Vorhersage von Infrarotspektren mittels neuronaler Netze zur Identifikation organischer Verbindungen

Dissertation

Paul Selzer

1998

Vorhersage von Infrarotspektren mittels neuronaler Netze zur Identifikation organischer Verbindungen

Den Naturwissenschaftlichen Fakultäten der

Friedrich-Alexander-Universität Erlangen-Nürnberg

zur

Erlangung des Doktorgrades

vorgelegt von

Paul Selzer

aus Ostrau

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

Tag der mündlichen Prüfung:	
Vorsitzender der Promotionskommission:	Prof. Dr. D. Kölzow
Erstberichterstatter:	Prof. Dr. J. Gasteiger
Zweitberichterstatter:	Prof. Dr. S. Schneider

Meinem Doktorvater

Herrn Prof. Dr. Johann Gasteiger

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

Weiteren Dank schulde ich

meinen Kollegen der Infrarot-Gruppe Herrn Jan Schuur, Herrn Markus Hemmer, Herrn Dr. Valentin Steinhauer und meiner Kollegin Frau Laura Hernández-Alpízar für die gute Zusammenarbeit und die vielen interessanten Diskussionen

Herrn Andreas Teckentrup, Herrn Dr. Robert Höllering und Herrn Dr. Markus Wagener für die gute Zusammenarbeit bei der Administration des SUN Clusters

Herrn Dr. Wolf-Dietrich Ihlenfeldt, Frau Dr. Susanne Bauerschmidt, Herrn Dr. Ralf Fick und Herrn Dr. Bruno Bienfait für die Hilfestellungen bei programmiertechnischen Problemen

Herrn Thomas Kostka für die gute Zusammenarbeit bei dem Projekt zur Identifikation der Pestizidabbauprodukte

Herrn Prof. Dr. Reiner Salzer und seinem Arbeitskreis für die gute Zusammenarbeit im Rahmen des TeleSpek-Projekts und die freundliche Unterstützung meiner experimentellen Arbeiten am Institut für Analytische Chemie der TU Dresden

Herrn Jan Schuur und Herrn Oliver Sacher für das Schaffen einer stabilen PC-Arbeitsumgebung

Frau Eva Kohl und Herrn Jochen Hardt für die Aufgabenstellungen zur Anwendung der Spektrensimulation, die Eingang in diese Arbeit gefunden haben

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

meiner Freundin Frau Petra Fischer für die Unterstützung und die Geduld, die sie mir im Bezug auf das Verfassen dieser Arbeit entgegengebracht hat.

Für die finanzielle Unterstützung dieser Arbeit gebührt Dank:

Der Deutschen Forschungsgemeinschaft DFG, dem Bundesministerium für Bildung und Forschung bmbf sowie dem Verein zur Förderung des deutschen Forschungsnetzes e.V. DFN.

Für meine Eltern,
meine zukünftige Ehefrau Petra,
sowie
Ivette, Achim, Felix, Babette und Rudolf

Abkürzungen und physikalische Größen

Abkürzungen:

3D-MoRSE Code	3D-Molecule Representation of Structures based on Electron diffraction Code
ANN	Artificial Neural Network
CPG-Netzwerk	Counterpropagation Netzwerk
DMA	N,N-Dimethylanilin
DMA-NO	N,N-Dimethylanilin-N-Oxid
FMO	Flavinhaltige Monooxigenase
IR	Infrarot
KNN	künstliches neuronales Netz
MB	Megabyte
NN	neuronales Netz
vs	versus

Mathematische und physikalische Größen:

A	Atomeigenschaft
B	Temperaturparameter bei der Radialcodierung
E	Absorbanz (Extinktion), Energie
F	Skalierungsfaktoren bei der Radialcodierung
K	Gerätekonstante
N	Anzahl der Atome eines Moleküls
P_{ij}	temperaturabhängige Wahrscheinlichkeitsverteilung des Abstands zwischen Atom *i* und Atom *j*
R	Radius, Atomabstand
R_{max}	maximaler Atomabstand innerhalb eines Moleküls, maximaler berücksichtigter Atomabstand bei der Strukturcodierung
T	Transmission
Z	Ordnungszahl
c	Lichtgeschwindigkeit
c_{ij}	Gewicht *i* von Neuron *j*
d	Neuronenabstand
f	Formfaktor
h	Plancksches Wirkungsquantum
k	Skalierungsfaktoren für den 3D-MoRSE Code
m	Masse
n	Anzahl
q_π	π-Ladung
q_σ	σ-Ladung
q_{tot}	Gesamtladung
r	Korrelationskoeffizient nach Bravais-Pearson
r_b	bereichsgewichteter Korrelationskoeffizient
rms	root mean square errror
s	beugungswinkelabhängige Größe bei der 3D-MoRSE Codierung

t ..... Zeit
α ..... Polarisierbarkeit
η ..... Lernrate des neuronalen Netzes
λ ..... Wellenlänge
ν ..... Frequenz
$\tilde{\nu}$ ..... Wellenzahl

1 Einleitung

Viele etablierte Verfahren in der chemischen Analytik beruhen vollständig oder zumindest teilweise auf spektroskopischen Methoden. Bei den verschiedenen spektroskopischen Verfahren wird die Analysenprobe mit Strahlung unterschiedlicher Wellenlänge bestrahlt, und aus dem Reflexions- oder Absorptionsverhalten Rückschlüsse auf die Probe gezogen. Die wohl am häufigsten durchgeführte spektroskopische Analyse findet im Bereich von 400 - 750 nm statt und wird als Sehen bezeichnet. Während jedoch die Interpretation unserer Sinneseindrücke durch unser Gehirn prompt und unmerklich geschieht, müssen die Rohdaten einer spektroskopischen Analyse zunächst aufgearbeitet werden. Erst durch das Wissen um die Grundlagen der spektroskopischen Methoden, der Wechselwirkung zwischen eingestrahlter Strahlung und der untersuchten Materie, also dem Zusammenhang zwischen Struktur und Spektrum, ist es möglich, die detektierten Signale zu interpretieren, um so Aussagen über die Probe treffen zu können. Die Untersuchung der Grundlagen der Wechselwirkung von Strahlung und Materie sowie der Effekte auf molekularer und atomarer Ebene sind die zentralen Fragestellungen in der Spektroskopie. Mit dem Wissen um den Zusammenhang zwischen Struktur- und Spektreninformation ist es möglich, Spektrensignale zu entschlüsseln und Spektrenvorhersagen zu treffen, um so Substanzen anhand deren Spektren zu analysieren und zu identifizieren. Die vorliegende Arbeit beschreibt die Entwicklung eines Systems zur Vorhersage infrarot-(IR)-spektroskopischer Daten. Mit diesem System kann für ein Anfragemolekül das entsprechende Infrarotspektrum vorhergesagt werden. Der Zusammenhang zwischen der Struktur des Anfragemoleküls und dem Infrarotspektrum wird mittels eines künstlichen neuronalen Netzes modelliert:

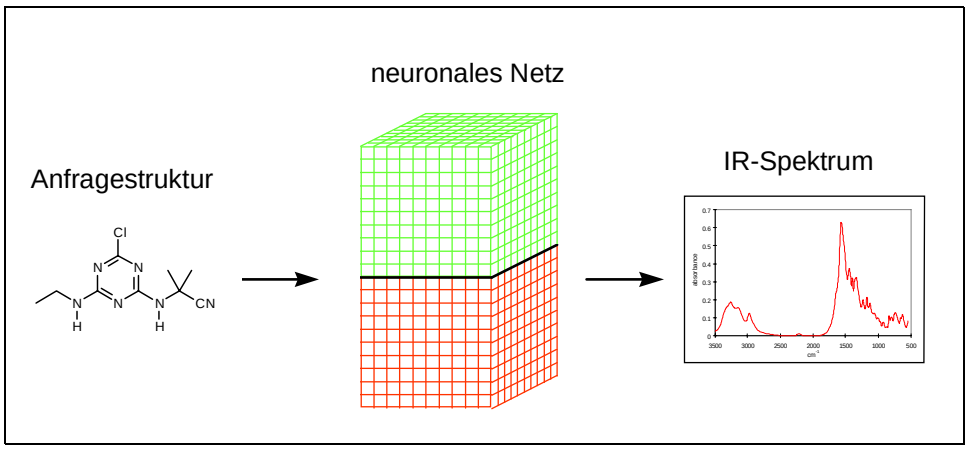


Abb. 1-1: Schematischer Ablauf der Spektrenvorhersage mittels eines neuronalen Netzes

Ein grundlegender Aspekt, in dem sich die hier vorgestellte Methode der Spektrenvorhersage von anderen Vorhersage- oder Berechnungsmethoden unterscheidet, ist, daß sie prinzipiell auf beliebige Moleküle anwendbar ist. Die Vorhersagegeschwindigkeit und Qualität ist unabhängig von der Molekülgröße. Da das neuronale Netz an Beispielen lernt, ist die Vorhersagequalität abhängig von der Qualität der zugrundeliegenden Datenbasis und wie gut die jeweilige Anfragestruktur durch die Datenbasis repräsentiert wird. Ein Schlüsselschritt dieser Methode ist dabei die Art der Strukturbeschreibung, mit der Molekülstrukturen dem neuronalen Netz präsentiert werden (vgl. Kap. 2.1.2.).

Durch die hohe Charakteristik und die gute Reproduzierbarkeit eines Infrarotspektrums erfolgt in der Infrarotspektroskopie eine Substanzidentifikation meist durch den Vergleich des experimentellen Spektrums mit einem Referenzspektrum aus einer Datenbank oder einem Spektrenkatalog. Dies kann natürlich nur funktionieren, wenn auch tatsächlich ein entsprechendes Referenzspektrum vorhanden ist. Gegenüber etwa 16 Millionen bekannten chemischen Verbindungen enthält die größte Infrarotspektrendatenbank nur ca. 100000 Spektren und Strukturen.[1] Die SpecInfo Infrarotdatenbank,[2] welche am Institut für Organische Chemie der Universität Erlangen-Nürnberg verwendet wird, enthält nur etwa 15000 Einträge. Durch dieses Mißverhältnis zwischen der Anzahl der Strukturen und den archivierten und zugänglichen Spektren, wird es sehr oft der Fall sein, daß für eine analysierte Verbindung zunächst kein Referenzspektrum in einer Datenbank oder einem Spektrenkatalog zu finden ist. Hier liegt das Haupteinsatzgebiet der Spektrenvorhersage. Für die vermessene Substanz bzw. eine Reihe in Frage kommender Kandidaten werden die Infrarotspektren vorhergesagt. Durch den Vergleich der vorhergesagten und der experimentellen Spektren können so die Substanzen identifiziert werden. Die vorgestellte Methode bietet einen schnellen Zugang zu beliebigen Referenzspektren.

1.1 Die Wechselwirkung zwischen infraroter Strahlung und Molekülen

Periodisch schwingende elektromagnetische Felder verursachen elektromagnetische Wellen, die sich, ohne äußere Einflüsse, mit Lichtgeschwindigkeit geradlinig im Raum ausbreiten. In Abhängigkeit von der Wellenlänge der Strahlung hat diese unterschiedliche Erscheinungen bzw. Einflüsse auf Materie und unsere Sinnesorgane. Der Spektralbereich der elektromagnetischen Strahlung reicht dabei von den hochenergetischen γ-Strahlen bis zu den niedrigenergetischen Radiowellen. Eine kennzeichnende Größe für die elektromagnetische Strahlung ist die Wellenlänge λ. Eine weitere Größe zur Beschreibung der Wellenbewegung ist die Schwingungsfrequenz ν. Sie ist definiert als die Anzahl der Schwingungen die der elektrische Vektor pro Zeiteinheit ausführt. Die Schwingungsfrequenz wird in s^{-1} oder auch Hertz

(Hz) angegeben. Die Wellenlänge λ und die Schwingungsfrequenz ν sind umgekehrt proportional. Die Konstante zur Beschreibung dieser Proportionalität ist die Lichtgeschwindigkeit *c*:

$$\nu = c / \lambda$$

Zusammenhang zwischen Frequenz, Lichtgeschwindigkeit und Wellenlänge (Gl. 1-1)

Die Schwingungsfrequenz hat eine direkte Beziehung zum Elementarphänomen der IR-Spektroskopie, nämlich der Wechselwirkung zwischen dem elektromagnetischen Wechselfeld und der Schwingungsbewegung der Atome im Molekül. Wegen der besseren Handhabbarkeit der Zahlenwerte wird jedoch nicht die Frequenz, sondern der Reziprokwert der Wellenlänge λ, die Wellenzahl $\tilde{\nu}$, verwendet.

$$\tilde{\nu} = \nu / c = \frac{1}{\lambda}$$

Definition der Wellenzahl $\tilde{\nu}$ (Gl. 1-2)

Infrarote Strahlung kann unter bestimmten Voraussetzungen Schwingungen der Atome sowie Atomgruppen eines Moleküls oder auch Rotation des Gesamtmoleküls anregen.[3][4] Wird ein geeigneter Lichtquant mit der Energie hν durch das Molekül absorbiert, so erfolgt ein Übergang des Systems vom Schwingungsgrundzustand auf einen schwingungsangeregten Zustand bzw. von einem schwingungsangeregten Zustand auf einen höheren Schwingungsangeregten Zustand. Dabei gilt die Resonanzbedingung:

$$\Delta E = h\nu$$

Resonanzbedingung (Gl. 1-3)

Die Resonanzbedingung besagt, daß die Energiedifferenz von Ausgangs- und Endniveau gleich der Energie des absorbierten Lichtquants hν sein muß. Damit eine Schwingung IR-aktiv ist, sie also im Spektrum zu beobachten ist, muß sich das Dipolmoment des Moleküls in einem Extrempunkt der Schwingung vom Dipolmoment des Moleküls im Grundzustand unterschei-

den. Beispielsweise sind von den vier möglichen Schwingungen des Kohlendioxidmoleküls, dessen Dipolmoment im Grundzustand 0 ist, nur drei IR-aktiv.

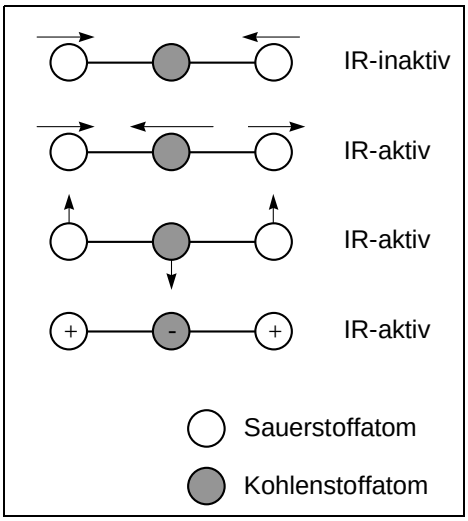


Abb. 1-2: Schwingungen des Kohlendioxidmoleküls

Wie bereits erwähnt wurde, ist infrarote Strahlung ebenso dazu geeignet eine Molekülrotation anzuregen. Bei Messungen in der Gasphase sind Rotationsfeinstrukturen detektierbar. Die nachfolgende Abbildung zeigt das Infrarotspektrum von Raumluft, wie es z.B. bei der Aufnahme eines Untergrundspektrums (engl.: background) vor der Vermessung der eigentlichen Probe entsteht. Das Spektrum zeigt deutlich die Rotationsfeinstruktur von Wasserdampf (gestrichelter Rahmen).

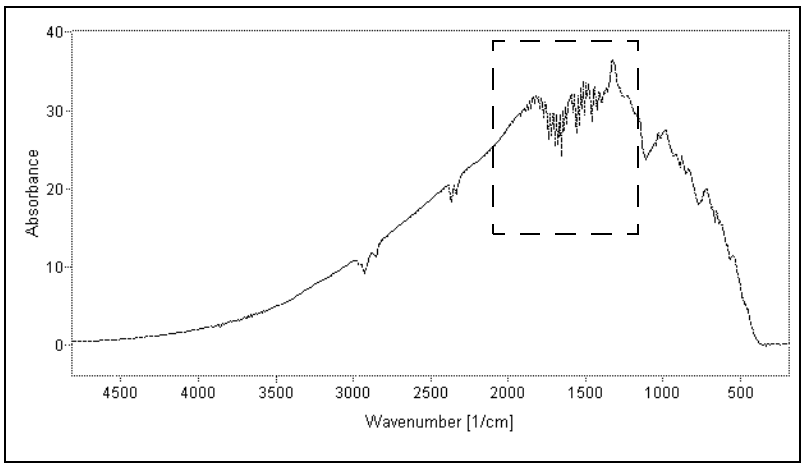


Abb. 1-3: Abbildung eines Untergrund-Spektrums von Luft mit Rotationsfeinstruktur (Backgroundspektrum). Entnommen aus dem virtuellen Infrarot-Spektroskopiekurs des Arbeitskreises von Prof. Salzer am Institut für Analytische Chemie, der TU Dresden. (http://analyt.chm.tu-dresden.de/analyt/praktikum/ir/irhom0.htm)

1.2 Warum Infrarotspektroskopie?

Wie im vorherigen Kapitel erwähnt wurde, lösen Strahlungen der verschiedenen Spektralbereiche in Wechselwirkung mit Materie unterschiedliche Effekte aus. Entsprechend lassen sich mit verschiedenen Strahlungen, also verschiedenen spektroskopischen Methoden, auch verschiedene Strukturmerkmale unterschiedlich gut analysieren. Wird nun im folgenden auf die Vorteile der Infrarotspektroskopie eingegangen, so soll dies nicht den Eindruck erwecken, daß es möglich wäre, mit dieser einen spektroskopischen Methode allen Problemstellungen der analytischen Chemie beizukommen. Tatsächlich leuchten die verschiedenen Analyseverfahren nur Teilbereiche aus. Eine komplexe Fragestellung wird somit nur durch die Kombination verschiedener Methoden zu beantworten sein, wobei die Infrarotspektroskopie jedoch sicherlich in einem Großteil der Fälle miteingebunden sein wird.

Die große Bedeutung der Infrarotspektroskopie beruht auf der Vielfalt der Probenmessung und Substanzpräparation sowie auf dem hohen Informationsgehalt der Spektren.[4][5][6][7] Anhand eines Infrarotspektrums können teilweise direkte Aussagen über die Konstitution eines Moleküls, beispielsweise die Nachbarschaft einzelner Strukturfragmente, getroffen werden. Direkte Aussagen bedeuten in diesem Zusammenhang solche Aussagen, die dem Spektrum einer unbekannten Probe allein basierend auf theoretisch abzuleitenden oder empirischen Zusammenhängen zu entnehmen sind. Da Lage und Intensität der Signale stoffspezifisch sind, kann eine unbekannte Probe oft durch den Vergleich von experimentellem Spektrum und Referenzspektrum identifiziert werden. Die hohe Spezifität wird durch eine gute Reproduzierbarkeit der Koordinaten der Absorptionsmaxima, also Wellenzahl und Absorbanz, gewährleistet. Daher läßt sich ein Infrarotspektrum, wie der Fingerabdruck beim Menschen als hochcharakteristische Eigenschaft, zur Identifizierung benutzen. Dies gilt besonders für jene Signale, die durch Schwingungen des Kohlenstoffgerüsts verursacht werden. Diese Signale sind im Bereich von 1500-1000 cm^{-1}, dem sogenannten Fingerprintbereich, zu beobachten. Wie bereits erwähnt wurde, setzt dieses Verfahren zur einfachen Substanzidentifikation durch den Vergleich von Spektren voraus, daß ein Referenzspektrum der vermessenen Substanz in einem Spektrenkatalog oder einer Datenbank enthalten ist. Bei dem bereits erwähnten Mißverhältnis zwischen der Anzahl bekannter Verbindungen und der Zahl zugänglicher archivierter Infrarotspektren, wird der beschriebene Weg der einfachen Substanzidentifikation mittels Spektrenvergleich oftmals nicht anwendbar sein, da für die unbekannte Probe eben kein entsprechendes Referenzspektrum vorhanden ist.

