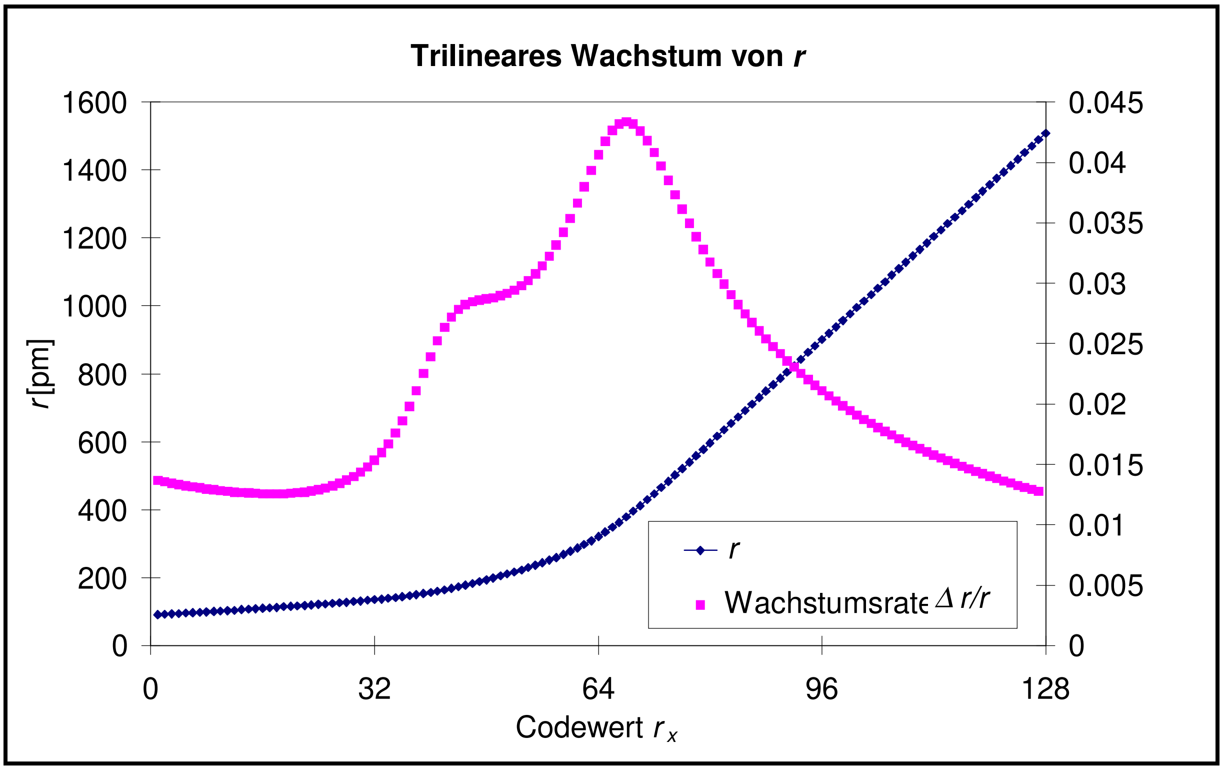


Abbildung 110: Trilinieares Wachstum von *r* mit zwei sigmoidalen Übergängen zwischen den Wachstumsraten. Die hier verwendeten Parameter waren: *Dmin* = 90, *Δmin* = 1.0 pm, *Δw1*= 3 pm, *Δw2*= 5 pm, $r_{w0,1}$= 150 pm, $r_{w1,2}$= 300 pm, $f_{s0,1}$= 10 pm, $f_{s1,2}$= 50 pm.

Abbildung 110 zeigt deutlich, daß eine Kombination nur aus linearen Wachstumsraten nicht das Mittel der Wahl sein kann. Vielmehr wird eine Kombination aus linearen und exponentiellen Wachstumsraten sinnvoll sein, denn im Distanzbereich von Bindungen wird man die Genauigkeit relativ konstant halten wollen. So könnte ein Ausgangspunkt für die Optimierung einer Codierung beispielsweise wie folgt aussehen:

Dmin = 90, *Δmin* = 1.0 pm, p_f = 0.25 %, *Δw1*= 3 pm, *Δw2*= 5 pm, $r_{w0,1}$= 150 pm, $r_{w1,2}$= 300 pm, $f_{s0,1}$= 10 pm, $f_{s1,2}$= 50 pm

9.1.2 Notwendigkeit der Skalierung

Fraglich ist, welche Bedeutung den einzelnen Bindungsabständen bzw. Codewerten zukommt, so dürfte die Zahl der CC-Einfachbindungen, die ein Intervall von 152-156 pm codieren würde, das somit immer einen hohen Wert hätte, für die biologische Wirkung aber relativ uninteressant sein, während dies für Intervalle, die für funktionelle Gruppen stehen, anders ist. Hieraus ergibt sich die Notwendigkeit zur Skalierung des Codes. Neben anderen Verfahren bietet sich insbesondere die Skalierung mit Hilfe eines Standarddatensatzes an, die sich auf die Skalierungsparameter für jeden einzelnen Codewert aus den minimalen und maximalen Codewerten eines Beispieldatensatzes bezieht und so eine Unterscheidung zwischen „üblichen“ und damit wenig interessanten Codewerten und „ungewöhlichen“ Codewerten erlaubt.

9.2 Die Zukunft der Simulation von Infrarotspektren

Die anfrageorientierte Methode zur Simulation von Infrarotspektren gibt dem Chemiker ein Werkzeug zur schnellen Simulation (innerhalb von 5 Minuten auf einer Sun 10-40) von Infrarotspektren für eine Vielzahl von Derivaten in die Hand. Wird dieses Werkzeug mit einer Datenbank kombiniert, die auch das Suchen und Anzeigen von Struktur-Spektren-Paaren erlaubt, kann diese Methode in vielen Fällen zur Simulation von Infrarotspektren genutzt werden.

Wichtig ist, daß der Chemiker die Grenzen dieser Methode im Auge behält:

1. Diese Methode kann nicht aus dem Datenraum der zugrundliegenden Datenbank heraus extrapolieren. Die Datenbank muß deshalb ausreichend ähnliche Moleküle für eine Interpolation des Infrarotspektrums der Anfrage-Struktur enthalten.

2. Für eine Reihe von Molekülen kann es keine ähnlichen Moleküle geben, die auch ein ähnliches Infrarotspektrum haben, da jedes Hinzufügen, Weglassen oder Ändern der Ordnungszahl von Atomen aus Gründen der Symmetrie und/oder des chemischen Aufbaus der Struktur, das Infrarotspektrum erheblich verändern würde. Beispiele für solche Verbindungen sind Methan, Ethan, Ethen, Ethin, Benzol und Anisol.

Die erste Grenze ist zugleich ein Aufruf an alle Chemiker jedes Infrarotspektrum einer neu synthetisierten Substanz der wissenschaftlichen Gemeinschaft in elektronischer Form zur Verfügung zu stellen. Dies würde gleich in zweifacher Hinsicht helfen, zum einen die schnelle und sichere Identifikation bekannter Substanzen mit Hilfe des Vergleichs der Infrarotspektren erlauben, und zum anderen die Datenbasis für die Simulation von Infrarotspektren bisher unbekannter Verbindungen stetig erweitern. Für letzteres ist es wichtig, daß die anfrageorientierte Simulation von Infrarotspektren eine Erweiterung der Datenbasis sofort nutzen kann, da sie direkt auf die Datenbank zurückgreift.