1.3 Wozu Struktur-Spektren-Korrelationen und warum mit neuronalen Netzen?

Eine der Hauptfragestellungen in der analytischen Chemie ist sicherlich die Identifikation einer unbekannten Probe. Eine einfache Identifikation durch den Vergleich von experimentellem Spektrum und einem entsprechenden Referenzspektrum ist jedoch aus den im vorherigen Kapitel beschriebenen Gründen oftmals nicht möglich. Zudem gibt es Fragestellungen, die über eine reine Substanzidentifikation hinausgehen. Im Kapitel 1.2 wurde bereits ausführlich dargelegt, welche Strukturinformation aus einem Infrarotspektrum herausgelesen werden kann. Ziel ist es daher, Verfahren zu entwickeln, die es erlauben, Struktur- und Spektreninformation zu korrelieren. Im Prinzip stellt die Auflistung von Tabellen mit Strukturfragmenten und den entsprechenden Spektrensignalen eine, wenn auch sehr einfache, Form der Korrelationsuntersuchung dar. Ziel der Korrelationsuntersuchungen ist dabei stets die Information aus Struktur und Spektrum zu verknüpfen, um anhand von Strukturinformation Spektrenvorhersagen zu treffen und umgekehrt. Im Falle der computergestützen Analytik schließt sich ein weiterer Schritt an, nämlich die Beschreibung dieser Zusammenhänge durch Regeln, die sich wiederum in Algorithmen transformieren lassen, um so eine Automatisierung der Spektrenanalyse zu ermöglichen. So gesehen bildet die Kombination von Struktur/Spektren-Tabellen mit den Augen eines erfahrenen Spektroskopikers ein Vorhersagesystem, an dessen Interpretationsfähigkeiten sich computergestützte Systeme messen lassen müssen.

Im Aufstellen dieser Regeln liegt bereits das grundlegende Problem: Durch die unüberschaubare Vielzahl an denkbaren Substrukturen ergeben sich eine unendliche Anzahl an sich überlappenden Signalmustern und -bereichen. Die Verknüpfungen von Struktur und Spektrum werden sich also schwerlich in universell einsetzbare und verläßliche Regeln fassen lassen. Selbst der oben erwähnte erfahrene Spektroskopiker wird oftmals Schwierigkeiten haben, die Einzelschritte der Interpretation explizit zu formulieren, da sich viele Spektrenmerkmale verschiedenen Strukturmerkmalen zuordnen lassen und die Identifikation oft zu einem großen Teil auf Intuition beruht. Desweiteren ist es möglich, Infrarotspektren mittels *ab initio* oder dichtefunktionaler Methoden zu berechnen. Die Ergebnisse zeigen zwar meist gute Übereinstimmung mit experimentellen Werten, jedoch sind die Rechnungen trotz der hohen Kapazität heutiger Rechnersysteme sehr zeitintensiv. Beispielsweise benötigt die Berechnung verschiedener Cyclohexenderivate mit acht nicht-H-Atomen auf einem SUN Ultra-Sparc Rechenserver etwa 36 Stunden. Semiempirische Verfahren arbeiten deutlich schneller, wobei die so ermittelten Wellenzahlen oft erst nach einer Skalierung mit einem empirisch ermittelten Faktor den Bereich der experimentellen Werte erreichen.

Die Korrelation von Struktur und Infrarotspektrum mittels eines neuronalen Netzes, wie

es das Thema der vorliegenden Arbeit ist, beschreitet einen ganz anderen Weg. Bei diesem datenbasierten Ansatz lernt das neuronale Netz den Zusammenhang zwischen Struktur und Spektrum selbständig anhand einer Reihe von Beispielstrukturen und Spektren. Ein trainiertes, neuronales Netz ist dann in der Lage für eine ihm bis dahin unbekannte Verbindung das Spektrum vorherzusagen. Da die Methode datenbasiert ist, lassen sich damit Infrarotspektren, einschließlich des charakteristischen Fingerprintbereichs, vorhersagen. Die Vorhersagequalität ist dabei ebenso wie die Rechenzeit nahezu unabhängig von der Größe des Moleküls. Die Simulation eines Infrarotspektrums benötigt auf einer SGI ORIGIN 200 je nach der angewendeten Methode (vgl. Kap. 2.4.2) etwa 1.5 Minuten.

2 Korrelation von Struktur und Infrarotspektrum

2.1 Beschreibung von Daten

Die verschiedenen Verfahren für Korrelationsuntersuchungen, z.B. statistische Verfahren oder neuronale Netze, benötigen jeweils eine einheitliche Beschreibung von Daten. Einheitlich bedeutet in diesem Fall, daß die Anzahl der Werte mit der beispielsweise ein Molekül beschrieben wird, konstant sein muß. Die Anzahl der Werte darf von Versuch zu Versuch differieren, jedoch nicht innerhalb eines Versuchslaufes. Vor einer Korrelationsuntersuchung gilt es also, die Daten in einen Code konstanter Länge zu transformieren. Jede derartige Transformation ist zwangsläufig mit einem mehr oder weniger großen Informationsverlust verbunden. Es gilt also ein Transformationsverfahren zu wählen, das den Verlust an für die Korrelationsuntersuchung relevanter Information so gering wie möglich hält.

2.1.1 Spektrenbeschreibung

Die Beschreibung von Spektren durch einen Code konstanter Länge ist verhältnismäßig einfach, da das Spektrometer die Spektren gewissermaßen bereits in codierter Form, also als eine Reihe von Werten, die bestimmten Wellenzahlen zugeordnet werden, ausgibt. Gebräuchliche Größen sind Transmission *T* und Absorbanz *E*, die wie folgt zusammenhängen:

$$T = I / I_0$$

$$E = lg\ (1/ T)$$

Zusammenhang von Transmission und Absorbanz (Gl. 2-1)

Bei den in den nachfolgenden Kapiteln beschriebenen Untersuchungen, werden die Infrarotspektren durchwegs als Graphen in Absorbanz gegen die Wellenzahl aufgetragen.

2.1.1.1 Datenreduktion

Wie bereits erwähnt wurde, können die Spektrendaten in der Form, wie sie durch das Spektrometer ausgegeben werden, zu Korrelationsuntersuchungen eingesetzt werden. Da bei vielen Verfahren jedoch die Rechenzeit nahezu exponentiell mit der Zahl der Variablen ansteigt, ist es oftmals sinnvoll, eine Datenreduktion durchzuführen.[8] Hierbei finden ver-

schiedene Verfahren Verwendung, z.B.:

- Mittelwertbildung eines Intervalls:
  Hier wird die Wellenzahlen-Skala des Infrarotspektrums in Intervalle unterteilt. Für jedes dieser Intervalle wird der jeweilige Mittelwert bestimmt. Die Mittelwerte stellen die Stützstellen des datenreduzierten Spektrums dar.

- Hadamard-Transformation
  Bei dieser Methode wird mit dem Spektrum eine Hadamard-Transformation durchgeführt. Diese ist der Fourier-Transformation sehr ähnlich, wobei jedoch statt der Sinusfunktion eine Rechteckfunktion verwendet wird.
  Von den so erhaltenen Hadamard-Koeffizienten wird eine bestimmte Anzahl an Koeffizienten gleich Null gesetzt. Die verbleibenden Koeffizienten werden in das Infrarotspektrum rücktransformiert.[9] Je mehr Koeffizienten zuvor auf Null gesetzt worden sind, desto geringer ist die Auflösung des datenreduzierten Spektrums.[10]

Die Infrarotspektren, die bei den nachfolgenden Korrelationsuntersuchungen eingesetzt wurden, werden durch 128 Absorbanzwerte beschrieben. Um diese 128 Werte aus den Datenbankspektren zu erhalten, wurden zunächst in den beiden Wellenzahlenbereichen von 3500-2000 cm^{-1} und 2000-552 cm^{-1} jeweils äquidistante Stützstellen durch Interpolation bestimmt. Der Abstand der Stützstellen beträgt danach 10 bzw. 4 cm^{-1}. Diese Spektren mit 512 Absorbanzwerten wurden in 512 Hadamard-Koeffizienten transformiert. Die Hadamard-Koeffizienten 129 bis 512 wurden gleich Null gesetzt und die verbleibenden 128 Koeffizienten wurden rücktransformiert. Das datenreduzierte Spektrum hat nun im Bereich von 3500-2000 cm^{-1} eine Auflösung von 40 cm^{-1} und im Bereich von 2000-552 cm^{-1} eine Auflösung von 16 cm^{-1}. Dies entspricht der analytischen Praxis, da die beiden Bereiche in der Regel eine unterschiedliche Signal- und damit auch Informationsdichte aufweisen. Der Vergleich eines Originalspektrums und eines datenreduzierten Spektrums ist in Abbildung 2-1 am Beispiel von Mesitylenspektren dargestellt:

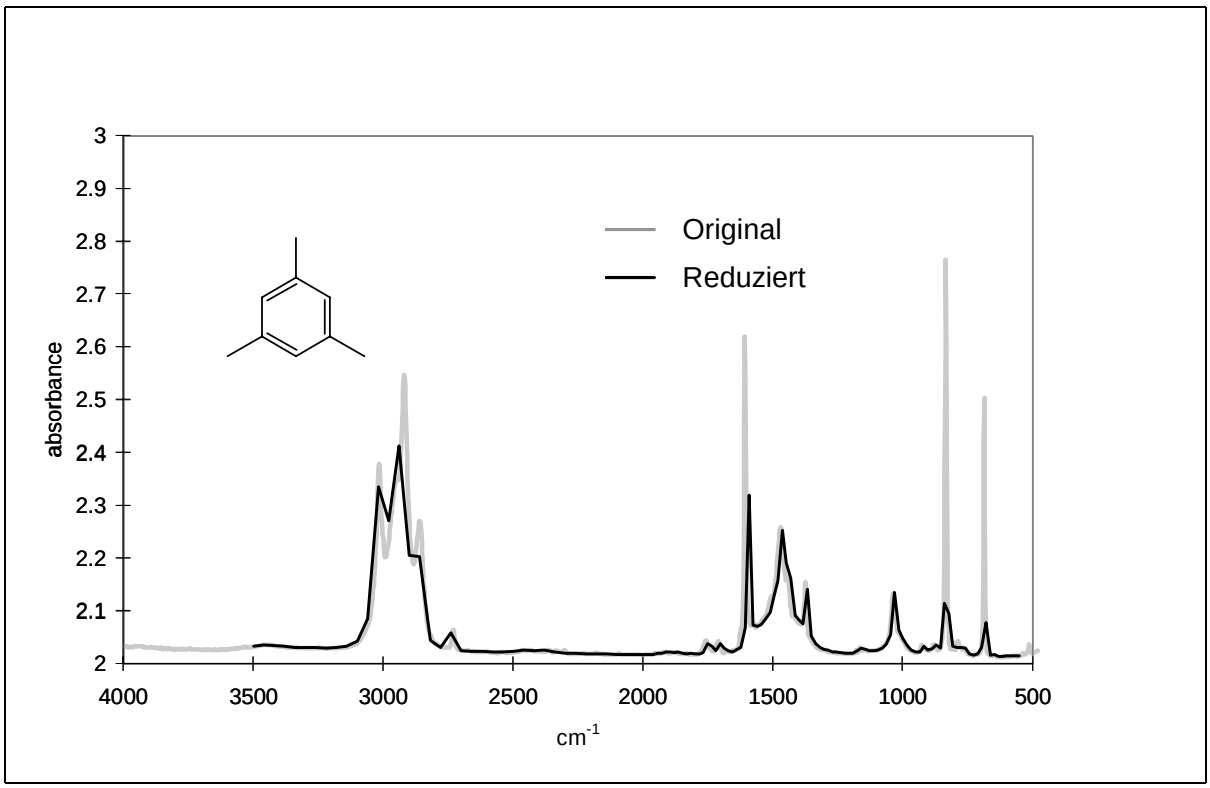


Abb. 2-1: Vergleich des Originalspektrums und des datenreduzierten Spektrums von Mesitylen

Es fällt auf, daß scharfe Signale mit einer geringen Halbwertsbreite, z.B. jenes bei 730 cm^{-1}, welches die 1,3,5- Substitution des Aromaten beschreibt, im datenreduzierten Spektrum mit zu geringer Intensität dargestellt werden. Über den gesamten Bereich betrachtet gibt das datenreduzierte Spektrum das Originalspektrum jedoch sehr gut wieder. Trotzdem ist die Beschränkung auf 128 Punkte eine sehr grobe Näherung der spektroskopischen Wirklichkeit. Da gerade bei der Methodenentwicklung eine Vielzahl von Simulationsexperimenten durchgeführt wurden und werden, galt es einen Kompromiß zwischen ausreichender Auflösung und akzeptabler Rechenzeit zu finden. Dies stellt jedoch kein prinzipielles Problem dar, da eine einzelne Rechnung, wie sie bei den weiter unten aufgeführten Anwendungen durchgeführt wurde, absolut nicht zeitkritisch ist und größenordnungsmäßig im Minutenbereich liegt.

2.1.2 Strukturbeschreibung

Die verschiedenen Formen der Strukturcodierung haben zum Ziel, eine Strukturbeschreibung zu finden, die es ermöglicht, Struktur- und Spektreninformation zu korrelieren. Die Ergebnisse der Korrelationsuntersuchungen hängen dabei in hohem Maße von der Art der Strukturcodierung ab, inwieweit also infrarotrelevante Strukturinformation im Strukturcode enthalten ist. Wie bereits im Kapitel zur Einleitung erwähnt wurde, haben folgende Merkmale

Einfluß auf die Änderung des Dipolmoments und damit auf die Absorption von infraroter Strahlung:

- Atomeigenschaften (Ordnungszahl, Ladung, Polarisierbarkeit, etc.)
- Bindungsstärken

Ein geeignetes Codierungsverfahren sollte diese Strukturaspekte bei der Codierung berücksichtigen. Während es jedoch noch relativ einfach ist, Infrarotspektren durch eine konstante Zahl von Werten zu beschreiben, ist dies bei der Beschreibung von Strukturen schon wesentlich komplizierter: Der Länge des Strukturcodes eines Moleküls muß unabhängig von der Anzahl der Atome in dem Molekül sein. Die gebräuchliche Beschreibung von Molekülstrukturen durch kartesische Koordinaten kann nicht verwendet werden, da hier die Anzahl der Codewerte von der Zahl der Atome abhängt. Hinzu kommt noch die oben erwähnte Anforderung, daß der Strukturcode infrarotrelevante Inhalte der Strukturinformation wiedergeben soll. Die Ergebnisse der Korrelationsuntersuchungen werden in großem Maße von der Erfüllung dieser Anforderung abhängen. Eine Reihe von Ansätzen zur Strukturcodierung werden in den nachfolgenden Kapiteln beschrieben.

2.1.2.1 Substrukturbasierte Codierung

Bei der fragment- oder substrukturbasierten Codierung wird das zu codierende Molekül nach einer Liste von Substrukturen durchsucht.[11][12][13][14] Jede Substruktur der Liste entspricht dabei einer Variablen des Strukturcodes. Ist eine bestimmte Substruktur in dem Molekül enthalten, so wird die korrespondierende Variable auf einen bestimmten Wert gesetzt.

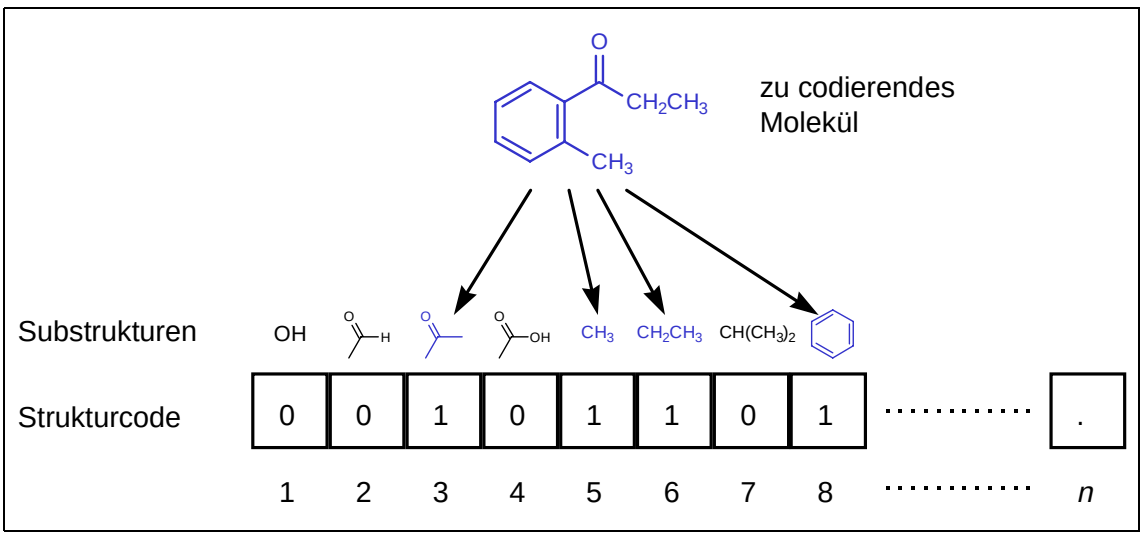


Abb. 2-2: Prinzip einer substrukturbasierten Strukturcodierung

Dieser Ansatz zur Strukturcodierung hat zwei wesentliche Nachteile: Zum einen ist die Anzahl an denkbaren Substrukturen prinzipiell unbegrenzt. Der Ansatz muß somit immer in gewissem Maße unvollständig bleiben. Zweitens ist eine *objektive* Auswahl an Substrukturen, die in die Liste aufgenommen werden sollen, sehr schwierig. Entsprechend variiert die Anzahl der berücksichtigten Substrukturen in verschiedenen Veröffentlichungen deutlich, z.B. 40 (vgl. Lit. [14]) oder 229 (vgl. Lit. [12]).

Untersuchungen von Affolter et al. [15] haben gezeigt, daß eine einfache Verknüpfung von einzelnen Strukturfragmenten und Spektrensignalen, wie sie z.B. bei Substruktur/Subspektrum-Datenbanken vorliegen, nicht zuverlässig funktioniert. Ein Grund liegt sicherlich darin, daß die Peaklage in hohem Maße von der Umgebung des Fragments, das diesem Signal entspricht, beeinflußt wird.