Die zweite Grenze zieht eine Trennlinie zwischen anfrageorientierter Simulation und *ab initio* Berechnungen von IR-Spektren. Letztere sind gerade zur Berechnung von IR-Spektren symmetrischer Moleküle geeignet, bieten doch alle *ab initio* Verfahren die Möglichkeit zu rechenzeitsparender Berücksichtigung der Symmetrie von Molekülen. So zeigt die Trennlinie auch, wo sich die beiden Methoden sinnvoll ergänzen.

Die Zukunft der anfrageorientierten Simulation hängt vor allem davon ab, ob es gelingt Infrarotspektren in hoher Qualität mit Angabe der Meßmethode zu sammlen und elektronisch verfügbar abzuspeichern.

Auf dem Wege dahin wird es noch einige Änderungen im Rahmen der anfrageorientierten Methodik geben. Zunächst wird es eine Berücksichtigung der Meßmethodik geben, wobei fraglich bleibt, ob in Form einer Warnung vor Abweichungen, wenn die Infrarotspektren aus dem Trainingsdatensatz mit unterschiedlichen Methoden gemessen wurden oder durch nach der Meßmethodik getrennten Datenbanken.

Die mit der Methode gesammelten Erfahrungen werden bei der Verbesserung der Codierung von 3D-Struktur und Infrarotspektren helfen. Ansätze zur Verbesserung der 3D-Strukturcodierung wurden in Kapitel 9.1 vorgestellt. Die Codierung von Infrarotspektren bzw. die Vergleichsmaße für Infrarotspektren und ihre Implementation in das zur Simulation verwendete CPG-Netz werden die Aussagekraft von IR-Spektren und ihre Simulation verbessern.

Bisher gab es im wesentlichen drei Verfahren zum Vergleich von Infrarotspektren. Bedingt dadurch, daß in der Literatur oft nur die Lage starker Banden angegeben wurde, war die Prüfung auf übereinstimmende Bandenlagen oft die einzige Möglichkeit zum Vergleich von Infrarotspektren. Die Angabe eines Infrarotspektrums mit 3-5 Bandenlagen, bedeutet aber eine ungeheure Datenreduktion, die die Unterscheidung ähnlicher Spektren sicher verhindert und die oft wichtige Information schwacher Banden über das Gerüst einer Verbindung verwirft.

Die euklidische Distanz zweier Infrarotspektren, ihr *rms*-Wert, das Fehlerquadrat oder ähnliche Vergleichsmaße für Infrarotspektren haben alle das gemeinsame Problem, daß sie auf den Intensitätswerten des Infrarotspektrums arbeiten. Diese hängen aber gerade in der Infrarotspektroskopie erheblich von den Aufnahmebedingungen des Spektrums ab, so daß Intensitätsdifferenzen selbst von 50% einer Bande, wenn die Aufnahmebedingungen nicht exakt identisch waren, wenig Aussagekraft haben.

Das dritte auch in dieser Arbeit verwendete Vergleichsmaß für Infrarotspektren, der Korrelationskoeffizient, macht einen relativen Vergleich der Spektren, und ist das beste der drei bisher verwendeten Verfahren.[102] Die Schwäche des Korrelationskoeffizienten ist aber, daß er eine identische Null-Linie der Länge x genauso bewertet wie eine identische Bande die an der Basis die Breite x hat. Außerdem bewertet er alle Intensitätsdifferenzen gleich, ohne Rücksicht auf den spektralen Verlauf. Dabei geht die Information über Schultern und Bandenaufspaltung leicht verloren, die obwohl signifikant, nur selten zu großen relativen Intensitätsdifferenzen führt.[103]

Weiterentwicklungen wären hier durch die Verwendung der Ableitung des Infrarotspektrums sowie durch die Einführung von Grenzwerten für Intensitätsdifferenzen normierter Spektren möglich.[104]

Die Entfernung des Rauschens der Basislinie durch Festschreibung der geringsten Absorption auf einen festen kleinen Wert, würde den Einsatz logarithmischer Vergleichsmaße möglich machen. Logarithmische Vergleichsmaße hätten den Vorteil, einen Vergleich auf der Basis logarithmischer Intensitätsdifferenzen zu erlauben.

So bieten sich viele Möglichkeiten zur Verbesserung des Verfahrens der anfrageorientierten Simulation an. Ob die Nutzung der anfrageorientierten Simulation aber einigen wenigen vorbehalten bleibt, die über eine Datenbank verfügen, welche ihr Forschungsgebiet einigermaßen abdeckt, oder ob es eine universelle Methode wird, hängt davon ab, ob es gelingt, genügend Spektren für einen breiten Einsatz zu sammeln.

CPG-Netze erlauben, einmal trainiert, die Vertauschung von Eingabe und Ausgabe. Dies bedeutet, daß man im Falle der Infrarotspektroskopie bei der Eingabe eines Infrarotspektrums die Vorhersage eines Strukturcodes erhalten könnte. Die Arbeiten zur Decodierung von Strukturcodes zu dreidimensionalen Molekülstrukturen sind im vollen Gange. Zwar enthalten die vorgestellten Strukturcodes in den meisten Fällen zuwenig Informationen für eine zweifelsfreie Lösung des Decodierungsproblems, dennoch ist häufig ein korrekter Strukturvorschlag möglich, da die Beschränkung auf chemisch sinnvolle Strukturen eine erhebliche Einschränkung des Lösungsraums bedeutet.

So besteht die Hoffnung, daß die Korrelation von Infrarotspektrum und 3D-Struktur einer Verbindung in Zukunft einen breiten Einsatz der Infrarotspektroskopie in der analytischen

Chemie erlauben. Mit den hier vorgestellten Verfahren einschließlich der 3D-Strukturvorhersage aus dem Infrarotspektrum sähe der mögliche Einsatz der Infrarotspektroskopie in der Analytik wie folgt aus:

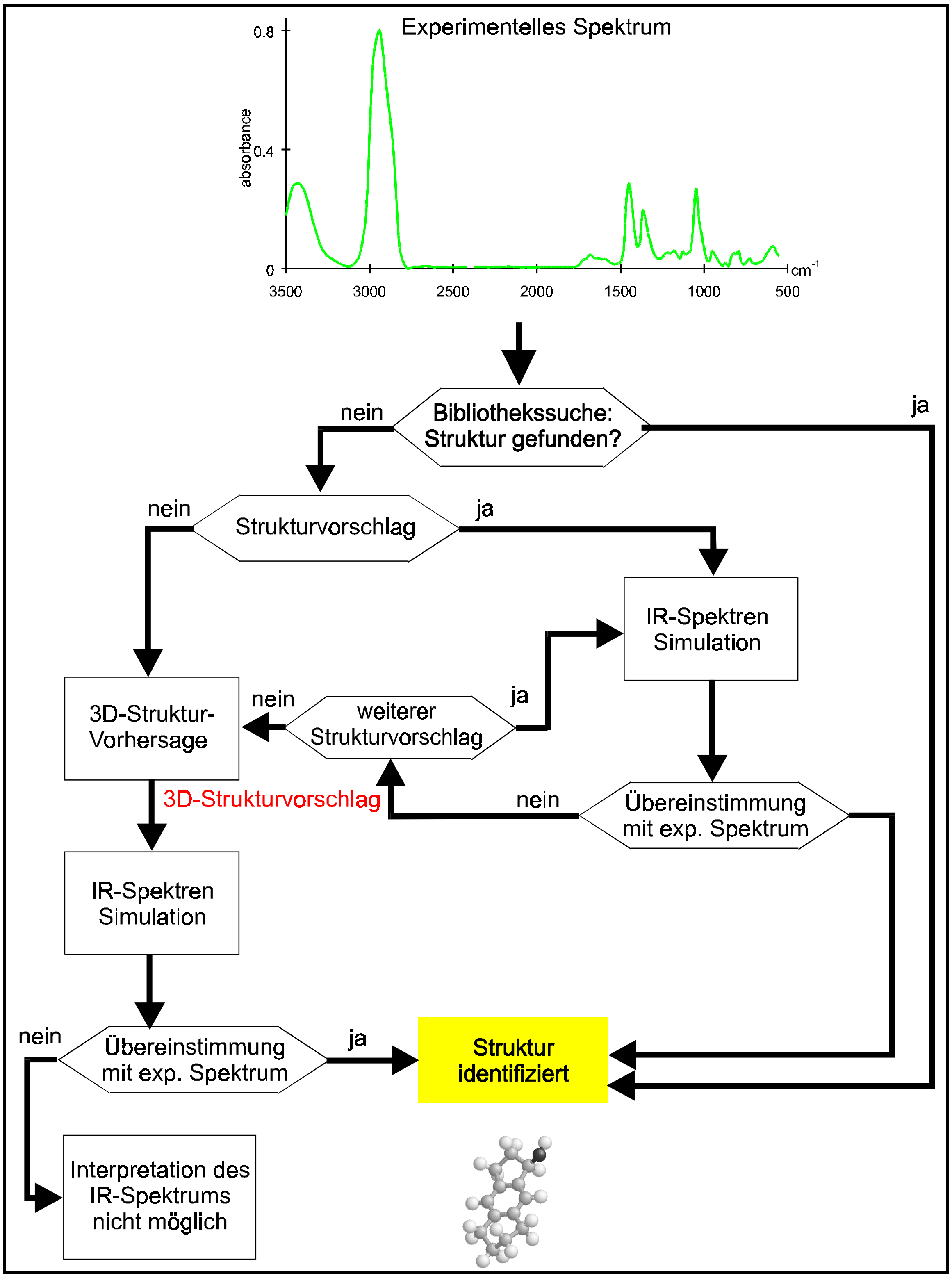


Abbildung 111: Zukünftige Einsatzmöglichkeiten der Infrarotspektroskopie zur Identifikation von Verbindungen.

9.3 QSAR/QSPR-Untersuchungen basierend auf einer 3D-Strukturcodierung

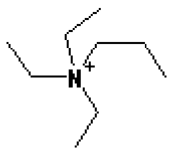
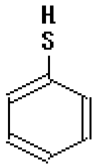
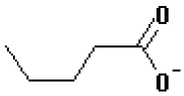
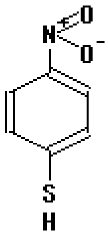
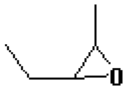
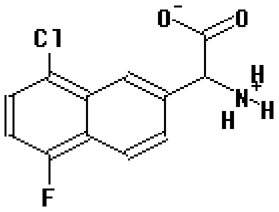
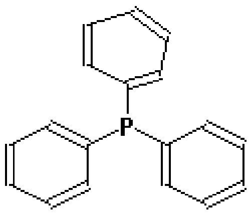
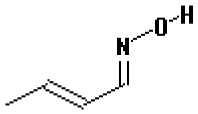
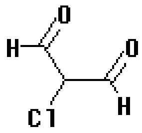
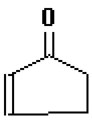
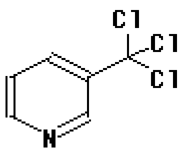
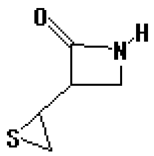
3D-Strukturgeneratoren liefern heute die 3D-Struktur für fast beliebig große Datensätze in kürzester Zeit. Damit ist die Suche nach 3D-Strukturen in nahezu jeder chemischen Strukturdatenbank möglich. Selbst künstliche 3D-Strukturdatenbanken könnten aus dem Output von Strukturgeneratoren wie Molgen[105] angelegt werden. 3D-Strukturcodierungen werden die schnelle Suche nach ähnlichen 3D-Strukturen selbst in größten Datenbanken erlauben und bieten Anlaß zur Hoffnung auf die zuverlässige Vorhersage von Struktureigenschaften. Experimente der kombinatorischen Chemie könnten so auf den Computer übertragen werden. Die Suche nach neuen Leitstrukturen für bekannte Wirkprinzipien würde am Computer möglich werden, indem man gezielt nach anderen Stoffklassen sucht, deren 3D-Strukturen aber ähnlich den bekannten Wirkstoffen sind.

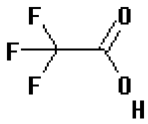
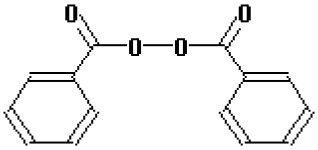
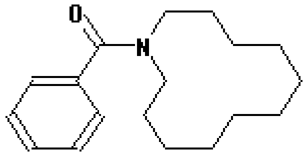
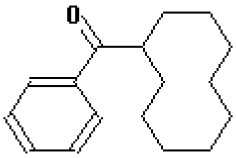
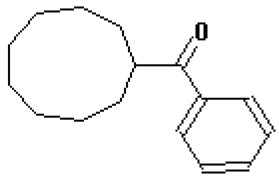
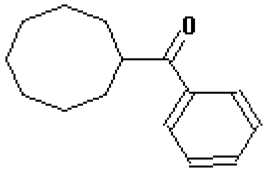
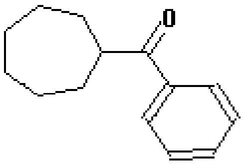
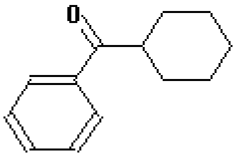
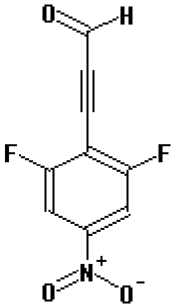
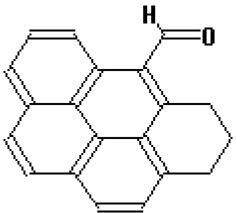
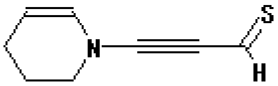
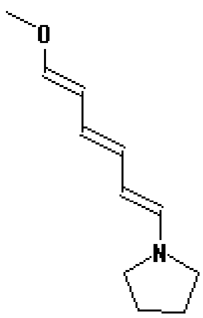
Das bedeutet aber keineswegs das Ende chemischer Forschung. Alle QSAR/QSPR-Untersuchungen setzen Bekanntes voraus. Ein neues Wirkprinzip oder neue Stoffeigenschaften werden auf diese Weise nicht gefunden werden. Extrapolationen aus dem bekannten Wissen heraus werden selten gelingen.

Deshalb werden neue Wirkprinzipien weiterhin nur durch das Prinzip von Versuch und Irrtum sowie durch die Beobachtung der Natur gefunden werden können. Das gleiche gilt für Stoffklassen mit überraschenden Eigenschaften.