Dubois et al. [16] beschreiben den FREL- (**F**ragment **R**educed to an **E**nvironment that is **L**imited) Ansatz, welcher als Erweiterung der oben beschriebenen Substrukturcodierungs-Methode gesehen werden kann. Durch diese Codierungsmethode soll obigem Problem entgegengewirkt werden, indem verschiedene Zentren (Fokusse) innerhalb des Moleküls und ihre Umgebung beschrieben werden. Diese Zentren können Atome oder Bindungen sein. Dubois et al. berücksichtigen bei der Codierung die Atomzentren H, C, N, O, F, Cl, Br sowie I und erhalten somit durch die Verknüpfung mit Einfach-, Doppel,- Dreifach,- und aromatischen Bindungen 43 FREL-Fokusse.

2.1.2.2 Pfadlängenbasierte Codierung

Bei diesem Ansatz wird das Konnektivitätsdiagramm als Graph interpretiert, bei dem die Knoten den Atomen und die Kanten den Bindungen entsprechen. Die Numerierung der Knoten erfolgt willkürlich. Der resultierende Code beschreibt somit die Konstitution des Moleküls, wobei implizite H-Atome unterdrückt werden können. Während des Codierungsprozesses werden die verschiedenen im Molekül auftretenden Pfadlängen gezählt, wobei in der Regel eine maximale Pfadlänge festgelegt wird, die noch berücksichtigt werden soll. Die Pfadlänge wird immer für ein Atompaar bestimmt, indem die Anzahl der Bindungen zwischen diesen beiden Atomen ermittelt wird. Die ermittelten Häufigkeiten für das Auftreten der verschiedenen Pfadlängen > 0, wird schließlich durch zwei dividiert, da jede Pfadlänge bei einem systematischen Abarbeiten der Struktur zweimal gezählt wurde.[17]

2.1.2.3 3D-basierte Codierung

Ausgehend von der Tatsache, daß die Infrarotspektroskopie Schwingungen von Atomen und Molekülteilen im dreidimensionalen Raum widerspiegelt, wurde nach einer Form der

Codierung gesucht, welche die dreidimensionale Anordnung der Atome eines Moleküls beschreibt. Zusätzlich zu dieser 3D-Strukturinformation sollte der Code weitere infrarotrelevante Information, wie z.B. die partiellen Atomladungen, enthalten.

Ein Verfahren zur Auswertung von Elektronenbeugungsbildern war Ausgangspunkt für die Entwicklung einer derartige Transformationsmethode. Grundlage dafür waren Arbeiten von Debye,[18] die sich mit der Streuung von Röntgenstrahlen an amorphen Festkörpern beschäftigen. Darauf basierend, entwickelte Wierl [19] eine Funktion, die eine Berechnung der Intensitäten der Beugungsreflexe bei Elektronenbeugungsexperimenten in Abhängigkeit vom Beobachtungswinkel ermöglicht.

$$I(s) = K \sum_{i=2}^{N} \sum_{j=1}^{i-1} f_i f_j \cdot \int_0^{\infty} P_{ij}(R) \frac{\sin(sR)}{sR} dR$$

Berechnung der Beugungsintensitäten nach Wierl (Gl. 2-2)

mit:

$I(s)$ .....Intensität der Streustrahlung
N.....Anzahl der Atome
K.....Gerätekonstante des Meßinstruments
R.....Atomabstände
$P_{ij}(R)$. ....temperaturabhängige Wahrscheinlichkeitsverteilung der Atomabstände
f_i,f_j.....Formfaktoren der Atome i und j
s.....beugungswinkelabhängige Größe (vgl. Gl. 2-3)

Der Beugungswinkel θ wird durch die Größe s beschrieben:

$$s = 4\pi \sin(\theta/2)/\lambda$$

Beschreibung des Beugungwinkels θ durch s (Gl. 2-3)

mit:

θ.....Beugungswinkel
λ.....Wellenlänge

Ausgehend von Gleichung 2-2 entwickelten Soltzberg und Wilkins [20] ein Transforma-

tionsverfahren zur Untersuchung der Zusammenhänge zwischen Molekülstruktur und pharmakologischer Aktivität. Sie trafen dabei folgende Vereinfachungen:

- Die Gerätekonstante wird gleich 1 gesetzt
- Das Molekül ist starr, wodurch das Integral über die Aufenthaltswahrscheinlichkeiten der Atome $P_{ij}(r)$ in eine einfache Abstandsfunktion $\delta(r - r_{ij})$ übergeht

Zusätzlich verwendeten Soltzberg und Wilkins anstelle der Formfaktoren f_i die Ordnungszahlen Z_i der Atome. Bei der eigentlichen Codierung verwendeten sie jedoch nicht die Funktionswerte, sondern beobachteten, ob die Funktion in den einzelnen Intervallen, in die sie den Wertebereich unterteilt hatten, Nulldurchgänge hatte oder nicht. Das Ergebnis war ein binärer Vektor. Bei der Strukturcodierung, wie sie in unserer Arbeitsgruppe eingesetzt wird, folgten wir dieser letzten Vereinfachung nicht. Für die Ordnungszahlen Z_i werden ganz allgemein Atomeigenschaften A_i, wie z.B. die Gesamtladung q_{tot} oder die Masse m gesetzt. Man gelangt dabei zu folgender Gleichung:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \cdot \frac{\sin(sR_{ij})}{sR_{ij}}$$

Berechnung des 3D-MoRSE Codes (Gl. 2-4)

Dieser Code erhielt den Namen 3D-MoRSE Code, was für **3D-Mo**lecule **R**epresentation of **S**tructures based on **E**lectron diffraction steht.[21][22][23][24] In Abbildung 2-3 ist der 3D-MoRSE Code von Benzol dargestellt:

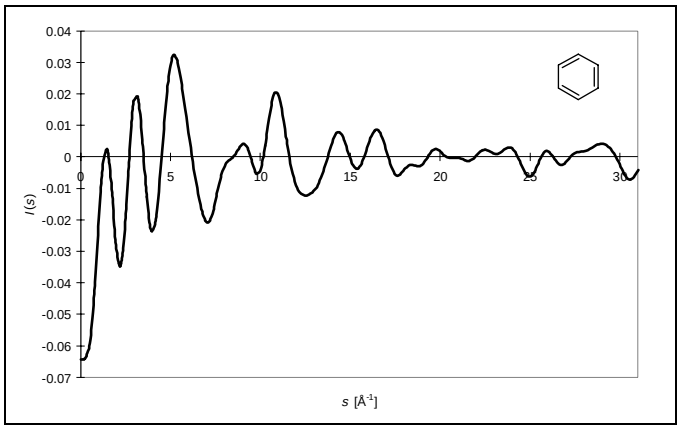


Abb. 2-3: 3D-MoRSE Code von Benzol

Zur Strukturcodierung wird obige Funktion diskretisiert, indem äquidistante Stützstellen der Funktion im Bereich von $s = 0$ bis s_{max} berechnet werden. Zusätzlich werden die Funktionswerte skaliert, worauf jedoch in Kapitel 2.4.1.1 näher eingegangen werden soll. In vorausgegangenen Arbeiten unserer Arbeitsgruppe konnte gezeigt werden, daß der Strukturcode sehr detaillierte Information über die 3D-Struktur von Molekülen enthält und sich beispielsweise auch zur Vorhersage von biologischen Eigenschaften eignet. [21][25] In Publikationen Náray-Szabó et al. und Csorvássy et al. wird beschrieben, wie sich ähnliche Strukturcodes zur Quantifizierung molekularer Ähnlichkeit einsetzen lassen.[26][27]

Arbeiten von V. Steinhauer [28][29] und M. C. Hemmer [30] beschreiben ebenfalls Korrelationsuntersuchungen zwischen Molekülstrukturen und Infrarotspektren mittels neuronalen Netzen. Ziel dieser Untersuchungen ist es, anhand des Infrarotspektrums Vorhersagen über die dreidimensionale Struktur zu treffen. Bei diesen Korrelationsexperimenten kommt die sogenannte Radialcodierung zum Einsatz. Der Radialcode wird wie folgt berechnet (vgl. Gl. 2-5):

$$g(R) = F_s \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i \cdot A_j \cdot e^{-B \cdot (R - R_{ij})^2}$$

Berechnung des Radialcodes (Gl. 2-5)

für: $R = 0, ..., R_{max}$

mit:
F_s ..... Skalierungsfaktor
N ..... Anzahl der Atome des Moleküls
A ..... Atomeigenschaften
B ..... Unschärfeparameter (entspricht dem Temperaturparameter des Experiments)
R_{max} ..... maximaler berücksichtigter Atomabstand
R_{ij} ..... Abstand zwischen Atom i und j

Beide Codierungsverfahren transformieren die kartesischen Atomkoordinaten unter Berücksichtigung physikochemischer Atomeigenschaften in den Strukturcode. Die zugrundeliegenden 3D-Molekülstrukturen werden mit dem 3D-Strukturgenerator CORINA (**CO**o**R**d**I-NA**tes) erzeugt. Die Erstellung der 3D-Strukturen mit CORINA geschieht anhand eines Regelsatzes, der wiederum auf kristallographischen Daten, Kraftfeldrechnungen und Molekülgeometrischen Grundregeln basiert. [31][32][33] Die entsprechenden physikochemischen Atomeigenschaften werden mit dem Programm PETRA (**P**arameter **E**valuation for the **T**reatment of **R**eactivity **A**pplications) berechnet.[34][35][36] Diese inkrementbasierte Methode ermöglicht eine schnelle Berechnung der gewünschten infrarotrelevanten physikochemischen

Größen, beispielsweise der Ladung q_σ, q_π und q_{tot} oder der Polarisierbarkeit α. Nachfolgende Abbildung zeigt den schematischen Ablauf der Erstellung der 3D-Struktur aus der Molekülzeichnung und die anschließende Berechnung der physikochemischen Atomeigenschaften (hier die Gesamtladung q_{tot}):

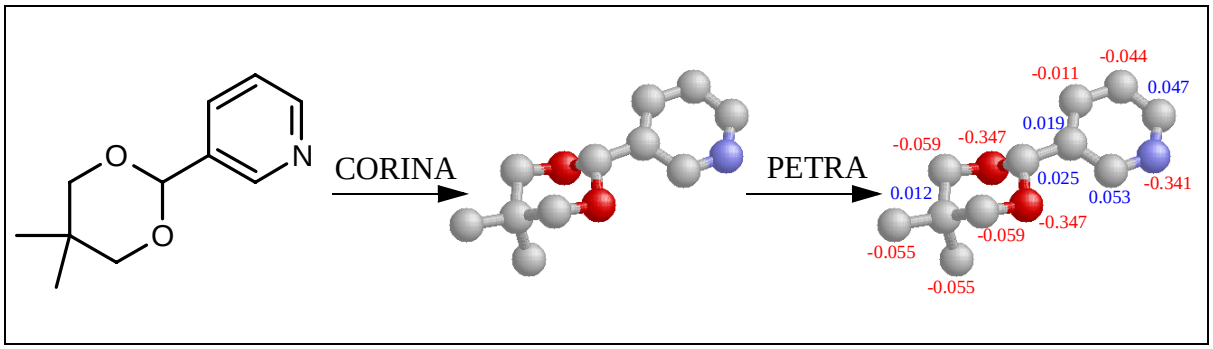


Abb. 2-4: Generierung der 3D-Struktur und Berechnung physikochemischer Atomeigenschaften (q_{tot})

2.2 Korrelationsmethoden

Mittels verschiedener Korrelationsmethoden können Struktur und Spektreninformation verknüpft und Vorhersagen getroffen werden.[37] Die Vorhersagen können derart sein, daß für eine bestimmte Molekülstruktur oder bei dem Vorhandensein verschiedener Strukturmerkmale, ein bestimmtes Spektrum oder den Strukturmerkmalen entsprechende Signale im Spektrum zu erwarten sind. Umgekehrt kann aus bestimmten Signalen im Spektrum auf gewisse Strukturmerkmale geschlossen werden. Im Prinzip ist das Verwenden von Tabellenwerken mit Substruktur/Subspektrenpaaren zur Interpretation von Infrarotspektren eine, wenn auch sehr einfache, in der qualitativen Analytik äußerst gängige Methode zur Korrelation von Struktur und Spektreninformation. Zwei Korrelationsmethoden, die in der Literatur Erwähnung finden, werden in den nachfolgenden Kapiteln näher erläutert.

2.2.1 Regelbasiert

Der Zusammenhang zwischen Struktur und Spektrum kann durch Regeln beschrieben werden. Der regelbasierte Ansatz ist dem manuellen Arbeiten mit Tabellenwerken sehr ähnlich, da das Aufstellen dieser Regeln anhand detaillierter Untersuchungen von Substruktur-Subspektrenpaaren erfolgt. In Arbeiten von Ricard et al. wird ein System beschrieben, das in der Lage ist, aus Struktur und Spektrendaten automatisch Regeln abzuleiten.[38] Die aufgestellten Regeln können positiv oder negativ sein: Das Vorhandensein eines bestimmten Signals im Spektrum zeigt das Vorhandensein eines bestimmten Strukturmerkmals an, ebenso, wie die

Abwesenheit eines Signals an einer bestimmten Stelle im Spektrum darauf hindeutet, daß auch das entsprechende Strukturmerkmal im Molekül fehlt. Die eben erläuterte Spektreninterpretation kann natürlich auch in umgekehrter Richtung betrieben werden und so zur Vorhersage von Banden, ausgehend von der Molekülstruktur dienen. In Arbeiten von Affolter et al. [15] wird beschrieben, daß diese einfache Verknüpfung von funktionellen Gruppen und Spektrensignalen jedoch nicht zuverlässig zur Spektreninterpretation einzusetzen ist. (vgl. auch Kap. 2.1.2.1)

2.2.2 Künstliche neuronale Netze

Das wesentliche Merkmal neuronaler Netze (NN) ist, daß sie in der Lage sind, über den Zusammenhang von zwei Merkmalen durch das selbständige Abarbeiten einer Reihe von Beispielen zu lernen.[39][40] Ausgehend von dem Gelernten können neuronale Netze dann Vorhersagen treffen. Wie ihre biologischen Gegenstücke, bestehen künstliche neuronale Netze (Artificial Neural Network ANN) aus Neuronen. Die Verbindung zwischen den Neuronen wird durch Gewichte beschrieben. Die Neuronen sind ebenso wie die Gewichte nicht tatsächlich physikalisch vorhanden. Sie sind vielmehr Variablen deren Inhalte sich, gesteuert durch Korrekturfunktionen, gegenseitig beeinflussen (Lernfähigkeit). Während des Lernvorgangs, dem Training, werden die Gewichte den Trainingsdaten angepaßt. Der Zusammenhang, in diesem Fall zwischen Molekülstruktur und Infrarotspektrum, wird nicht in einer expliziten Gleichung niedergelegt, sondern ist in den Gewichten des trainierten Netzes gespeichert. Der Vorteil dieses induktiven Lernverfahrens ist, daß kein a priori Wissen über den untersuchten Zusammenhang nötig ist. Ein neuronales Netz kann sich also die Fähigkeiten für verschiedene Aufgaben (z.B. Klassifikation, Spektrenvorhersage) anhand von Trainingsbeispielen selbst aneignen und muß nicht für jede Aufgabe neu programmiert werden. Weiterhin besteht gegenüber herkömmlichen Computersystemen der Vorteil, daß die Speicherung von Information assoziativ und nicht adreßbezogen stattfindet, wodurch eine Robustheit gegenüber verrauschten Daten erreicht wird. Diese Eigenschaften machen neuronale Netze zu einem hervorragend geeigneten Werkzeug zur Datenverarbeitung in der Chemie.[41][42] In vielen Bereichen der Chemie liegen experimentelle Daten (z.B. Strukturen) vor, die mit anderen experimentellen Daten (z.B. Spektren, chemische Reaktionsfähigkeiten, biologische Aktivitäten) in engem Zusammenhang stehen, welcher sich jedoch nicht über eine einfache Funktion beschreiben läßt. Derartige Korrelationen lassen sich mittels eines neuronalen Netzes gut modellieren.[43] Trainierte neuronale Netze stellen somit ein Mittel zur Transformation von Mustern dar. Es wird ein Eingabemuster, in diesem Fall ein Vektor dessen Werte die Molekülstruktur beschreiben, in ein Ausgabemuster, ein Vektor dessen Werte die Absorbanzwerte des vorhergesagten Infrarotspektrums darstellen, transformiert.

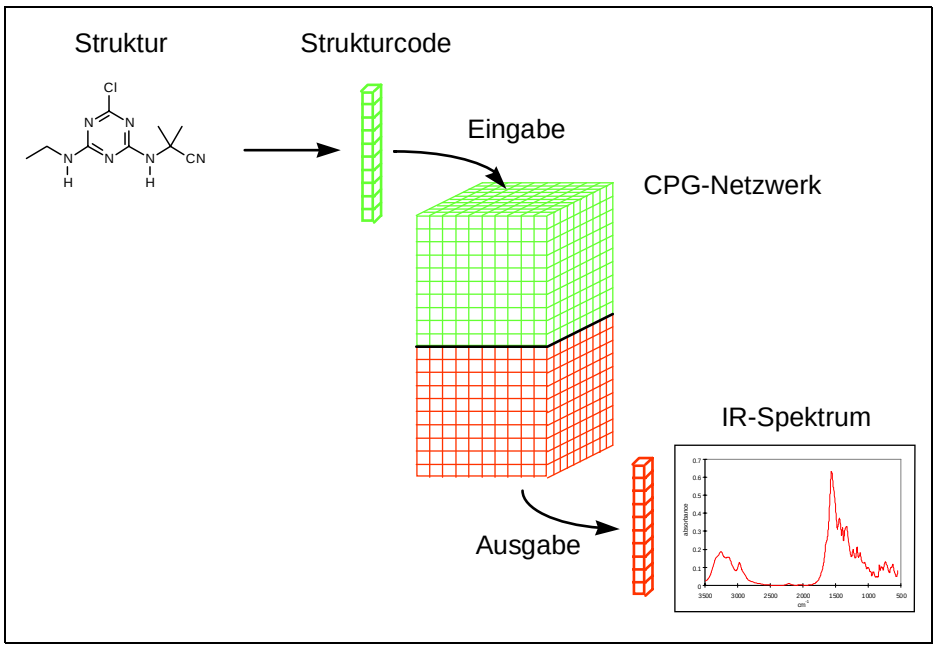


Abb. 2-5: Schema der Vorhersage eines Infrarotspektrums aus dem Strukturcode eines Moleküls

Es existieren verschiedene Netzwerktypen, z.B. Hopfield-,[44] Kohonen-,[45][46][47] Backpropagation-[40][42] oder Counterpropagation-Netzwerke,[50] die sich in ihrer Architektur und ihrer Funktionsweise sowie in der Art des Trainings unterscheiden. Bei den in dieser Arbeit beschriebenen Untersuchungen wurden ausschließlich Counterpropagation Netze eingesetzt. Der Aufbau dieses Netzwerktypus sowie das für die Vorhersageergebnisse wesentliche Training wird im nachfolgenden erläutert.

Counterpropagation-(CPG)-Netze sind in der Regel aus einer Kohonenschicht und einer Ausgabeschicht aufgebaut. Die Kohonenschicht ist aus x mal y z-dimensionalen Variablen, den Neuronen, aufgebaut. Jedem Neuron der Kohonen-(Eingabe)-schicht entspricht ein Neuron in der Ausgabeschicht.

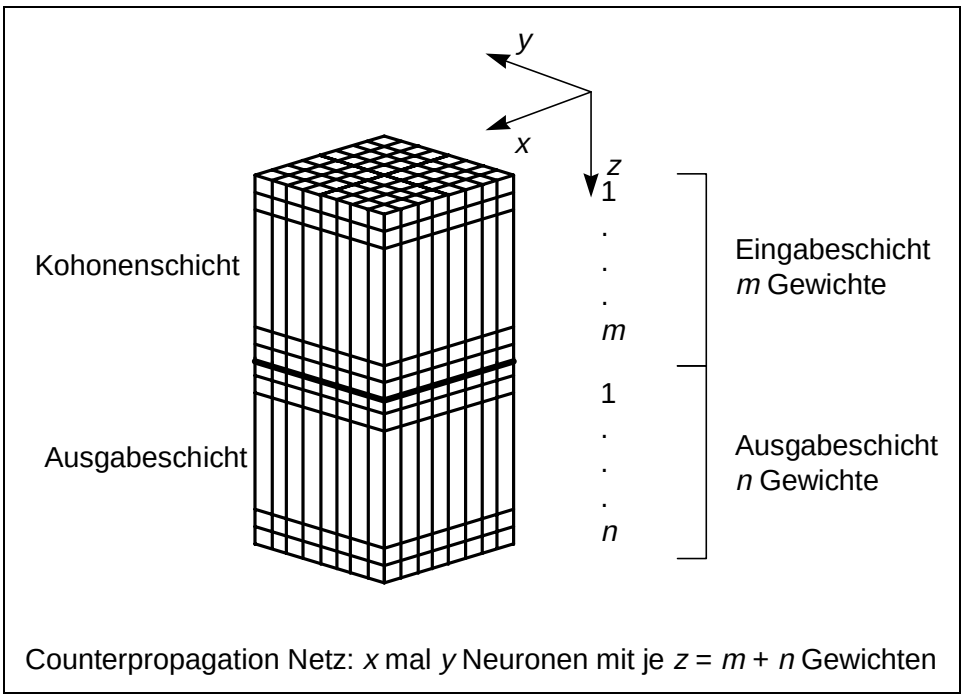


Abb. 2-6: Schematischer Aufbau eines Counterpropagation-(CPG)-Netzes

Während des Netztrainings wird nun für jeden Datenpunkt (Strukturcode des Moleküls und Infrarotspektrum) des Trainingsdatensatzes das ähnlichste Neuron bestimmt. Das ähnlichste Neuron ist jenes, mit der geringsten euklidischen Distanz zum jeweiligen Trainingsdatenpunkt. Wird die euklidische Distanz nur zwischen dem Strukturteil des Eingabevektors und dem Eingabeblock des CPG-Netzes bestimmt, so spricht man von unüberwachtem (unsupervised) Lernen. Überwachtes (supervised) Lernen liegt vor, wenn die euklidische Distanz zwischen dem gesamten Eingabevektor (Strukturcode und Infrarotspektrum) und dem Ein- und Ausgabeblock des CPG-Netzes bestimmt wird.

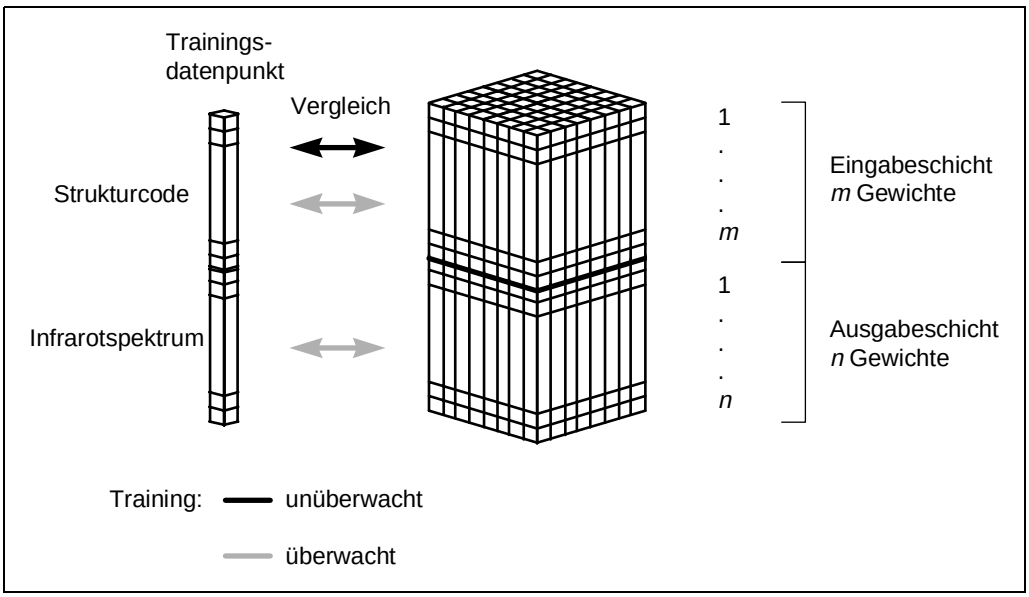


Abb. 2-7: Auswahl des ähnlichsten Neurons bei unüberwachtem und überwachtem Lernen

Ist das ähnlichste Neuron ermittelt, so werden die Gewichte der Neuronen den Werten des Trainingsdatenpunktes angepaßt. Der Anpassungsgrad ist für das ähnlichste Neuron *c* am größten und nimmt mit zunehmendem topologischen Abstand der Neuronen zum Neuron *c* ab. Die Anpassung der Gewichte erfolgt dabei gemäß Gleichung 2-6.[41]

$$c_{ji}^{angepasst} = c_{ji}^{alt} + \eta(t)a(d_c - d_j)(y_i - c_{ji}^{alt})$$

Anpassung der Neuronengewichte (Gl. 2-6)

mit:
c_{ji} ..... Gewicht *i* von Neuron *j*
t ..... Zeit
$\eta(t)$ ..... zeitabhängige Lernrate
y_i ..... Dimension *i* des Trainingsdatenpunktes *y*
$a(d_c - d_j)$ ..... nachbarschaftsabhängige Funktion zur Beschreibung des Abstands zwischen Neuron *j* und dem Gewinnerneuron *c*

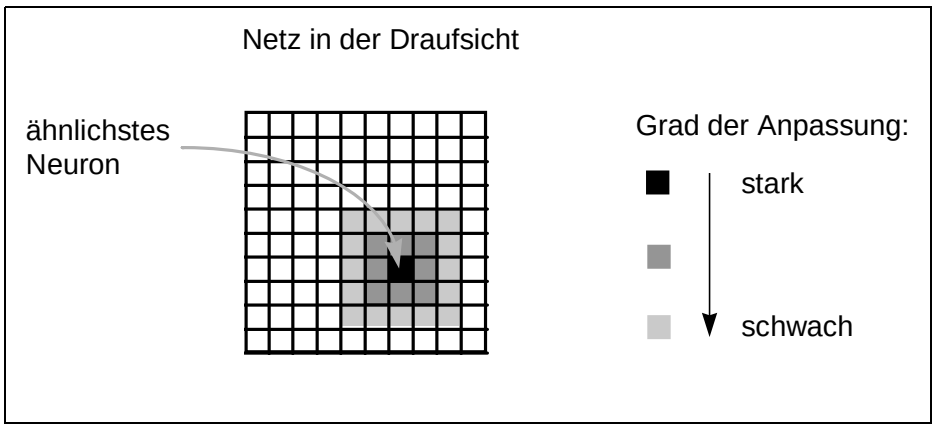


Abb. 2-8: Anpassung der Neuronengewichte

Angepaßt werden sowohl Eingabe-(Struktur)- als auch Ausgabe-(Spektren)-Block des Netzes. Durch diese Anpassung gelangt Information über den Zusammenhang zwischen Struktur und Spektrum in das Netz.

Bei der Vorhersage eines Infrarotspektrums wird der Strukturcode des Anfragemoleküls mit dem Strukturteil des trainierten CPG-Netzes verglichen. Aus der Ausgabeschicht des ähnlichsten Neurons wird das simulierte Infrarotspektrum entnommen.

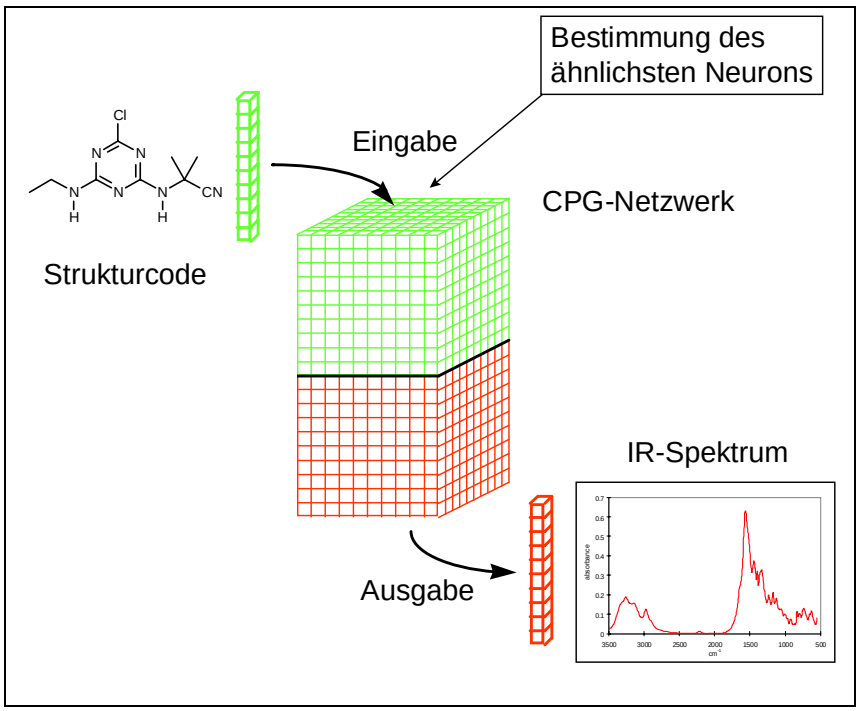


Abb. 2-9: Schema der Simulation eines Infrarotspektrums mittels eines neuronalen Counterpropagation-(CPG)-Netzes

Diese Vorhersagemethode hat den großen Vorteil, daß sie prinzipiell für nahezu beliebige Moleküle eingesetzt werden kann. Da die Methode datenbasiert ist, sich ihr Wissen also selbständig aus Beispielen ableitet, ist die Qualität und die Geschwindigkeit der Vorhersage unabhängig von der Molekülgröße. So können also auch für Moleküle mit komplizierten Gerüsten sehr gute Simulationsergebnisse erzielt werden, sofern ähnliche Moleküle im Trainingsdatensatz enthalten sind. Wie bereits erwähnt, ist die benötigte Rechenzeit unabhängig von der Molekülgröße.

2.3 Datenvergleich

Der Vergleich von Daten ist ein zentraler Punkt der Methode, da er an zwei Schlüsselschritten entscheidend eingreift: der Vergleich von Strukturdaten ist das dominierende Kriterium für den Ablauf des Netztrainings, bei dem es in erster Linie darum geht, Ähnlichkeiten zwischen Datenpunkten zu erkennen und diese im Netz abzubilden. Ebenso basiert die Auswahl des Neurons, aus dessen Ausgabeschicht das simulierte Spektrum entnommen wird, auf dem Vergleich des Strukturcodes der Anfragestruktur mit dem Strukturteil des trainierten neuronalen Netzes. Weiterhin ist der Vergleich (siehe Gleichung 2-7 und 2-8) von Infrarotspektren das Kriterium für Entscheidungen bei der Methodenoptimierung, da die Simulationsgüte durch den Vergleich von experimentellem und simuliertem Infrarotspektrum bestimmt wird. Aufgrund der unterschiedlichen Natur der Daten werden auch unterschiedliche Anforderungen an das Vergleichsmaß gestellt. Dieser Aspekt wird in den beiden folgenden Kapiteln näher beleuchtet.

2.3.1 Spektrendaten

Automatisierte Spektrenvergleiche gestalten sich schwierig, da viele für den Spektroskopiker augenfälligen Vergleichskriterien zwar vorhanden, jedoch schwierig objektiv beschreibbar sind. Bei der Suche nach einem geeigneten Vergleichsmaß für Infrarotspektren gilt es sich zu vergegenwärtigen, welche Merkmale in einem Spektrum auftreten und wie diese verursacht werden. Entsprechend muß abgewogen werden, welche Merkmale Information über die Struktur des vermessenen Moleküls liefern und somit vom Vergleichsmaß berücksichtigt werden müssen und welche wenig oder nichts über das Probemolekül aussagen und somit auch vom Vergleichmaß möglichst nicht beachtet werden sollten. Ein geeignetes Vergleichsmaß für IR-Spektren sollte die Besonderheiten von IR-Spektren berücksichtigen. Es sollte also bestimmte spektralen Unterschiede (z.B. absolute Intensitäten) möglichst ignorieren, da diese auf präparative Ursachen, wie z.B. unterschiedliche Schichtdicken, zurückzuführen sind. Auf andere

Spektrendifferenzen (z.B. Fehlen eines Signals, unterschiedliche Bandenmuster), die durch strukturelle Unterschiede verursacht werden, sollte das Vergleichsmaß wiederum sehr empfindlich reagieren. Weiterhin sollte dieses Vergleichsmaß, da es bei jedem Simulationsexperiment eine Vielzahl von Vergleichen anzustellen gilt, einfach und schnell zu berechnen sein. In der analytischen Praxis werden zwei Vergleichsmaße vermehrt zum Vergleich von IR-Spektren eingesetzt:

- Der *rms*-(root mean square error)-Wert:

$$rms = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{x,i} - E_{y,i})^2}$$

Berechnung des *rms*-Werts (Gl. 2-7)

mit:

n.....Anzahl der Spektrenpunkte (hier $n = 128$)
$E_{x,i}$, $E_{y,i}$..Absorbanzwerte der zu vergleichenden Spektren *x* bzw. *y*

- Der Korrelationskoeffizient *r* nach Bravais-Pearson: [51][52][53][54]

$$r = \frac{\sum_{i=1}^{n} E_{x,i}E_{y,i} - \left(\frac{\sum_{i=1}^{n} E_{x,i} \sum_{i=1}^{n} E_{y,i}}{n}\right)}{\sqrt{\sum_{i=1}^{n} E_{x,i}^2 - \frac{\left(\sum_{i=1}^{n} E_{x,i}\right)^2}{n}}\sqrt{\sum_{i=1}^{n} E_{y,i}^2 - \frac{\left(\sum_{i=1}^{n} E_{y,i}\right)^2}{n}}}$$

Berechnung des Korrelationskoeffizienten *r* (Gl. 2-8)

mit:

n.....Anzahl der Spektrenpunkte (hier $n = 128$)
$E_{x,i}$, $E_{y,i}$..Absorbanzwerte der zu vergleichenden Spektren *x* bzw. *y*

Der *rms*-Wert, den wir bei unseren anfänglichen Arbeiten verwendeten, zeigte bezüglich der Bewertung von Spektrendifferenzen Schwächen. So sind oftmals Simulationen zunächst durch das Raster des Vergleichsmaßes durchgefallen, da der hohe *rms*-Wert eine schlechte Simulation anzeigte. Eine visuelle Inspektion ergab dann jedoch, daß sich simuliertes und experimentelles Spektrum zwar in ihren absoluten Intensitäten unterschieden, im Bandenmuster jedoch sehr ähnlich waren. In anderen Fällen wurden intensitätsschwache Spektren als ähnlich bewertet, obwohl sich einzelne Signale in ihren Lagen deutlich unterschieden oder in einem der beiden Spektren fehlten.

Der Korrelationskoeffizient *r* (Bravais Pearson) ist prinzipiell unempfindlich gegen Unterschiede in den absoluten Intensitäten des Gesamtspektrums. Das Fehlen oder die Verschiebung intensitätsschwacher Signale wird jedoch oftmals nur ungenügend berücksichtigt. Bezüglich dieser Fragestellung wurde bei den nachfolgenden Experimenten untersucht, in welchem Maße die beiden Vergleichsmaße auf gezielte Veränderungen spektraler Merkmale reagieren. Beide Werte haben die Eigenschaft, daß sie mit den reinen Absorbanzwerten des Spektrums arbeiten und keine Kennung oder Wichtung bezüglich Peakpositionen enthalten. Dies kann einerseits ein Vorteil sein, da keine Vorbehandlung des Spektrums mit einem Algorithmus zur Signalauswahl (engl. Peak-Picker), mit all den Fehlerquellen, wie Subjektivität bei der Signalauswahl und sonstigen Schwachpunkten, mit denen ein solches System behaftet sein kann, notwendig ist. Andererseits ist zu erwarten, daß rein mathematische Verfahren, die auf keinerlei spektroskopisch-chemischem Wissen aufsetzen, gerade feine Bandenmuster, wie sie für ein Infrarotspektrum typisch sind und welche auch wesentliche Information tragen können, nicht in ausreichendem Maße beachten. Um zu untersuchen, wie die beiden Vergleichsmaße auf Spektrenunterschiede reagieren, wurde ein experimentelles Spektrum auf verschiedene Weisen systematisch verändert. Die Veränderung erfolgte schrittweise, wobei der Grad der Veränderung zunahm. Zwischen den modifizierten Spektren und dem Ausgangsspektrum wurden dann jeweils die Vergleichsmaße berechnet. Als experimentelles Spektrum wurde das Spektrum eines Cyclohexyl-substituierten Esters genommen, da diese Verbindung neben den typischen Grundschwingungen auch eine Vielzahl komplexer Gerüstschwingungen aufweist, was zu einer hohen Signaldichte im Fingerprintbereich führt. Zur Untersuchung der Vergleichsmaße wurden folgenden Experimente durchgeführt.

Experiment 1: Veränderung der Intensität des gesamten Spektrums

Die Absorbanzwerte des Gesamtspektrums werden schrittweise um jeweils 20% reduziert (vgl. Abb. 2-10).