10 Anhänge und Literaturverzeichnis

10.1 Anhang 1 Standarddatensatz

10.2 Anhang 2: Atomeigenschaften und Skalierungsfaktoren

Ordnungszahl, Skalierungsfaktor 1/17.0;

Elektronegativität der freien Elektronenpaare, Skalierungsfaktor 1/8.6925;

Elektronegativität der p-Orbitale, Skalierungsfaktor 1/9.21920;

Elektronegativität der s-Orbitale, Skalierungsfaktor 1/15.340;

Anzahl der freien Elektronen, Skalierungsfaktor 1/6.0;

Stabilisierungsenergie der freien Elektronenpaare, Skalierungsfaktor 1/167.110;

Anzahl der benachbarten Atome, Skalierungsfaktor 1/4.0;

Anzahl der benachbarten nicht Wasserstoffatome, Skalierungsfaktor 1/4.0;

Partielle Ladung der p-Orbitale, Skalierungsfaktor 1/0.5;

Partielle Atomladung, Skalierungsfaktor 1/0.60411;

Partielle Ladung der s-Orbitale, Skalierungsfaktor 1/0.479410;

Atommasse, Skalierungsfaktor 1/35.5;

Flag Bestandteil eines Rings, Skalierungsfaktor 1/1.0;

Atom = 1, Skalierungsfaktor 1/1.0;

10.3 Skalierungsfaktoren für die Werte des 3D-MoRSE Codes

Die Werte des 3D-MoRSE Codes nach Gleichung (11) sind entsprechend ihres Index mit dem korrespondierenden Multiplikationsfaktor zu multiplizieren, anschließend ist der additive Wert zu addieren. Es folgen nun in Tabellenform die Skalierungsfaktoren sortiert nach der Atomeigenschaft A_i, s_{max} und der Anzahl der Werte n.

10.3.1 Skalierungsfaktoren für den 3D-MoRSE Code mit $A_i=1$, $n=32$, $s_{max}=31$ Å^{-1}

Wert	Multiplikationsfaktor	additiver Wert
0	0.001899	-1.053181
1	0.059632	-1.528429
2	0.273005	1.144434
3	0.380064	0.020969
4	0.442235	1.262159
5	0.234176	1.166447
6	0.368331	-1.029467
7	0.662819	-0.963404
8	0.355073	-0.541738
9	0.482703	-0.409363
10	0.574144	0.419264
11	0.567202	0.773683
12	1.080876	-0.067486
13	0.60701	-0.374062
14	0.755747	-1.032286
15	1.334685	0.465189
16	1.122697	-0.552876
17	0.681654	0.965193
18	0.832126	0.718003
19	0.394128	-0.956607
20	0.673101	0.22396
21	0.691088	0.567546
22	1.919132	-0.472361
23	1.081566	0.912867
24	3.24794	-0.346346
25	0.876406	-0.956823
26	2.135839	0.237133
27	1.858686	-0.115191
28	1.089327	0.15254
29	0.967769	0.89079
30	2.090965	-0.044209
31	1.154323	-0.882405

10.3.2 *Skalierungsfaktoren für den 3D-MoRSE Code mit* A_i*=1,* n*=32,* s_{max}*=9.42 Å^{-1}*

Wert	Multiplikationsfaktor	additiver Wert
0	0.001899	-1.053181
1	0.002908	-1.07379
2	0.009387	-1.1758
3	0.038232	-1.417869
4	0.127015	-1.590632
5	0.151045	-0.052767
6	0.261924	0.865466
7	0.225908	1.144698
8	0.232908	1.038061
9	0.477835	0.932443
10	0.358699	-0.08853
11	0.411549	-0.534616
12	0.580554	-0.105477
13	0.451538	1.152991
14	0.369014	1.323828
15	0.295327	1.307637
16	0.263518	1.234262
17	0.197469	1.083453
18	0.178818	0.191145
19	0.298725	-0.93373
20	0.312573	-0.919889
21	0.360825	-0.954582
22	0.461548	-1.255995
23	0.654193	-0.985718
24	0.503207	-0.396445
25	0.642725	-0.432256
26	0.418703	-0.467831
27	0.259623	-0.304911
28	0.435497	-0.403348
29	0.485452	-0.201188
30	0.550761	-0.397184
31	0.708493	0.224474

10.3.3 Skalierungsfaktoren für den 3D-MoRSE Code mit $A_i=q_{tot,i}$, $n=32$, $s_{max}=31$ *Å*$^{-1}$

Wert	Multiplikationsfaktor	additiver Wert
0	0.001899	-1.053181
1	0.059632	-1.528429
2	0.273005	1.144434
3	0.380064	0.020969
4	0.442235	1.262159
5	0.234176	1.166447
6	0.368331	-1.029467
7	0.662819	-0.963404
8	0.355073	-0.541738
9	0.482703	-0.409363
10	0.574144	0.419264
11	0.567202	0.773683
12	1.080876	-0.067486
13	0.60701	-0.374062
14	0.755747	-1.032286
15	1.334685	0.465189
16	1.122697	-0.552876
17	0.681654	0.965193
18	0.832126	0.718003
19	0.394128	-0.956607
20	0.673101	0.22396
21	0.691088	0.567546
22	1.919132	-0.472361
23	1.081566	0.912867
24	3.24794	-0.346346
25	0.876406	-0.956823
26	2.135839	0.237133
27	1.858686	-0.115191
28	1.089327	0.15254
29	0.967769	0.89079
30	2.090965	-0.044209
31	1.154323	-0.882405

10.3.4 Skalierungsfaktoren für den 3D-MoRSE Code mit $A_i = q_{tot,i}$, $n = 32$, $s_{max} = 9.42$ Å^{-1}

Wert	Multiplikationsfaktor	additiver Wert
0	0.75142	0.062564
1	1.1274	0.031472
2	1.473991	0.21313
3	1.916104	0.070852
4	4.627834	0.57137
5	5.795319	0.745946
6	6.918949	-0.757389
7	8.065846	-0.763106
8	6.593377	-0.482139
9	6.136175	-0.829903
10	11.454682	-0.28134
11	10.365054	0.194448
12	8.749268	0.564273
13	13.305958	0.730617
14	11.579918	0.92116
15	23.653682	0.05127
16	21.016724	0.034488
17	30.243701	0.032062
18	17.214747	-0.687891
19	13.2952	-0.958267
20	17.811195	-0.677397
21	20.327253	0.61953
22	10.924135	0.536195
23	13.673067	0.539619
24	19.2857	0.761136
25	21.941047	-0.151499
26	17.774352	-0.552425
27	26.171777	-0.362235
28	43.779703	-0.122565
29	22.286502	-0.432832
30	32.670289	-0.679166
31	51.366794	-0.487332

10.3.5 *Skalierungsfaktoren für den 3D-MoRSE Code mit* $A_i = q_{tot,i}$, n=*64*, s_{max}=*15.5 Å^{-1}*

Wert	Multiplikationsfaktor	additiver Wert
0	0.75142	0.062564
1	0.869391	0.011467
2	1.1274	0.031472
3	1.413451	0.248252
4	1.473991	0.21313
5	1.54445	0.108413
6	1.916104	0.070852
7	3.85765	0.394922
8	4.627834	0.57137
9	3.807376	0.765361
10	5.795319	0.745946
11	7.79518	-0.005494
12	6.918949	-0.757389
13	8.44468	-0.701089
14	8.065846	-0.763106
15	7.236194	-0.586419
16	6.593377	-0.482139
17	5.543417	-0.65099
18	6.136175	-0.829903
19	8.443187	-0.914399
20	11.454682	-0.28134
21	10.111802	0.14771
22	10.365054	0.194448
23	9.483457	0.52498
24	8.749268	0.564273
25	9.796739	0.619949
26	13.305958	0.730617
27	10.607694	0.804107
28	11.579918	0.92116
29	16.338186	0.6588
30	23.653682	0.05127
31	21.241781	-0.066979
32	21.016724	0.034488
33	30.430342	0.151331
34	30.243701	0.032062
35	42.773307	-0.424685
36	17.214747	-0.687891
37	14.12965	-0.787383
38	13.2952	-0.958267
39	14.733154	-0.884095
40	17.811195	-0.677397
41	22.766159	-0.398343
42	20.327253	0.61953
43	13.044188	0.655076
44	10.924135	0.536195
45	12.076035	0.420159
46	13.673067	0.539619
47	18.786204	0.55906
48	19.2857	0.761136
49	19.827196	0.656117
50	21.941047	-0.151499
51	18.633077	-0.481862
52	17.774352	-0.552425
53	19.439756	-0.48596
54	26.171777	-0.362235
55	47.464152	-0.064219
56	43.779703	-0.122565
57	28.82279	-0.404991
58	22.286502	-0.432832
59	28.412291	-0.429043
60	32.670289	-0.679166
61	33.845189	-0.669783
62	51.366794	-0.487332
63	54.397851	-0.277251

10.3.6 *Skalierungsfaktoren für den 3D-MoRSE Code mit* $A_i=q_{tot,i}$, n=*120,* s_{max}=*30.0 Å*$^{-1}$

Wert	Multiplikationsfaktor	additiver Wert
0	0.75142	0.062564
1	0.878161	0.007559
2	1.141846	0.042992
3	1.448625	0.273069
4	1.467574	0.192115
5	1.577748	0.097513
6	2.070662	0.082421
7	4.78561	0.59224
8	4.291129	0.622092
9	4.113393	0.802022
10	6.65249	0.5536
11	7.067693	-0.349446
12	7.003264	-0.777865
13	9.169996	-0.748385
14	7.76004	-0.760747
15	7.331244	-0.46548
16	5.806833	-0.569361
17	5.673133	-0.746095
18	7.387932	-0.95101
19	9.885564	-0.572111
20	9.98968	0.009118
21	10.151699	0.203191
22	9.532265	0.41561
23	8.763094	0.555637
24	9.390655	0.602486
25	12.46624	0.722166
26	10.846302	0.787551
27	11.177255	0.903082
28	15.636144	0.711734
29	23.813211	0.076392
30	21.33417	-0.071105
31	21.016724	0.034488
32	31.069257	0.154503
33	31.176247	-0.011702
34	37.065429	-0.526433
35	16.473063	-0.69826
36	13.899615	-0.811063
37	13.206906	-0.960601
38	15.9726	-0.82173
39	19.612993	-0.641039
40	21.712734	-0.059749
41	17.570673	0.683519
42	11.726977	0.61663
43	11.260587	0.476837
44	12.228108	0.454223
45	15.380869	0.565704
46	20.891397	0.668153
47	19.16745	0.781939

48	20.691796	0.276567
49	20.393639	-0.435346
50	17.722383	-0.536135
51	18.433912	-0.515438
52	23.189324	-0.412712
53	38.070827	-0.1816
54	47.346019	-0.123767
55	32.825605	-0.334175
56	22.335826	-0.432165
57	26.620569	-0.42868
58	33.152301	-0.629337
59	33.02831	-0.682946
60	51.531946	-0.484764
61	54.285698	-0.306897
62	48.663123	0.570849
63	34.775133	0.768718
64	28.731569	0.653351
65	29.761111	0.365519
66	36.432543	0.34879
67	32.37915	0.439174
68	24.415452	-0.101809
69	23.119961	-0.401043
70	25.900323	-0.591305
71	32.181112	-0.648273
72	48.495073	-0.671415
73	67.334733	-0.582867
74	48.930872	-0.52159
75	32.841061	-0.149588
76	28.662013	0.095022
77	29.420146	0.182203
78	42.431791	0.346673
79	76.509203	0.295953
80	56.756516	-0.303253
81	33.137555	-0.395504
82	29.681196	-0.491956
83	34.847518	-0.554155
84	48.790743	0.295629
85	49.241343	0.615631
86	57.407598	0.614185
87	54.925848	0.571021
88	54.825907	0.518021
89	73.859504	0.402648
90	57.849416	0.383349
91	60.048514	0.100548
92	85.944766	0.029444
93	68.110454	-0.229826
94	60.475559	0.05194
95	39.486078	0.242472
96	36.451517	0.369265
97	38.77988	0.304978
98	51.810593	0.100301

99	68.962237	0.045291
100	56.013588	-0.026522
101	42.793627	-0.375871
102	36.954332	-0.64841
103	36.379666	-0.826432
104	34.252916	-0.629103
105	43.069673	-0.580931
106	55.487805	-0.497864
107	47.199619	-0.532208
108	39.367549	-0.115052
109	44.943618	0.744503
110	40.266408	0.904062
111	39.907459	0.847867
112	44.400913	0.648637
113	63.535636	0.507915
114	102.476188	0.137191
115	129.162973	-0.247709
116	98.226737	-0.220987
117	102.219589	-0.262287
118	94.421265	-0.455692
119	58.702484	-0.475263

10.3.7 *Skalierungsfaktoren für den 3D-MoRSE Code mit* A_i=m, n=*32*, s_{max}=*31.0 Å*$^{-1}$

Wert	Multiplikationsfaktor	additiver Wert
0	0.076467	-1.174666
1	1.187563	-1.726375
2	3.629296	0.888764
3	3.248595	0.144326
4	4.673041	1.124524
5	3.604124	-0.077736
6	3.108435	-0.483381
7	6.819788	0.447255
8	7.437121	0.665541
9	10.837669	0.154849
10	9.372645	-0.172193
11	7.69913	-0.781446
12	9.41064	0.124073
13	9.561177	0.827896
14	7.139623	-0.56198
15	11.755286	-0.272035
16	15.154067	-0.132974
17	26.430392	0.498358
18	11.01314	0.728121
19	10.655832	-0.651486
20	13.463135	-0.304808
21	24.905711	0.740808
22	21.273167	0.162507
23	21.361441	0.477942
24	18.26412	0.228524
25	22.720189	-0.427188
26	20.224224	0.220882
27	15.603572	0.279293
28	19.799596	0.439731
29	16.748797	-0.166856
30	33.55457	-0.513427
31	30.54818	0.280878

10.3.8 *Skalierungsfaktoren für den 3D-MoRSE Code mit* A_i=Z, n=*32*, s_{max}=*31.0 Å*$^{-1}$

Wert	Multiplikationsfaktor	additiver Wert
0	0.057724	-1.168381
1	1.078937	-1.834849
2	3.473611	0.881108
3	3.123165	0.199034
4	4.240211	1.147003
5	3.31734	0.069222
6	2.847812	-0.56006
7	7.058457	0.401578
8	7.071331	0.683186
9	10.164875	0.229941
10	8.965585	-0.210609
11	7.128799	-0.614036
12	8.466498	0.090757
13	9.024223	0.733326
14	6.519457	-0.59173
15	11.842647	-0.19511
16	13.860198	-0.19632
17	19.74965	0.663712
18	9.892807	0.721788
19	9.527557	-0.726159
20	11.931902	-0.29343
21	23.438946	0.830191
22	19.979474	0.169453
23	19.575389	0.588201
24	17.632271	0.194242
25	22.477707	-0.563767
26	19.613185	0.230536
27	15.204273	0.322637
28	19.023113	0.440681
29	15.310831	-0.077286
30	32.606928	-0.529157
31	25.868862	-0.039021