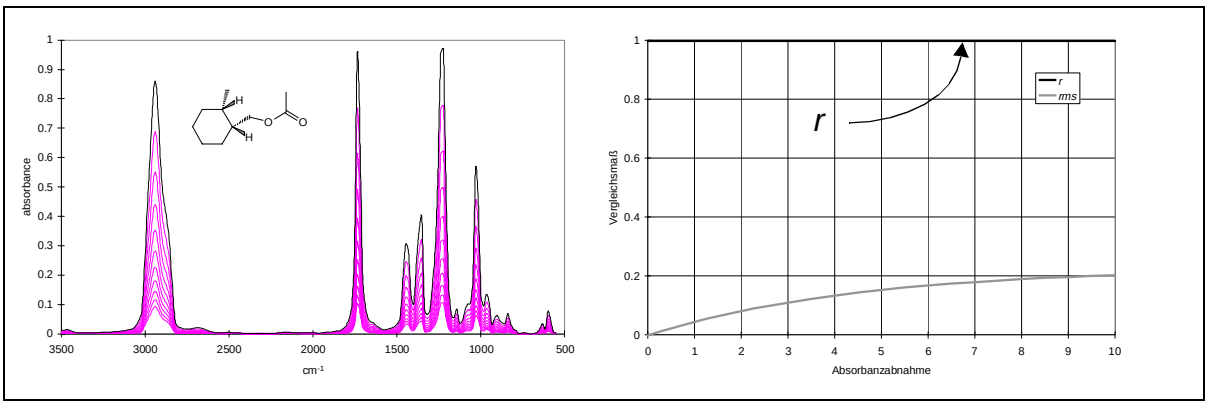


Abb. 2-10: Veränderung der Intensität des gesamten Spektrums

Diskussion der Ergebnisse von Experiment 1:

Der Korrelationskoeffizient *r* reagiert auf diese Art der Spektrenveränderung überhaupt nicht. Der *rms*-Wert nähert sich dem Wert von 0.229 an, was einem Vergleich des Spektrums mit der Nullinie entspricht. Die Steigung des *rms*-Graphen wird zunehmend flacher, was einerseits daran liegt, daß die Abnahme von Spektrum zu Spektrum jeweils prozentual erfolgt und andererseits daran, daß die Intensitäten des Gesamtspektrums abnehmen. Derartige Spektrendifferenzen werden im Experiment durch unterschiedliche Schichtdicken bzw. Konzentrationen verursacht und beschreiben keine Strukturmerkmale. Dieses Computerexperiment ist mit einer Absorbanzabnahme auf 11% des Ursprungsspektrums sehr weit gegangen. Ein Spektroskopiker wird in der Regel die Messung mit mehr Substanz wiederholen. Ein Vergleich solcher Spektren wird also selten vorkommen, kann allerdings auch nicht ausgeschlossen werden. Ein Vergleich von Spektren in der Ursprungsform und dem ersten oder zweiten Reduktionsschritt entsprechen der analytischen Praxis. Der *rms*-Wert reagiert bereits bei den ersten beiden Schritten und signalisiert möglicherweise eine zu hohe Spektrendifferenz. Zusammenfassend kann bemerkt werden, daß der Korrelationskoeffizient *r* auf die Abweichungen in Experiment 1 besser reagiert als der *rms*-Wert und den Vergleich realistischer bewertet.

Experiment 2: Veränderung der Intensität eines dominanten Signals

Die Absorbanzwerte eines intensitätsstarken Signals werden schrittweise um jeweils 20% reduziert (vgl. Abb. 2-11).

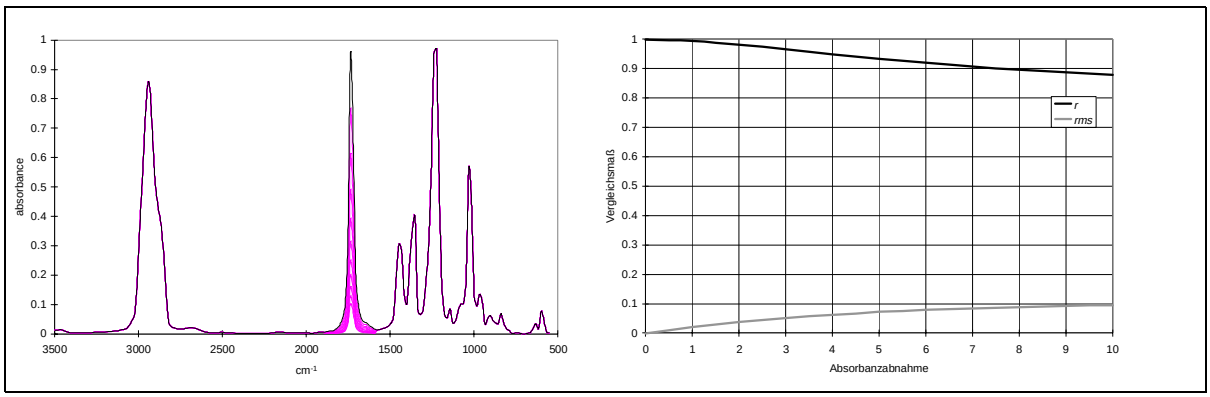


Abb. 2-11: Veränderung der Intensität eines dominanten Signals

Diskussion der Ergebnisse von Experiment 2:

Beide Vergleichsmaße reagieren bei den ersten fünf Schritten relativ ähnlich auf die Veränderung des Spektrums. Bei den weiteren Veränderungen strebt der *rms*-Wert deutlich einem Grenzwert von etwa 0.1 zu, der einem *rms*-Wert zwischen dem Ursprungsspektrum und dem Spektrum, bei dem das Carbonylsignal bei 1720 cm^{-1} völlig fehlt, entspricht. Für den Korrelationskoeffizient *r* liegt der entsprechende Wert bei 0.85, wobei sich ein asymptotischer Verlauf des Graphen erst etwa bei Schritt neun bis zehn andeutet. Derartige Veränderungen von Spektrensignalen können in der Praxis durch strukturelle Veränderungen sowie durch Lösungsmitteleinflüsse verursacht werden. Ersteres muß durch das Vergleichsmaß angezeigt werden, während der zweite Fall eher toleriert werden sollte. Da dies jedoch nicht durch das System entschieden werden kann, muß es für derartige Spektrenveränderungen eine Spektrendifferenz signalisieren. Dies wird durch den *rms*-Wert besser erfüllt, da dieser eher seinem Grenzwert entgegenstrebt.

Experiment 3: Veränderung der Intensität eines intensitätsschwachen Signals

Die Absorbanzwerte eines intensitätsschwachen Signals werden schrittweise um jeweils 20% reduziert:

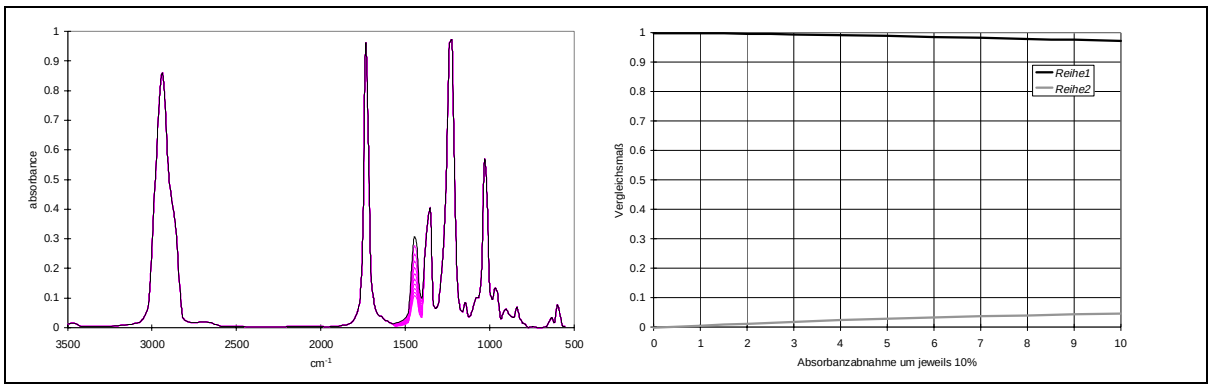


Abb. 2-12: Veränderung der Intensität eines intensitätsschwachen Signals

Diskussion der Ergebnisse von Experiment 3:

Beide Vergleichsmaße reagieren sehr wenig auf das Verschwinden eines intensitätsschwachen Signals. Dies ist zwar verständlich, da ein sehr großer Teil des Spektrums dekkungsgleich ist, ist jedoch als sehr ungünstig zu bewerten, da kleine Signale oftmals Aussagen über die An- oder Abwesenheit von funktionellen Gruppen, z.B. Nitrilfunktionen, ermöglichen.

Experiment 4: Absenken der Intensitäten eines Teilbereichs mit geringen Intensitäten

Bei diesem Experiment wurde der langwellige Randbereich des Spektrums (1128 - 552 cm^{-1}), ein Bereich mit weitgehend geringen Absorbanzwerten, schrittweise um 20% abgesenkt.

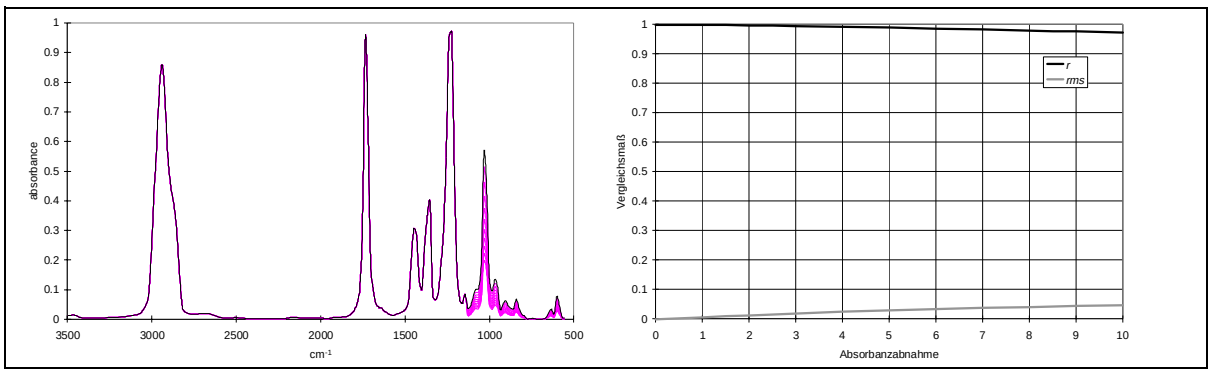


Abb. 2-13: Absenken der Intensitäten eines Teilbereichs mit geringen Intensitäten

Diskussion der Ergebnisse von Experiment 4:

Eine derartige Spektrenveränderung ist mehr auf apparative Ursachen wie Defekte an der Lichtquelle bzw. am Detektor oder auch unterschiedliches Küvettenmaterial zurückzuführen als auf strukturelle Unterschiede. Der Korrelationskoeffizient *r* reagiert hier etwas schwächer als der *rms*-Wert, was wegen den unveränderten Signalmustern als günstig anzusehen ist. Objektiv betrachtet ist das Verhalten der beiden Vergleichsmaße hier zu ähnlich, um eine Wertung vornehmen zu können.

Experiment 5: Hypsochrome Verschiebung des gesamten Spektrums

Das gesamte Spektrum wird hypsochrom schrittweise im Abstand der Stützstellen verschoben. Dies führt im Bereich von 3500-2000 cm^{-1} um eine Verschiebung von je 40 cm^{-1} und im Bereich von 2000-552 cm^{-1} zu einer Verschiebung von je 16 cm^{-1}.

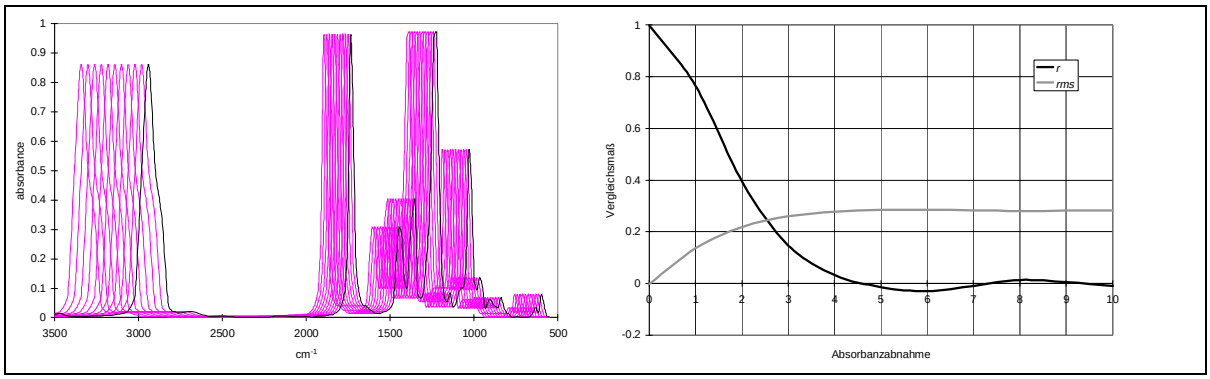


Abb. 2-14: Hypsochrome Verschiebung des gesamten Spektrums

Diskussion der Ergebnisse von Experiment 5:

Beide Vergleichsmaße reagieren sehr stark auf die Veränderung des Spektrums und pendeln sich jeweils bei Schritt fünf bei ihren jeweiligen Werten für maximale Unähnlichkeit ein. Daß dies gerade beim fünften Schritt geschieht ist leicht zu verstehen, da ab dieser Stufe der Spektrenveränderung die veränderten Spektren nichts mehr mit dem Ursprungsspektrum gemein haben. Die einzelnen Signale sind dann so stark verschoben, daß sie auch in ihren Flankenbereichen nicht mehr überlappen. Der Korrelationskoeffizient *r* wird zwar ab dem sechsten Schritt wieder besser und deutet so auf eine vermeintlich geringere Spektrendifferenz hin. Tatsächlich ist dieser Effekt jedoch zufällig und wird dadurch verursacht, daß Signale soweit verschoben werden, so daß sie bereits wieder mit anderen Signalen des Ursprungsspektrums überlagern. Die Lage und die relativen Intensitäten von Signalen sind hochcharakteristische Spektrenmerkmale. Änderungen dieser Merkmale deuten auf erhebliche

Strukturveränderungen hin und sollten daher von einem Spektrenvergleichsmaß entsprechend stark gewertet werden, wie es sowohl für den *rms*-Wert als auch für den Korrelationskoeffizienten *r* der Fall ist.

Experiment 6: Hypsochrome Verschiebung eines intensitätsstarken Signals

Der Carbonylpeak, ein intensitätsstarkes Signal bei 1720 cm^{-1}, wird schrittweise um je 16 cm^{-1} hypsochrom verschoben.

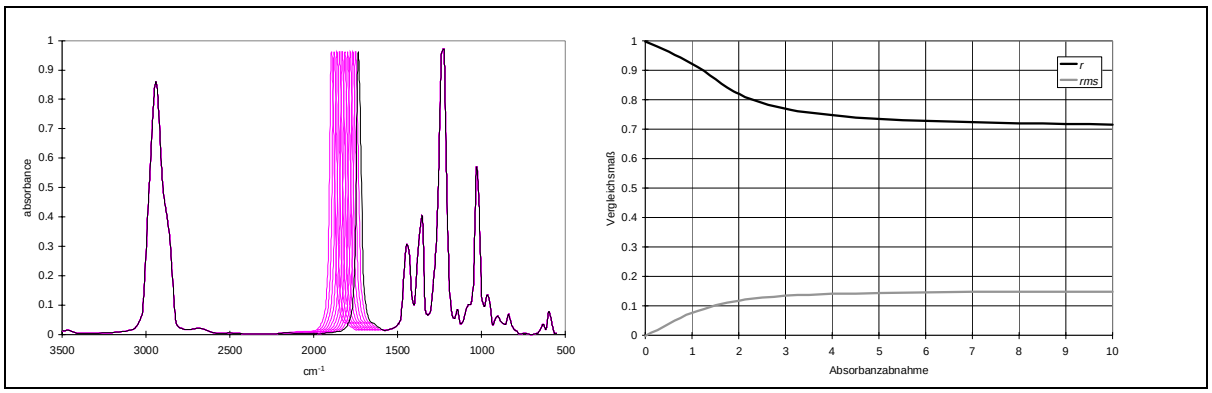


Abb. 2-15: Hypsochrome Verschiebung eines dominanten Signals

Diskussion der Ergebnisse von Experiment 6:

Verschiedene Lösungsmitteleinflüsse oder die Aufnahme in verschiedenen Medien können die Verschiebung eines Signals im Spektrum bewirken. Ebenso können unterschiedliche funktionelle Gruppen, die z.B. eine unterschiedliche Elektronegativität besitzen, die Lage des Signals einer benachbarten Gruppe beeinflussen. Letzteres hätte jedoch sicherlich noch weitere Spektrendifferenzen zur Folge als nur die Verschiebung eines Signals. Leichte Verschiebungen eines Signals durch Lösungsmitteleinflüsse sollten durch ein Spektrenvergleichsmaß möglichst toleriert werden. Dies wird durch beide Vergleichsmaße erfüllt. Bei weiteren Verschiebungen fallen auch die Reaktionen der Vergleichsmaße stärker aus, bis sie bei Werten von 0.71 (Korrelationskoeffizient *r*) und 0.15 (*rms*-Wert) eine Art Sättigung erreichen. Während ein Korrelationskoeffizient von r = 0.7 bereits eine sehr hohe Spektrendifferenz anzeigt, ist der *rms*-Wert mit 0.15 noch relativ gering.

Experiment 7: Hypsochrome Verschiebung eines intensitätsschwachen Signals

Ein intensitätsschwaches Signal wird schrittweise um 16 cm^{-1} hypsochrom verschoben.

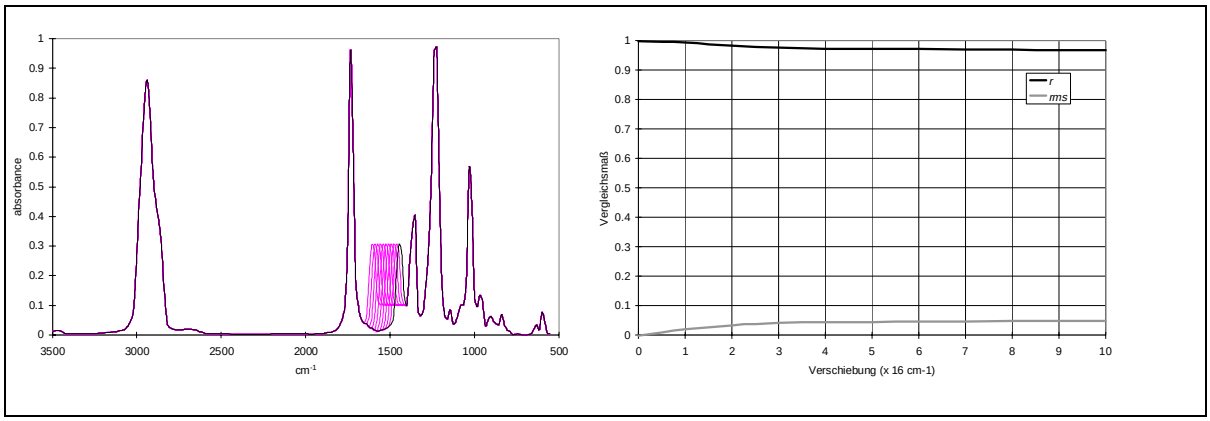


Abb. 2-16: Hypsochrome Verschiebung eines intensitätsschwachen Signals

Diskussion der Ergebnisse von Experiment 7:

Derartige Signalverschiebungen können ähnliche Ursachen haben, wie sie bereits unter Experiment 6 diskutiert wurden. Hier reagieren beide Vergleichsmaße sehr schwach. Eine stärkere Reaktion würde der analytischen Praxis besser entsprechen.

Experiment 8: Entstehen eines Signals aus der Flanke einer Bande

Aus der Flanke des Carbonylsignals wurde ein zusätzliches Signal aufgebaut. Die Zunahme erfolgte schrittweise um je 20% bezogen auf das vorherige Spektrum.