10.3.9 Skalierungsfaktoren für den 3D-MoRSE Code mit A_i=Z, n=*32,* s_{max}=*9.42 Å*$^{-1}$

Wert	Multiplikationsfaktor	additiver Wert
0	0.057724	-1.168381
1	0.085992	-1.229524
2	0.231242	-1.391809
3	0.819811	-1.752259
4	1.449969	-1.074655
5	1.680193	-0.120662
6	2.220116	0.299477
7	4.306285	0.981418
8	3.21162	0.963104
9	2.966678	0.265521
10	3.421875	0.210637
11	7.034954	0.496557
12	5.123545	1.039071
13	4.434218	1.143168
14	3.521501	1.123289
15	2.865504	0.941367
16	3.118969	0.344456
17	4.271267	-0.411715
18	2.554411	-1.111168
19	2.86784	-1.013017
20	2.893387	-0.47423
21	4.582466	-0.352819
22	9.119356	-0.676124
23	7.087875	0.373171
24	5.81877	0.504768
25	5.541973	0.581806
26	6.692945	0.652381
27	9.553075	0.685313
28	9.785645	0.577351
29	11.205839	0.580244
30	9.238366	-0.037473
31	8.128462	-0.561208

10.3.10 Skalierungsfaktoren für den 3D-MoRSE Code mit A_i=Z, n=*32*, s_{max}=*15.5 Å*$^{-1}$

Wert	Multiplikationsfaktor	additiver Wert
0	0.057724	-1.168381
1	0.153428	-1.31563
2	1.078937	-1.834849
3	1.712467	-0.125544
4	3.473611	0.881108
5	3.43403	1.001892
6	3.123165	0.199034
7	6.849045	0.974851
8	4.240211	1.147003
9	2.925769	0.977227
10	3.31734	0.069222
11	2.52276	-1.129974
12	2.847812	-0.56006
13	5.924452	-0.495872
14	7.058457	0.401578
15	5.414179	0.560576
16	7.071331	0.683186
17	9.76126	0.572976
18	10.164875	0.229941
19	8.792039	-0.629924
20	8.965585	-0.210609
21	10.756467	-0.101831
22	7.128799	-0.614036
23	8.038913	-0.370532
24	8.466498	0.090757
25	12.115758	0.782066
26	9.024223	0.733326
27	5.972383	0.166042
28	6.519457	-0.59173
29	7.798928	-0.951356
30	11.842647	-0.19511
31	9.637879	-0.098631

10.4 Anhang 4: Verzeichnis der Publikationen

Die mit * gezeichneten Veröffentlichungen sind Teil der Dissertation.

1. Schuur, J. H.; Selzer, P.; Gasteiger, J.;
   „The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity“
   J. Chem. Inf. Comput. Sci. **1996**, *36*, 334-344. *

2. Schuur, J. H.; Gasteiger, J.;
   „3D-MoRSE Code - A Method for Coding the 3D Structure of Molecules“
   in Software Development in Chemistry 10, J. Gasteiger (Editor); Gesellschaft Deutscher Chemiker: Frankfurt am Main, **1996**; S. 67. *

3. Schuur, J. H.; Selzer, P.; Gasteiger J.;
   „Simulation of IR Spectra with Neural Networks Using the 3D-MoRSE Code“
   in Software Development in Chemistry 10¸ J. Gasteiger (Editor); Gesellschaft Deutscher Chemiker: Frankfurt am Main, **1996**; S. 293. *

4. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V.;
   „Chemical Information in 3D space“
   J. Chem. Inf. Comput. Sci. **1996**, *36*,1030-1037. *

5. Schuur, J.; Selzer, P.; Steinhauer, V.; Gasteiger, J.;
   Kooperative, rechnergestützte IR-Spektreninterpretation - neue Wege für die Infrarotspektroskopie“
   GIT Labor-Fachzeitschrift **1997**, 283-286. *

6. Schuur, J.; Gasteiger, J.;
   „Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation“
   Anal. Chem. **1997**, *69*, 2398-2405. *

7. Schuur, J. H.; Selzer, P.; Steinhauer, V.; Gasteiger, J.;
„3D structure coding opens new applications for IR spectroscopy“
Linking and Interpreting Spectra through Molecular Structures, *LISMS*, Charlton, Chicester, **1997**, 15-28. *

8. Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V.;
„Finding the 3D Structure of a Molecule in Its IR Spectrum“
Fresenius J. Anal. Chem. **1997**, *359*, 50-55. *

9. Schuur, J.;
„Spektren-Interpretation und -Verwaltung mit SpecInfo“
Nachr. Chem. Tech. Lab. **1997**, *45*, 369-405.

10. Schuur, J.;
„Informationsorganisation fürs Internet“
Nachr. Chem. Tech. Lab. **1997**, *45*, 518-520.

11. Selzer, P.; Hemmer, M. C.; Schuur, J. H.; Steinhauer, V.; Gasteiger, J.;
„TeleSpek - Telekooperation in der Spektroskopie“
Nachr. Chem. Tech. Lab. **1998**, 46, A78 - A82.*

12. Hemmer, M. C.; Selzer, P.; Schuur, J. H.; Gasteiger, J.;
„TeleSpek - Telekooperation in der Spektroskopie“
DFN-Mitteilungen **1998,** *47*, 8-9.*

10.5 Anhang 5: Lebenslauf

Vor- und Zuname	Jan Heinrich Schuur
Geburtsdatum und -ort:	16.11.68 in Bonn
Eltern:	Dr. med. Klaus Schuur und Annegret Schuur, geb. Gördes
Staatsangehörigkeit:	deutsch
Familienstand:	verheiratet, ein Sohn
Schulbildung:	1975-1979 Grundschule 1979-1988 Gymnasium am Kattenberg Buchholz i. d. N. 27.05.1988 Abitur, Duchschnittsnote 2,4, Gymnasium am Kattenberg Buchholz i. d. N.
Auslandsaufenthalte:	1984 3 Wochen als Austauschschüler in Bolton/England 1985 2 Wochen als Austauschschüler in Toulon/Frankreich 1994 2 Wochen Studienaufenthalt am National Institute of Chemistry Ljubljana/Slowenien
Studium:	November 1988 bis Oktober 1990 Studium der Chemie an der Universität Würzburg November 1990 bis Sommer 1993 Studium der Chemie an der Universität Göttingen Sommer 1992 - Mai.1993 Diplomarbeit bei den Professoren Hüttermann/Lackner, Note „sehr gut“ 05.07.1993 Diplom mit der Note „Gut“ an der Universität Göttingen
Promotion:	Begonnen im Oktober 1993 bei Professor Dr. J. Gasteiger an der Technischen Universität München, Februar 1994 Wechsel mit dem Doktorvater an die Friedrich-Alexander Universität Erlangen-Nürnberg 23.7.98 Promotion, Note „sehr gut“
Berufliche Tätigkeiten:	Sommer 1988 zweimonatiges Praktikum Nordmark Pharma GmbH, Üttersen Frühjahr 1990 acht Wochen Werkstudent Bayer AG Leverkusen 1994 - heute PC Betreuung im Arbeitskreis Prof. Gasteiger 1996 - heute nebenberufliche Tätigkeit als freier Journalist, Veröffentlichungen in Fachzeitschriften: „Nachrichten aus Chemie, Technik und Laboratorium“, „GIT Fachzeitschrift für das Laboratorium“ und der Tagespresse.
Auszeichnungen:	1988 GDCh Buchpreis für Abiturleistungen im Fach Chemie 1995 3. Preis im Wettbewerb um das beste Poster bei der 10. Fachtagung der Fachgruppe Chemie-Information-Computer der GDCh.

LITERATURVERZEICHNIS

[1] Cambridge Christallographic Database, Ref. Code APSEUR

[2] M. C. Garry, Z. Stanek, Nicolet Technical Note TN-9259, Nicolet Spectroscopy Research Center, Madison, USA. 1996

[3] K. V. Sarkanen, H. Chang, G. G. Allan, *Tappi* **1967**, *50*, 587-590.

[4] E. Pretsch, J. Seibel, W. Simon, T. Clerc; *Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden*, zweite Auflage, Springer Verlag, Berlin Heidelberg New York, **1981.**

[5] M. Hesse, H. Meier, B. Zeeh, *Spektroskopische Methoden in der Organischen Chemie*, Thieme Verlag, Stuttgart, Germany **1991**, S 29-67.

[6] S. Hünig, G. Märkl, J. Sauer, Integriertes Organisches Praktikum, Verlag Chemie, Weinheim Germany, **1979**.

[7] L.-F. Tietze, T. Eicher; *Reaktionen und Synthesen im organisch-chemischen Praktikum*, Thieme Verlag, Stuttgart, **1981**.

[8] C. Middelberg, v. Jürgensonn, Werk Uetersen, Knoll AG, BASF Pharma, persönliche Mitteilung, **1998**.

[9] V. E. Turula, J. A. de Haseth, *Applied Spectrocopy* **1994**, *48*, 1255-1264.

[10] SpecInfo Version 3.1.6.0, Chemical Concepts, Weinheim, Bundesrepublik Deutschland.

[11] SpecInfo Version 3.1.6.0, Chemical Concepts, Weinheim, Bundesrepublik Deutschland, Spektrum MI-IG00000127.

[12] SpecInfo Version 3.1.6.0, Chemical Concepts, Weinheim, Bundesrepublik Deutschland, Spektrum MI-IV00020446.

[13] A. P. Scott, L. Radom, *J. Phys. Chem.* **1996,** *100*, 16502-16513.

[14] CrossFire plus Reactions, Beilstein Informationssysteme GmbH, Frankfurt.

[15] E. Negishi, S. J. Holmes, J. M. Tour, J. A. Miller, F. E. Cederbaum, D. R. Swason, T. Takahashi; *J. Am. Chem. Soc.* **1989,** *111*, 3336-3346.

[16] http://www.uni-erlangen.de/docs/RRZE/dienste/dioalog/

[17] R. Salzer, S. Thiele, H. Thomas; *Fresenius J. Anal. Chem.* **1997,** *359*, 126-131.

[18] P. Willet; *Three Dimensional Structure Handling*, XXXX

[19] L. J. Soltzberg, C. L. Wilkins; *J. Am. Chem. Soc* **1977**, *99*, 439-443.

[20] L. J. Soltzberg, C. L. Wilkins; *J. Am. Chem. Soc* **1976**, *98*, 4006.

[21] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer; *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030-1037.

22 D. Bawden, P. Willett, Preface in *Three-Dimensional Chemical Structure Handling*, Hrsg. P. Willett, Research Studies Press Ltd., Sommerset, **1991**.

23 Daylight Smiles-String

24 A. M. Lesk; *Communications of the ACM* **1979**, *22*, 219-224.

25 J. Figueras; Journal of Chemical documentation **1972**, 12, 237-244.

26 J. R. Ullmann; *Journal of the ACM* **1976**, *16*, 31-42.

27 B. Bienfait, J. Gasteiger, in print.

28 B. Bienfait persönliche Mitteilung. 30.07.96.

29 L. Steinhauer, V. Steinhauer, J. Gasteiger; *Obtaining the 3D Structure from Infrared Spectra of Organic Compounds Using Neural Networks*, in: Software Development in Chemistry 10, J. Gasteiger (Hrsg.), GDCh, Frankfurt/Main, **1996**, S. 315-322.

30 J. Schuur, P. Selzer, V. Steinhauer, J. Gasteiger; *3D Structure Coding Opens New Applications for IR Spectroscopy*, in: Linking and Interpreting Spectra through Molecular Structures, LISMS, Charlton, Chichester, 1997, S. 15-28.

31 M. Wagener, J. Sadowski, J. Gasteiger; *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.

32 J. Zupan, J. Gasteiger, *Neural Networks for Chemists - An Introduction*, VCH Verlag, Weinheim, **1993**.

33 J. Polánski, J. Gasteiger, M. Wagener, J. Sadowski, *Quant. Struct.-Act.* **1998**, *17*, 27-36.

34 S. Anzali, G. Barnickel, M. Krug, J. Sadowski, M. Wagener, J. Gasteiger, J. Polanski; *Journal of Computer-Aided Molecular Design* **1996**, *10*, 521-534.

35 J. Karle; J. Chem. Inf. Comput. Sci. **1994,** 34, 381-390.

36 R. Wierl; *Ann. Phys.* (Leipzig), **1931**, *8*, 521-564.

37 L. J. Soltzberg, C. L. Wilkins; *J. Am. Chem. Soc.*, **1977,** *99*, 439-443.

38 I. Csorvássy, L. Töszsér, L. Kárpáti, G. Náray-Szabó; *J. Math. Chem.,* **1993,** *13,* 343-357.

39 G. Náray-Szabó, V. Harmat; *MATCH*, **1997***, 35*, 29-40.

40 HyperChem Release 4 for Windows, Hypercube Inc. Gainesville, Florida, USA. **1995.**

41 M. Hemmer persönliche Mitteilung.

42 B. Schrader; *Raman/Infrared Atlas of Organic Compounds*, VCH Verlag, Weinheim, **1989**.

43 EPA-Werte: Nist Chemistry Webbook, http://webbok.nist.gov/chemistry/

44 H. Roth, N. v. E. Hommes, persönliche Mitteilung.

45 M. Hesse, H. Meier, B. Zeeh, in: Spektroskopische Methoden der organischen Chemie, Georg Thieme Verlag, Stuttgart, New York, **1991**, S. 50.

46 AM1 Implementation: HyperChem Release 4 for Windows, Hypercube Inc. Gainesville, Florida, USA. **1995.**

47 S. Hünig, G. Märkl, J. Sauer, Integriertes Organisches Praktikum, Verlag Chemie, Weinheim Germany, **1979**. Die Spektren wurden gescannt und mittels des Programms UN-SCAN-IT digitalsiert. Die Darstellung erfolgt mit Hilfe von MS EXCEL.[79]

48 E. F. Healy, A. Holder; *J. Mol. Struct. (Theochem)* **1993**, *281*, 141-156.

49 S. R. Langhoff; *J. Phys. Chem.* **1996**, *100*, 2819-2841.

50 C. W. Bauschlicher, J. Partridge, H. Partridge; *J. Chem. Phys.* **1995**, *103*, 1788-1791.

51 H. Huixiao, X. Xinquan; *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 290-302.

52 SpecInfo Version 3.0 Manuel, Limitationen des HOSE Codes, Chemical Concepts, Weinheim, **1995**.

53 C. Affolter, K. Baumann, J. T. Clerc, H. Schriber, E. Pretsch; *Mikrochim. Acta* **1997**, *14*, 143-147.

54 J.E. Dubois; G. Mathieu; P. Peguet; A. Panaye; J. P. Doucet; *J. Chem. Inf. Comput. Sci* **1990**, *30*, 290-302.

55 H. Huixiao, X. Xinquan; *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 203-210.

56 J. T. Clerc, A. L. Terkovics; *Anal. Chim. Acta.* **1990**, *235*, 93-102.

57 J. T. Clerc, Vortrag Analytika'96 in München.

58 C. Hiller, J. Sadowski, C. Schwab, J. Gasteiger; CORINA Version 1.7, Universität Erlangen-Nürnberg 1996

59 J. Gasteiger, C. Rudolph, J. Sadowski; *Tetrahedron Comput. Method.* **1992**, *3*, 537-547.

[60] J. Sadowski, C. Rudolph, J. Gasteiger; *Anal. Chem. Acta.* **1992**, *165*, 233-241.