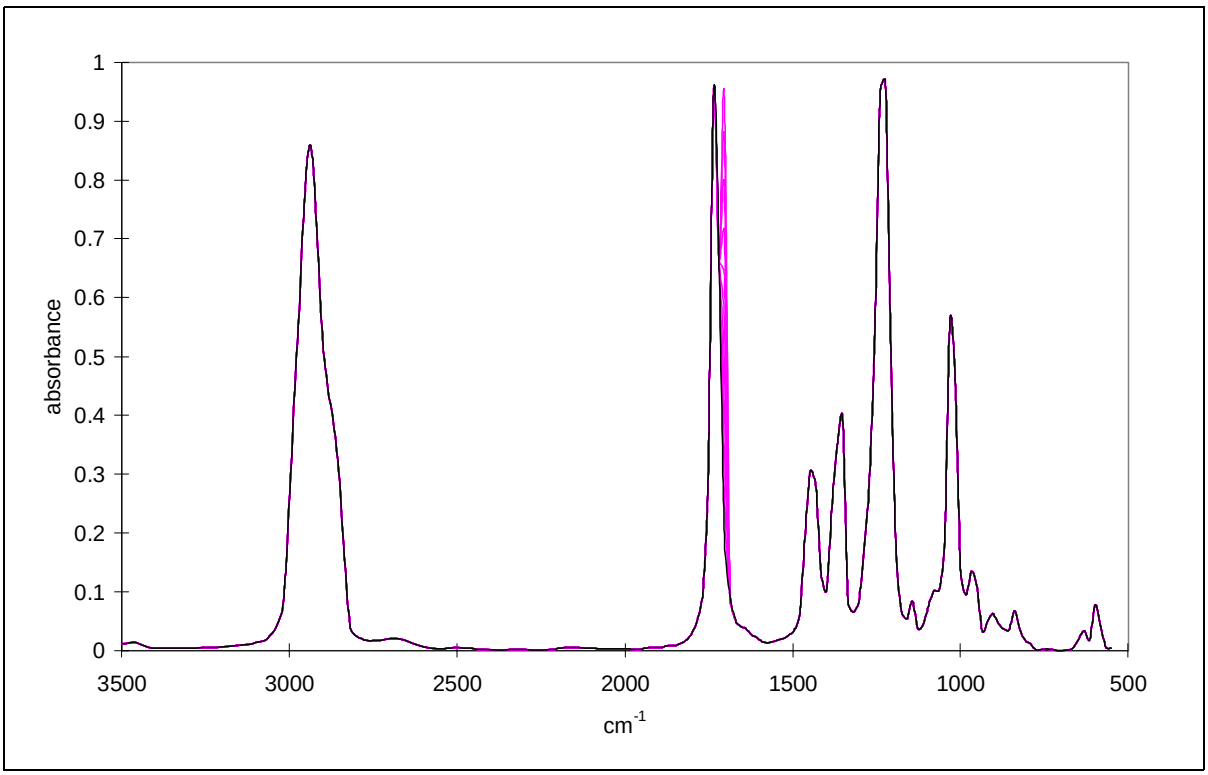


Abb. 2-17: Entstehen eines Signals aus der Flanke einer Bande

Im Gesamtspektrum sind die einzelnen Schritte kaum zu sehen. Aus diesem Grund wurde der Bereich von 2000 - 1500 cm^{-1} vergrößert dargestellt:

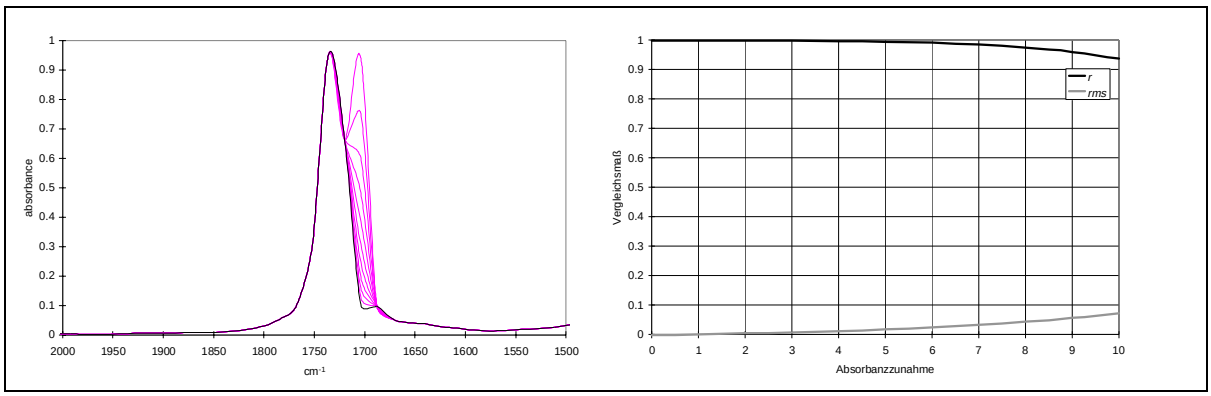


Abb. 2-18: Entstehen eines Signals aus der Flanke einer Bande (vergrößerter Ausschnitt)

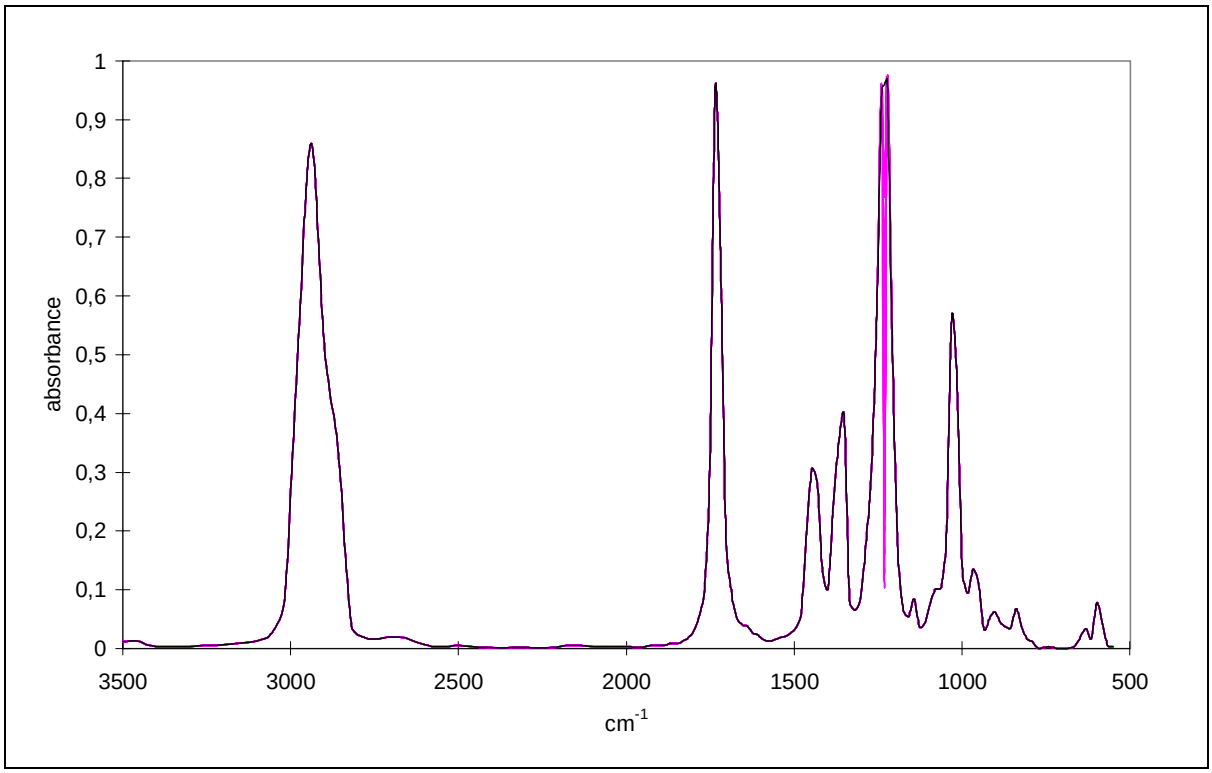
Diskussion der Ergebnisse von Experiment 8:

In IR-Spektren von Esterverbindungen sind häufig Carbonylsignale zu beobachten, die eine Schulter tragen. Feinstrukturen dieser Art stellen also wichtige Information dar. Beide

Vergleichsmaße reagieren erst sehr schwach, mit fortschreitender Spektrenveränderung jedoch zunehmend stärker. Der *rms*-Wert reagiert bereits ab Schritt 4 bis 5 als sich die Bande beginnt zu verbreitern. Der Korrelationskoeffizient *r* zeigt erst ab Schritt 8 bis 9 eine Reaktion als sich langsam ein Extremum im Spektrenverlauf auszubilden beginnt. Die zunehmende Reaktion ab dem achten Schritt ist bei beiden Vergleichsmaßen zu beobachten, weshalb deren Verhalten bei diesem Experiment als gut bewertet wird.

Experiment 9: Aufspaltung eines Signals in zwei Einzelsignale

Ein intensitätsstarkes Signal im Fingerprint-Bereich bei 1232 cm^{-1} (ν C-O-C) wird in zwei Einzelsignale aufgespalten. Der Einschnitt in der Mitte des Ursprungssignals erfolgt in 20%-Schritten.



Zwecks einer übersichtlicheren Darstellung wird auch hier eine Vergrößerung des Bereichs von 1500 - 1000 cm^{-1} abgebildet:

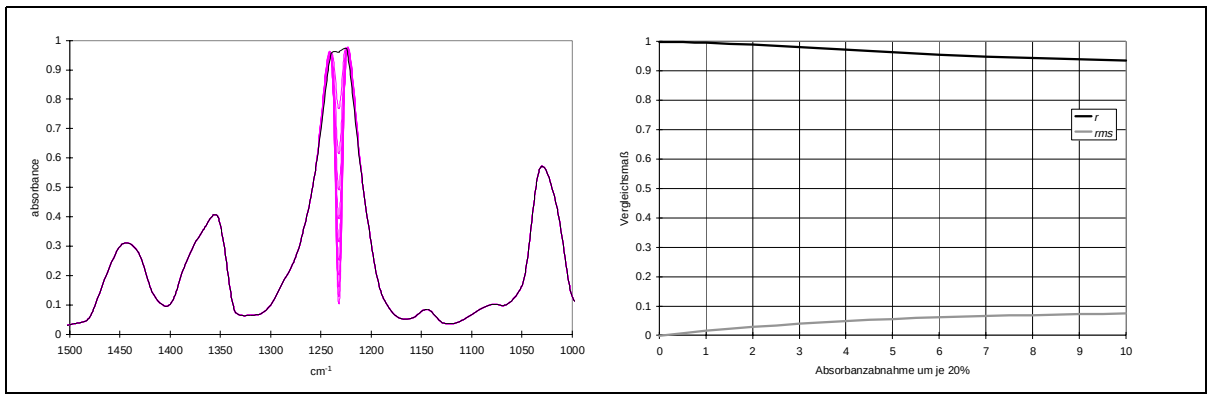


Abb. 2-19: Aufspaltung eines Signals in zwei Einzelsignale

Diskussion der Ergebnisse von Experiment 9:

Eine derartige Aufspaltung eines Signal ist weniger auf strukturelle Unterschiede zurückzuführen als auf eine Verbesserung der Auflösung. Umgekehrt kann das Zusammenfließen zweier Signale durch eine ungünstige Lage der Stützstellen nach einer Datenreduktion verursacht werden (vgl. Kap. 2.1.1.). Die Reaktion beider Vergleichsmaße fällt deutlich aus. Dies ist als negativ zu bewerten, da sich im Ursprungsspektrum bereits zwei Maxima andeuten und so auf das mögliche Vorhandensein zweier Einzelsignale hinweisen.

Experiment 10: Verschiebung der Basislinie

Bei diesem Experiment wurde der Verlauf der Basislinie verändert. Dazu wurde der Absorbanzwert des hochfrequenten Rands des Spektrums schrittweise um 0.04 angehoben.

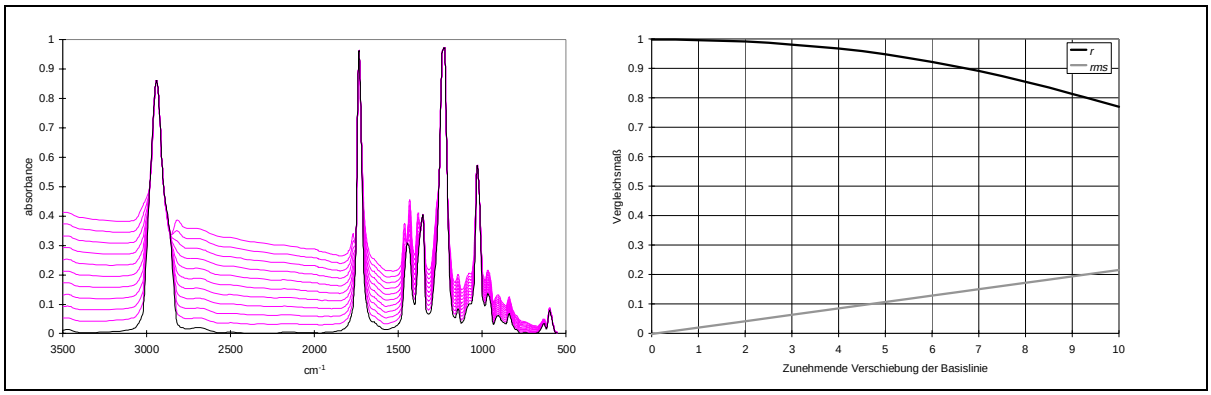


Abb. 2-20: Verschiebung der Basislinie

Diskussion der Ergebnisse von Experiment 10:

Ein Verschiebung der Basislinien wird nicht durch strukturelle Unterschiede verursacht,

sondern ist auf experimentelle Parameter, wie z.B. Defekte an Lichtquelle oder Detektor sowie eine mangelhafte Probenpräparation zurückzuführen. Inwieweit es sinnvoll ist, Spektren derartig schlechter Qualität überhaupt zu einem Vergleich heranzuziehen, ist eine ganz andere Frage. Während der erste und zweite Schritt der Spektrenveränderung noch toleriert werden kann, sollte in den anderen Fällen das Spektrum ein weiteres Mal aufgenommen werden. Steht kein Probenmaterial zur Verfügung so besteht weiterhin die Möglichkeit, das Spektrum elektronisch nachzubearbeiten und eine Basislinienkorrektur durchzuführen, wobei dann stets darauf hingewiesen werden muß, daß es sich nunmehr um ein modifiziertes Spektrum handelt. Interessant ist die unterschiedliche Reaktion der beiden Vergleichsmaße: Während der rms-Wert linear ansteigt, verläuft die Kurve des Korrelationskoeffizienten hyperbelförmig. Der Verlauf des *rms*-Werts ist leicht nachvollziehbar, stellt jedoch eine schlechte Beschreibung der Realität dar. Das Verhalten des Korrelationskoeffizienten *r* kommt einer praxisgerechten Einschätzung der Spektrenähnlichkeiten wesentlich näher: Während die beiden ersten Schritte der Spektrenveränderung nahezu keine Auswirkung auf *r* haben, fällt der Wert bei größeren Veränderungen rapide ab.

In nachfolgender Tabelle sind die Ergebnisse der Experimente 1-10 zusammengefaßt und bewertet. Die Bewertung ist mit gut (+), mittel (o) und schlecht (-) eingeteilt.

Tab. 2-1: Bewertung der Vergleichsmaße

Beschreibung des Experiments	*rms*	*r*
Veränderung der Intensität des gesamten Spektrums	+	-
Veränderung der Intensität eines dominanten Signals	o	+
Veränderung der Intensität eines intensitätsschwachen Signals	-	-
Absenken der Intensitäten eines Teilbereichs mit geringen Intensitäten	o	o
Hypsochrome Verschiebung des gesamten Spektrums	+	+
Hypsochrome Verschiebung eines intensitätsstarken Signals	+	o
Hypsochrome Verschiebung eines intensitätsschwachen Signals	o	o
Entstehen eines Signals aus der Flanke einer Bande	+	+
Aufspaltung eines Signals in zwei Einzelsignale	-	-
Verschiebung der Basislinie	+	o

Die verschiedenen Untersuchungen konnten Schwachstellen und Stärken der beiden Vergleichsmaße bei den Bewertungen von Spektrenähnlichkeiten aufzeigen. Die Bewertung

ist jedoch zu einem gewissen Grade sicher subjektiv und kann in manchen Fällen nur in Zusammenhang mit einer konkreten Fragestellung getroffen werden. Im Fall der Spektrensimulation ist die Anforderung an das Vergleichsmaß relativ genau definiert: Einem simulierten Spektrum einer Verbindung soll bei einem Vergleich mit verschiedenen experimentellen Spektren das entsprechende experimentelle Spektrum als ähnlichstes zugeordnet werden. Um also objektiv bewerten zu können, welches der beiden Vergleichsmaße Ähnlichkeiten von IR-Spektren besser erkennt, wurde ein Datensatz von 403 Molekülen der SpecInfo IR-Datenbank, die genau zwei Spektreneinträge enthalten (ca. 12970 Moleküleinträge enthalten nur ein Spektrum), untersucht. Diese 806 IR-Spektren wurde mit allen 13373 Spektren der Datenbank verglichen. Für jedes Spektrum wurde eine Liste mit den zehn ähnlichsten IR-Spektren der Datenbank erstellt. Zur Bewertung eines Vergleichsmaßes wurde untersucht, auf welchen Platz das jeweilige Zweitspektrum in der Top-10-Liste gesetzt wurde. Die Plazierungen wurden registriert und abschließend über alle 806 Vergleiche gemittelt. Weiterhin wurde gezählt, wieviele der Zweitspektren als so unähnlich bewertet wurden, daß sie gar nicht unter den ähnlichsten zehn Spektren zu finden waren.

Bei diesem Vergleich schneidet der Korrelationskoeffizient *r* deutlich besser ab als der *rms*-Wert.

Tab. 2-2: Vergleich von *rms*-Wert und Korrelationskoeffizient

Vergleichsmaß	mittlere Plazierung (gemittelt über 806 Vergleiche)	Anzahl der Moleküle außerhalb der Top-10-Liste
rms-Wert	5.52	351
r	3.27	145

IR-Spektren zeigen in weiten Bereichen oftmals keine Signale. Nimmt man eine Darstellungsbreite von 40 cm pro Spektrum an, so enthält die SpecInfo IR-Datenbank mit rund 13000 Einträgen an ungeladenen H, C, N, O, Hal-Verbindungen etwa 1.5 km Grundlinie. Dieser wenig aussagekräftige Grundlinienvergleich verzerrt die Spektrenbeurteilung zusätzlich. In weiteren Versuchen wurde deshalb untersucht, inwieweit das Ausblenden eines Spektrenbereichs die Trefferquote des Korrelationskoeffizienten beim Auffinden des Zweitspektrums verbessern kann. Es wurden Versuche angestellt, bei denen der Bereich zwischen 2540 und 1832 cm^{-1} oder 2740 und 1800 cm^{-1} ausgeblendet wurden. Um die Grenzen für den auszublendenden Bereich festzulegen, wurde ein Mittelwertspektrum über alle IR-Spektren der SpecInfo Datenbank ermittelt. Im ersten Versuch wurde der Spektrenbereich ausgeblendet, in welchem das Mittelwertspektrum eine Absorbanz $E < 0.04$ hat. Im zweiten Versuch wurde der Spektrenbereich mit einer Absorbanz $E < 0.02$ beim Berechnen des Korrelationskoeffizienten ausge-

nommen.