[61] J. Sadowski, J. Gasteiger; *Chem. Rev.* **1993***, 93*, 2567-2581.

[62] J. Sadowski, J. Gasteiger, G. Klebe; *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.

[63] J. Gasteiger, M. Marsili; *Tetrahedron*, **1980**, *36*, 3219-3228.

[64] M. G. Hutchings, J. Gasteiger; *Tetrahedron Lett.* **1983**, *24*, 2541-2544.

[65] J. Gasteiger, H. Saller; *Angew. Chemie*, **1985**, *97*, 699-701.

[66] Kohonen, T. *Self-Organisation and Associative Memory*, 3rd Edition; Springer: Berlin, 1989.

[67] Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59-69.

[68] M. Novic, J. Zupan; *J. Chem. Inf. Comput. Sci* **1995**, *35*, 454-466.

[69] M. D. Guillen, M. G. Hutchings, M. Marsili, H. Saller, A. Fröhlich, O. Dammer, K. Rafeiner, J. Gasteiger, PETRA 6.0, Technische Universität München 1992

[70] X. Li, M. Wagener, J. Gasteiger, *kmap* Version 3.0, Universität Erlangen-Nürnberg 1996.

[71] J. Gasteiger, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer; Fresenius J. Anal. Chem. **1997**, *359*, 50-55.

[72] AM1 Implementation: HyperChem Release 4 for Windows, Hypercube Inc. Gainesville, Florida, USA. **1995.**

[73] P. Selzer, J. Schuur, J. Gasteiger; *Simulation of IR Spectra with Neural Networks Using the 3D-MoRSE Code,* in: Software-Entwicklung in der Chemie 10, J. Gasteiger (Hrsg.), GDCh, Frankfurt/Main, 1996, S. 293-303.

[74] CSED, W.-D. Ihlenfeldt, erhältlich unter http://www2.ccc.uni-erlangen.de/cactvs, siehe auch: W. D. Ihlenfeldt, Y. Takahashi, H. Abe, S. Sasaki, *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 109-116.

[75] M. D. Guillen, R. Höllering, M. G. Hutchings, M. Marsili, H. Saller, A. Fröhlich, O. Dammer, K. Rafeiner, J. Gasteiger, PETRA, Technische Universität München 1992

[76] J. Schuur, J. Gasteiger, Code3D, Universität Erlangen-Nürnberg 1996.

[77] P. Selzer, J. Gasteiger, CHOOSE V. 1.1, Universität Erlangen-Nürnberg 1996.

[78] J. Schuur, J. Gasteiger, j2jcamp, Universität Erlangen-Nürnberg 1996

[79] MS Excel Version 7.0a, Microsoft GmbH, 85716 Unterschleißheim

[80] W.-D. Ihlenfeldt, CSIR, Universität Erlangen-Nürnberg 1996.

[81] E. Pretsch, J. Seibel, W. Simon, T. Clerc in Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden, zweite Auflage, Springer Verlag Berlin Heidelberg New York, **1981**, I100.

[82] W.-D. Ihlenfeldt, J. Gasteiger, *submtc*, Technische Universität München 1991.

[83] PM3 Implementation: HyperChem Release 4 for Windows, Hypercube Inc. Gainesville, Florida, USA. **1995**.

[84] C. J. Pouchert*; The Aldrich Library of Infrared Spectra Edition III*, Aldrich Chemical Company, Milwaukee, **1981**, S. 197.

[85] WWW-Adresse: www2.ccc.uni-erlangen.de/IR

[86] J. March; in Advanced Organic Chemistry, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, **1985**, S 499 ff.

[87] E. Pretsch, J. Seibel, W. Simon, T. Clerc in Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden, zweite Auflage, Springer Verlag Berlin Heidelberg New York, **1981**, I55.

[88] A. Manjarrez, T. Rios, A. Guzmán, *Tetrahedron* **1964**, *20*, 333-339.

[89] E. Pretsch, J. Seibel, W. Simon, T. Clerc in Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden, zweite Auflage, Springer Verlag Berlin Heidelberg New York, **1981**, I275.

[90] DFN-Projekt TeleSpek, Leitung Prof. Dr. J. Gasteiger, Mitarbeiter: Paul Selzer, Jan Schuur, Markus Hemmer, Prof. Dr. R. Salzer, siehe http://www2.ccc.uni-erlangen.de/IR.

[91] H.-J. Böhm, G. Klebe, H. Kubinyi, Wirkstoffdesign, Heidelberg, Berlin, Oxford, Spektrum Akad. Verlag, **1996**, S 88 ff.

[92] H. B. Niznik, R. K. Sunahara, Z. B. Pristupa, K. R. Jarvie, *Molekulare Grundlagen der Interaktion zwischen Dopamin-(D1-/D2-)Rezeptoren,* in Schizophrenie - Dopaminrezeptoren und Neuroleptika, Jes Gerlach (Hrsg), Springer Verlag, Berlin, Heidelberg, New York, **1995**, S. 2.

[93] MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, USA.

[94] H.-J. Böhm, G. Klebe, H. Kubinyi, Wirkstoffdesign, Heidelberg, Berlin, Oxford, Spektrum Akad. Verlag, **1996**, S 168 ff

[95] H.-J. Böhm, G. Klebe, H. Kubinyi, Wirkstoffdesign, Heidelberg, Berlin, Oxford, Spektrum Akad. Verlag, **1996**, S 162 ff

[96] Alchemy 2000, Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144-2913, USA.

[97] Beispielsweise: Spartan V. 5.0, 18401 Von Karman Ave., Suite 370, Irvine, CA 92612, USA.

[98] Markus Hemmer, Computer-Chemie-Centrum Universität Erlangen-Nürnberg, Erlangen, persönliche Mitteilung.

[99] H. R. Christen, Grundlagen der Organischen Chemie, Salle und Sauerländer, Berlin, Frankfurt, München, Salzburg, 6. Auflage, **1985**, S. 27.

[100] K. P. C. Vollhardt, Organische Chemie, VCH Verlag, New York, **1988**, S. 115.

[101] J. Zupan, J. Gasteiger, *Neural Networks for Chemists - An Introduction*, VCH Verlag, Weinheim, **1993**, S. 24ff.

[102] Paul Selzer, persönliche Mitteilung

[103] Markus Hemmer, persönliche Mitteilung

[104] Markus Hemmer, Jan Schuur unveröffentlichte Ergebnisse.

[105] Strukturgenerator Molgen, Universität Bayreuth, 1994-1998.