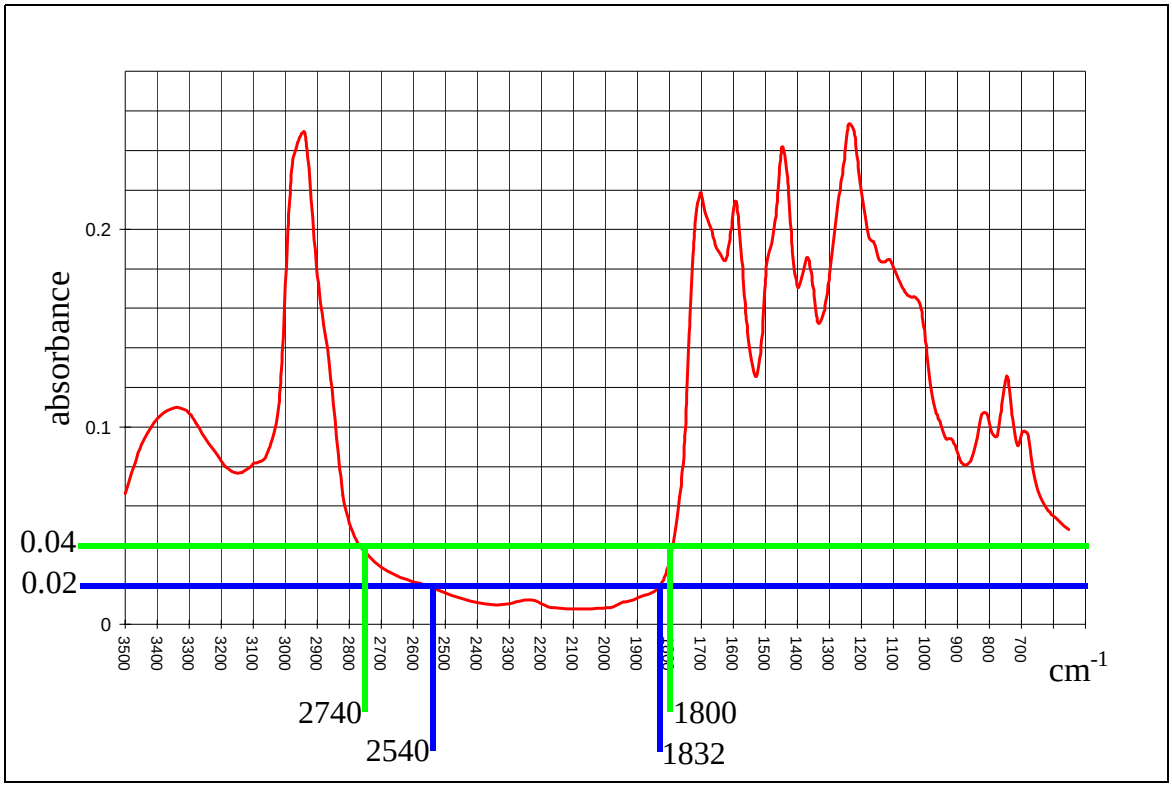


Abb. 2-21: Gemitteltes Spektrum über alle IR-Spektren der SpecInfo Datenbank

Die Maßnahme der Ausblendung ist vor dem Hintergrund, daß eine Vielzahl der Spektren in diesem Bereich keine Signale aufweisen, sinnvoll. Andererseits birgt das Ausblenden das Problem, daß gerade in diesem Bereich die Signale für Dreifachbindungen, Nitrile und Alkine zu finden sind. Welcher dieser konträren Aspekte den dominanteren Einfluß auf das Wiederfindungsverhalten des Korrelationskoeffizienten, hat, soll durch den nachfolgenden Versuch geklärt werden. Auch hier wurde, wie bereits oben beschrieben, ein Spektrenvergleichsexperiment mit den 403 Spektrenpaaren durchgeführt. Die Ergebnisse sind in Tabelle 2-3 aufgeführt.

Tab. 2-3: Vergleich der bereichsgewichteten Korrelationskoeffzienten

Grenzwert der Absorbanz *E*	ausgeblendeter Spektrenbereich	mittlere Plazierung	Anzahl der Moleküle außerhalb der Top-10-Liste
0.04	2740 - 1800 cm^{-1}	3.24	142
0.02	2540 - 1832 cm^{-1}	3.25	142
ohne Ausblendung		3.27	145

Eine Ausblendung des Spektralbereichs von 2740 - 1800 cm^{-1} führt zu einer weiteren wenn auch extrem geringen Verbesserung des Vergleichsmaßes. Bei der Auswahl geeigneter Codierungs- und Simulationsparameter wird daher der bereichsgewichtete Korrelationskoeffizient r_b gemäß Gleichung verwendet.

$$r_b = \frac{\sum_{i=1}^{n} w_i [E_{x,i} - \bar{E}_x][E_{y,i} - \bar{E}_y]}{\sqrt{\left(\sum_{i=1}^{n} w_i (E_{x,i} - \bar{E}_x)^2\right)\left(\sum_{i=1}^{n} w_i (E_{y,i} - \bar{E}_y)^2\right)}}$$

Berechnung des bereichsgewichteten Korrelationskoeffizienten r_b (Gl. 2-9)

mit:

n.....Anzahl der Spektrenpunkte (hier n = 128)
$E_{x,i}$, $E_{y,i}$..Absorbanzwerte der zu vergleichenden Spektren x bzw. y
$\bar{E}_x$, $\bar{E}_y$.....Mittelwerte der Absorbanzwerte der zu vergleichenden Spektren x bzw. y
w_i.....Bereichsgewichtungen

Zur Ausblendung des Wellenzahlenbereichs von 2740 - 1800 cm^{-1} werden die Wichtungen w_{20}-w_{50} gleich 0 gesetzt. Aufgrund der oben angesprochenen Problematik, daß durch das Ausblenden dieses Spektralbereiches beim Vergleich auch relevante Signale ignoriert werden könnten und zugleich um eine erhöhte Transparenz und Vergleichbarkeit, z.B. mit anderen Arbeiten bei Publikationen, zu gewährleisten, wird bei den nachfolgenden Simulationsexperimenten der Korrelationskoeffizient über den gesamten Spektrenbereich von 3500 cm^{-1} bis 552 cm^{-1} berechnet. Lediglich bei den Anwendungsbeispielen in Kapitel 3.3, welches die Fragestellung einer Substanzidentifikation behandelt, soll der bereichsgewichtete Korrelationskoeffizient verwendet werden.

2.3.2 Strukturdaten

Ganz ähnlich wie bei dem Vergleich von Infrarotspektren gilt es bei dem Vergleich von Strukturcodes die Merkmale des Codes zu berücksichtigen, die infrarotrelevante Strukturinformation beschreiben. Sowohl Radial- als auch 3D-MoRSE Code beschreiben die dreidimensionale Anordnung der Atome eines Moleküls im Raum. Eine Einschränkung wie sie bei dem Vergleich von Infrarotspektren, nämlich dem Ignorieren von absoluten Intensitäten getroffen

werden konnte, ist hier nicht mehr sinnvoll, da die absoluten Intensitäten des Strukturcodes die Häufigkeit für das Auftreten von bestimmten interatomaren Distanzen beschreibt. Als Vergleichsmaß wird der *rms*-Wert berechnet. Er entspricht der auf die Anzahl der Codewerte normierten euklidischen Distanz. In den nächsten Abbildungen sind die 3D-MoRSE Codes und Radialcodes für Benzol, Ethylbenzol und n-Butylbenzol dargestellt. Im ersten Experiment wurden die Radialcodes mit der Atomeigenschaft $A_i = q_{tot}$ und den Codierungsparametern B = 100 und R = 12.8 berechnet.

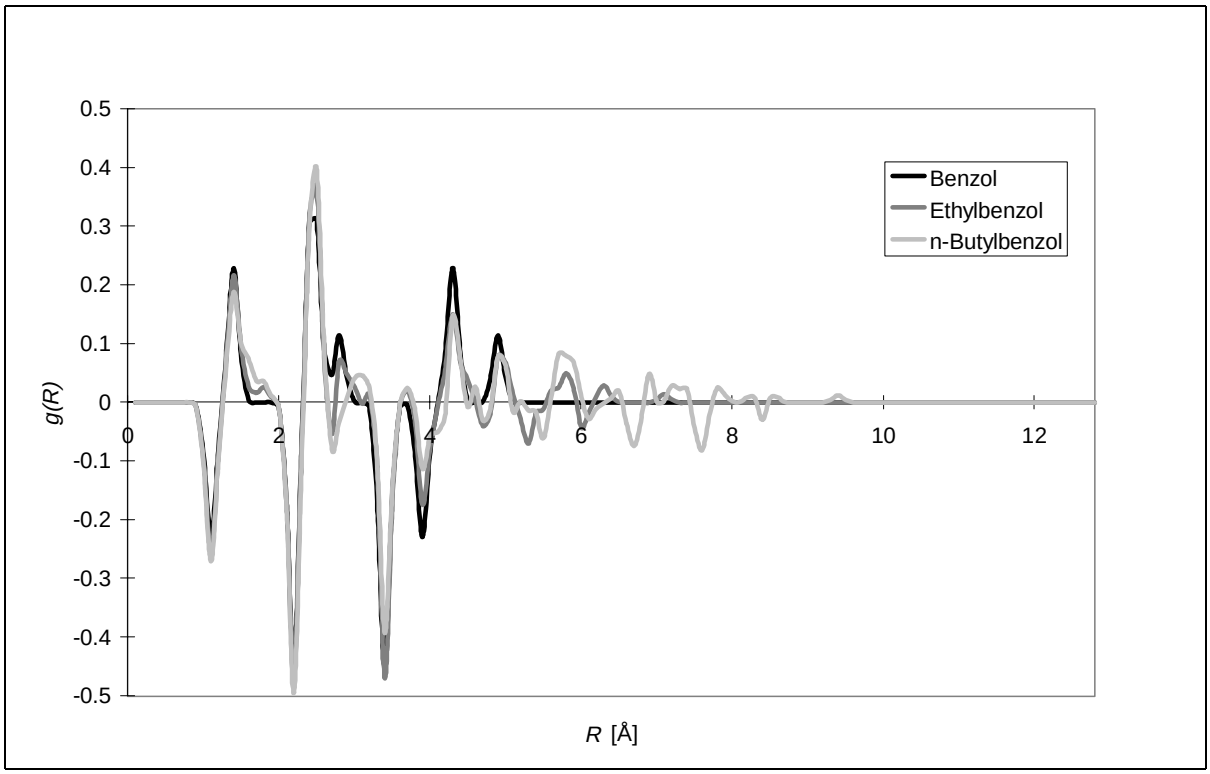


Abb. 2-22: Radialcodes von Benzol, Ethylbenzol und n-Butylbenzol

Für die verschiedenen Molekülpaare wurden die *rms*-Werte zwischen den entsprechenden Radialcodes berechnet und in nachfolgender Tabelle aufgetragen:

Tab. 2-4: *rms*-Werte zwischen den verschiedenen Radialcodes

	Benzol	Ethylbenzol	n-Butylbenzol
Benzol	0	0.022	0.036
Ethylbenzol	0.022	0	0.026
n-Butylbenzol	0.036	0.026	0

Bei den *rms*-Werten ist zu beobachten, daß die Radialcodes von Benzol und Ethylbenzol

sowie von Ethylbenzol und n-Butylbenzol als ähnlicher gewertet werden, als die von Benzol und n-Butylbenzol. Diese Wertung der Ähnlichkeiten entspricht dem chemischen Empfinden, da der Substituent systematisch aufgebaut wird ($H \rightarrow C_2H_5 \rightarrow C_4H_9$) und somit die Ähnlichkeit der Moleküle über die Ähnlichkeit der Substituenten bestimmt wird.

Das gleiche Experiment wurde mit dem 3D-MoRSE Code (vgl. Gl. 2-4) durchgeführt, wobei dieser mit $s = 31.0$ Å^{-1} und der Atomeigenschaft $A_i = q_{tot}$ berechnet wurde. Die entsprechenden 3D-MoRSE Codes sind in nachfolgender Abbildung dargestellt:

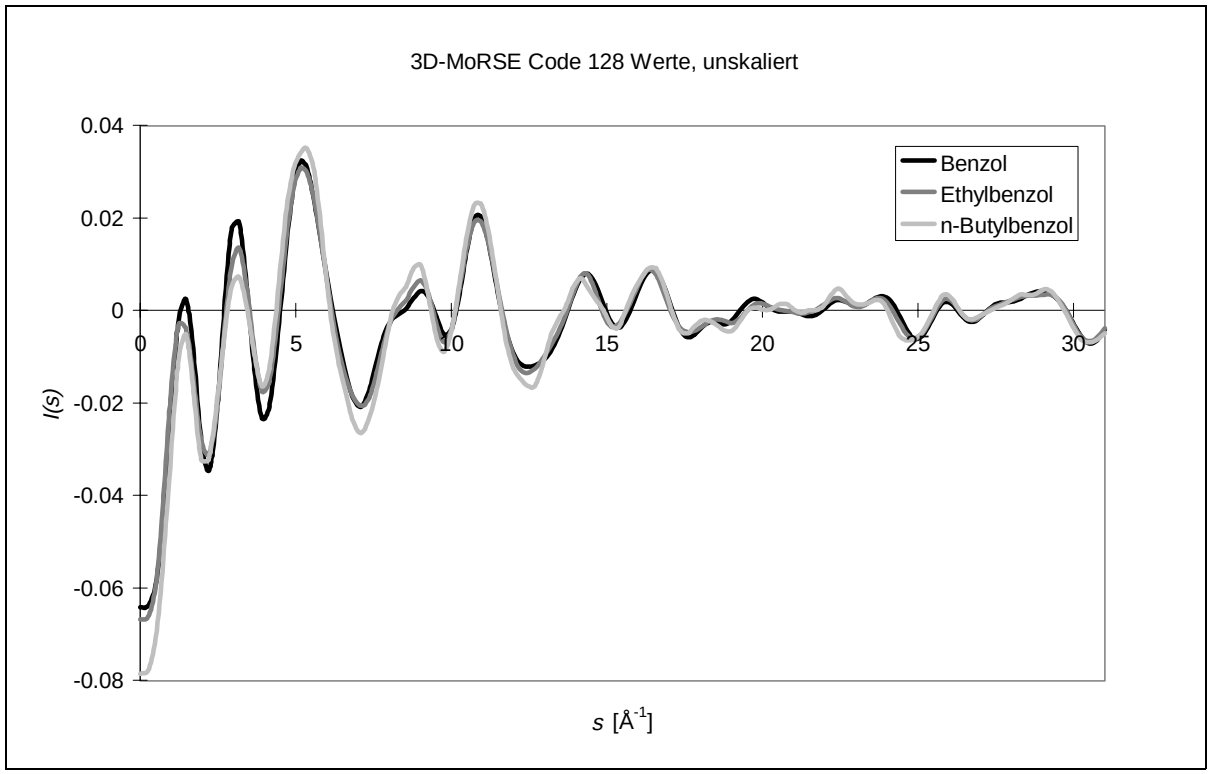


Abb. 2-23: 3D-MoRSE Codes von Benzol, Ethylbenzol und n-Butylbenzol

Für die verschiedenen Molekülpaare wurden die *rms*-Werte zwischen den entsprechenden 3D-MoRSE Codes berechnet und in nachfolgender Tabelle aufgetragen:

Tab. 2-5: *rms*-Werte zwischen den verschiedenen 3D-MoRSE Codes

	Benzol	Ethylbenzol	n-Butylbenzol
Benzol	0	0.032	0.065
Ethylbenzol	0.032	0	0.049
n-Butylbenzol	0.065	0.049	0

Auch hier ist zu beobachten, daß die *rms*-Werte zwischen den Radialcodes von Benzol und Ethylbenzol sowie von Ethylbenzol und n-Butylbenzol niedriger sind, als die von Benzol und n-Butylbenzol. Weiterhin fällt auf, daß in den Extrema der Codewerte, die Kurve für Ethylbenzol zwischen den Kurven von Benzol und n-Butylbenzol liegt. Die strukturellen Ähnlichkeiten, nämlich, daß die Paare Benzol/Ethylbenzol und Ethylbenzol/n-Butylbenzol einander ähnlicher sind als das Paar Benzol/n-Butylbenzol entsprechen also der Bewertung durch den *rms*-Wert zwischen den Funktionswerten des Codes sowie dem Verlauf der Kurven.

2.4 Simulation von Infrarotspektren mit neuronalen Netzen

Wie in den vorausgegangenen Kapiteln bereits erklärt wurde, wird bei der Spektrensimulation mittels eines neuronalen Netzes das Netz mit einer Reihe von Trainingsmolekülen und den dazugehörigen Spektren trainiert. Während des Netztrainings werden die Gewichte des Netzes den Trainingsdatenpunkten angepaßt, wodurch das Netz lernt, den Zusammenhang zwischen Molekülstruktur und Infrarotspektrum zu modellieren. Die tragende Rolle, die in diesem Zusammenhang die Funktionen zum Vergleich von Strukturcodes und Infrarotspektren spielen, wurde schon mehrfach erwähnt. Die angestellten Untersuchungen zur Auswahl geeigneter Vergleichsmaße für Infrarotspektren wurden bereits in Kapitel 2.3.1 beschrieben. In den Kapiteln zur Strukturcodierung (vgl. Kap. 2.1.2) wurde bereits darauf hingewiesen, daß der Strukturcode möglichst viel infrarotrelevante Information enthalten soll. Der Informationsgehalt des Codes, läßt sich durch die Variation der Codierungsparameter, wie z.B. die verwendete Atomeigenschaft, beeinflussen. Die folgenden Untersuchungen hatten zum Ziel, möglichst geeignete Codierungsparameter für die IR-Spektrensimulation zu ermitteln.

2.4.1 Auswahl geeigneter Codierungsparameter

2.4.1.1 3D-MoRSE Code

Bei der Berechnung des 3D-MoRSE Codes gemäß Gleichung 2-4 kann der Informationsgehalt des Codes durch die drei folgenden Parameter variiert werden:

- Atomeigenschaft A_i ; Wichtungsfaktor für die jeweilige Atomposition
- Anzahl der Codewerte n
- s_{max}; Maximalwert der Laufvariable s (vgl. Gl. 2-4)

Die minimale Differenz zwischen interatomaren Abständen ΔR_{min}, die noch aufgelöst und damit durch den Code beschrieben werden kann, ist abhängig von s_{max}:[49]

$$\Delta R_{min} = \frac{\pi}{s_{max}}$$

s_{max} Abhängigkeit von ΔR_{min} (Gl. 2-10)

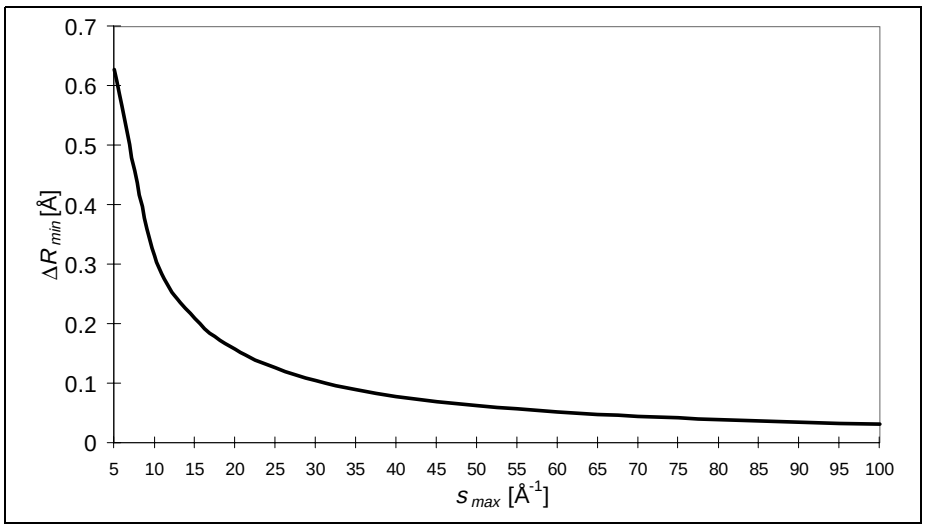


Abb. 2-24: s_{max} Abhängigkeit von ΔR_{min}

Der maximale Abstand zweier Atome R_{max}, der noch durch die Codierung erfaßt werden kann, ist von s_{max} und der Zahl der Codewerte n abhängig:

$$R_{max} = \frac{n\pi}{s_{max}}$$

Abhängigkeit des maximal erfaßbaren Atomabstands R_{max} von s_{max} und n (Gl. 2-11)

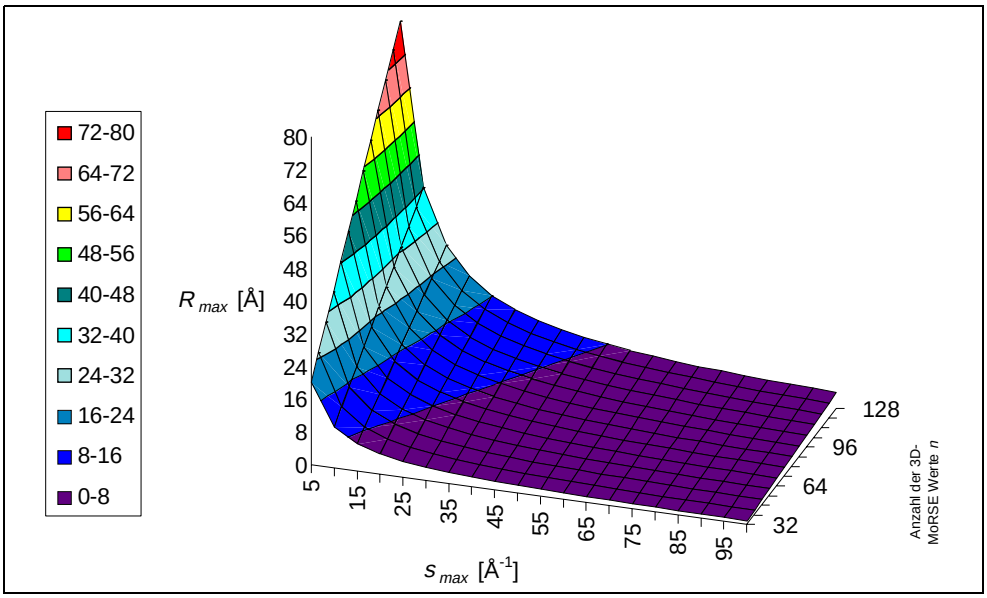


Abb. 2-25: s_{max} und n Abhängigkeit von R_{max}

Die Transformation führt zwangsläufig zu einem Informationsverlust. Durch die Wahl geeigneter Codierungsparameter soll versucht werden, den infrarotrelevanten Informationsgehalt des Codes so hoch wie möglich zu halten. Bei den nachfolgenden Simulationsexperimenten wurde untersucht, welche Parameterkombination zu den besten Simulationsergebnissen führt. Die Parameter wurden dabei wie folgt variiert:

Tab. 2-6: Variation der Codierungsparameter

Codierungsparameter	Variation
Atomeigenschaft A_i	α, q_σ, q_π, q_{tot}, m
s_{max}	5, 10, 15, ..., 100 Å^{-1}
Anzahl der 3D-MoRSE-Werte n	32, 40, 48, ..., 128

Neben den Codierungsparametern hat die Skalierung der 3D-MoRSE-Werte einen großen Einfluß auf die Simulationsqualität. Es werden hier zwei Skalierungsmethoden verglichen:

- Minimum/Maximum-Skalierung anhand eines Datensatzes ausgewählter Moleküle:
  Dazu werden für die Moleküle des Skalierungsdatensatzes (vgl. Anhang A.1) die 3D-MoRSE-Werte berechnet. Für die einzelnen Codewerte des Skalierungsdatensatzes werden Faktoren k_i bestimmt um die Werte in einen Bereich von -1 bis +1 zu skalieren. Diese Faktoren werden dann gemäß Gleichung 2-12 auf den zu codierenden Datensatz angewendet. Bei den nachfolgenden Experimenten wird diese Skalierungsmethode als **Min/Max-Methode** bezeichnet.

$$x_i^{skaliert} = k_i x_i^{unskaliert}$$

Min/Max-Skalierung (Gl. 2-12)

- Normierung des Betrags des Eingabevektors auf den Wert 1:
  Jeder Datenpunkt des zu codierenden Datensatzes wird gemäß Gleichung 2-13 auf den Betrag 1 normiert.

$$x_i^{skaliert} = \frac{x_i^{unskaliert}}{\sqrt{\sum_{j=1}^{n} x_j^2}}$$

Betragsskalierung (Gl. 2-13)

Bei den nachfolgenden Experimenten wird diese Skalierungsmethode als **Betragsskalierung** bezeichnet.

Um die Einflüsse der verschiedenen Codierungsparameter zu testen, wurden für Datensätzen von Cyclohexan-, Pyridin- und Naphthalinderivaten Spektrensimulationsexperimente durchgeführt. Die Datensätze wurde nach einem Verfahren, wie es in Kapitel 2.4.2.1 beschrieben wird in je einen Trainings und einen Testdatensatz aufgeteilt.

Tab. 2-7: Verwendete Datensätze

Datensatz	Trainingsdatensatz	Testdatensatz
Cyclohexane	66 Moleküle	78 Moleküle
Pyridine	69 Moleküle	80 Moleküle
Naphthaline	32 Moleküle	46 Moleküle

Diese Substanzgruppen wurden ausgewählt, um einen möglichst großen Bereich an struktureller Vielfalt abzudecken. Mit den Molekülen der Trainingsdatensätze wurden dann jeweils CPG-Netze trainiert und für die entsprechenden Moleküle der Testdatensätze die Spektrensimulationen durchgeführt. Die Simulationsparameter sind in Tabelle 2-8 aufgeführt.

Tab. 2-8: Simulationsparameter

Strukturcodierung	3D-MoRSE-Werte variierende Atomeigenschaft A_i variierende Anzahl der Codewerte n
Neuronen	10 x 10
Netzwerkform	planar
Training	unüberwacht

Bei den Experimenten wurden die Moleküle mit den verschiedenen Parameterkombinationen codiert, die IR-Spektren für die jeweiligen Testdatensätze simuliert und der mittlere Korrelationskoeffizient zwischen simulierten und experimentellen Spektren berechnet. Die Ergebnisse sind im nachfolgenden als s_{max} vs n Graphiken dargestellt. Die Farbe beschreibt die Höhe des bereichsgewichteten Korrelationskoeffizienten r_b, der für die jeweilige Parameterkombination ermittelt wurde. Die purpurfarbenen Kreise markieren die Maximalwerte für r_b, die im nachfolgenden als $r_{b,max}$ bezeichnet werden.

Simulationsexperimente für Cyclohexanderivate

Bei den Cyclohexanderivaten wurden keine Versuche mit $A_i = q_\pi$ durchgeführt, da dies bei Molekülen mit rein aliphatischen Substituenten, Nullvektoren als Code ergibt.

Ergebnisse der Simulationen mit Min/Max-Skalierung:

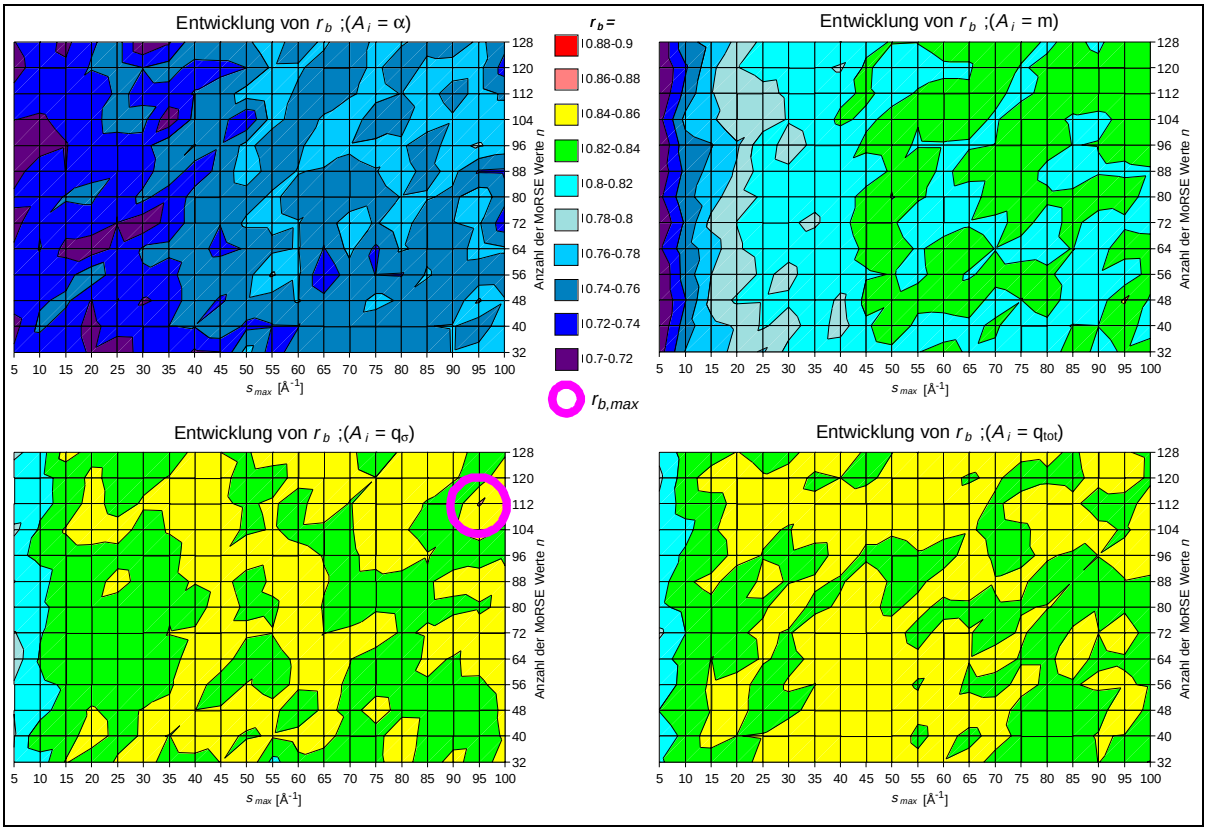


Abb. 2-26: Simulationsergebnisse mit der Min/Max-Skalierung (Cyclohexanderivate). Die Abbildung zeigt eine Auftragung von s_{max} gegen die Anzahl der 3D-MoRSE-Werte *n*. Die Farben beschreiben die erzielten Simulationsqualitäten anhand des Korrelationskoeffizienten *r* zwischen simulierten und experimentellem Spektren. Der purpurfarbene Kreis markiert den Höchstwert $r_{b,max}$.

Diskussion der Ergebnisse für die Min/Max Skalierung der Cyclohexandatensätze:

Allgemein nimmt die Güte der Simulationen von A_i = α über $A_i = m$ nach $A_i = q_\sigma$ sowie q_{tot} zu. Bei allen Experimenten ist weiterhin zu beobachten, daß die Ergebnisse für zu kleine s_{max}-Werte abnehmen. Die Grenze für s_{max}, ab welcher schlechtere Simulationsergebnisse zu beobachten sind, verläuft bei allen Experimenten nahezu parallel zur Abszisse und liegt für A_i = α bei s_{max} = 50 Å^{-1}, für $A_i = m$ bei s_{max} = 20 Å^{-1}, bei $A_i = q_\sigma$ und $A_i = q_{tot}$ bei s_{max} = 10 Å^{-1}. Dieses Verhalten scheint daraufhinzudeuten, daß bei dieser Form der Skalierung die Ergebnisse zu einem gewissen Grad unabhängig von der Anzahl der Strukturcodewerte sind.

Ergebnisse der Simulationen mit der Betragsskalierung:

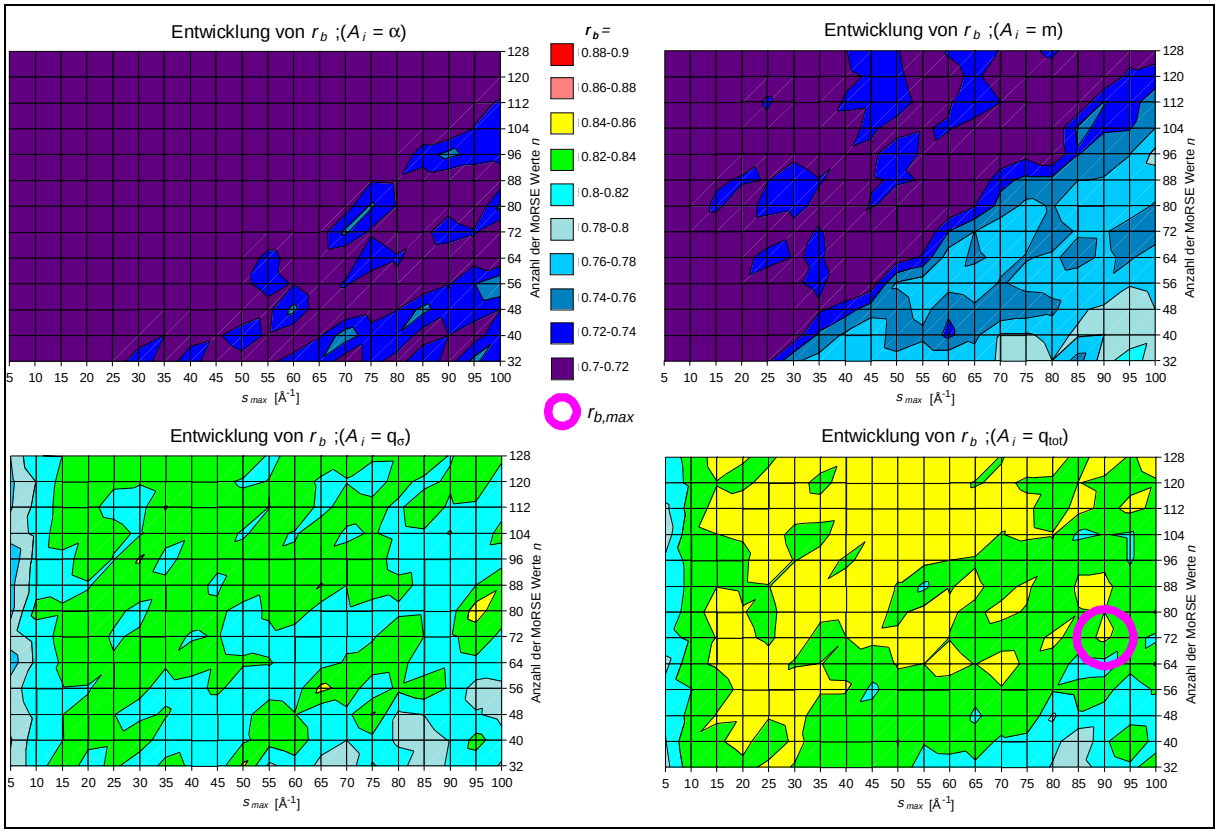


Abb. 2-27: Simulationsergebnisse mit der Betragsskalierung (Cyclohexanderivate). Die Abbildung zeigt eine Auftragung von s_{max} gegen die Anzahl der 3D-MoRSE-Werte *n*. Die Farben beschreiben die erzielten Simulationsqualitäten anhand des Korrelationskoeffizienten *r* zwischen simulierten und experimentellem Spektren. Der purpurfarbene Kreis markiert den Höchstwert $r_{b,max}$.

Diskussion der Ergebnisse für die Betragsskalierung der Cyclohexandatensätze:

Bei diesem Experiment nimmt die Güte der Simulationen von A_i = α über A_i = *m* nach A_i = q_σ sowie q_{tot} ebenfalls zu. Bei den Experimenten mit A_i = q_σ und A_i = q_{tot} ist auch wieder eine Grenze für s_{max} bei 10 - 15 Å^{-1} zu beobachten: Für s_{max} < 15 Å^{-1}werden die Ergebnisse wieder deutlich schlechter. Für diese beiden Atomeigenschaften fällt weiterhin auf, daß die Korrelationskoeffizienten für hohe s_{max} Werte und kleine Anzahlen von 3D-MoRSE-Werten abnehmen. Im gleichen Bereich werden die Simulationsergebnisse für A_i = α und A_i = *m* zwar zunehmend besser, bleiben jedoch deutlich unter denen für A_i = q_σ und A_i = q_{tot}. Bei allen vier Atomeigenschaften ist zu beobachten, daß die Isolinien diagonal durch die Abbildungen verlaufen. Um also in Bereichen gleicher Simulationsgüte zu bleiben, muß bei einer Erhöhung von s_{max} auch die Zahl n der 3D-MoRSE-Werte erhöht werden und umgekehrt. Es zeigt sich also eine wesentlich größere Abhängigkeit dieser beiden Codierungsparameter voneinander als

bei den Experimenten mit den Min/Max-skalierten Datensätzen.

Zusammenfassung der beiden Experimente:

In nachfolgender Tabelle sind die gemittelten r_b-Werte für die beiden Skalierungsmethoden mit den verschiedenen Atomeigenschaften aufgelistet.

Tab. 2-9: Cyclohexanderivate

Atom-eigenschaft	Min/Max-Skalierung		Betragsskalierung	
	$r_{b,\,max}$	$\bar{r}_b$	$r_{b,\,max}$	$\bar{r}_b$
α	0.782	0.746	0.760	0.684
m	0.842	0.805	0.815	0.727
q_σ	0.861	0.836	0.851	0.817
q_{tot}	0.860	0.839	0.856	0.833

Es zeigt sich, daß die r_b-Werte für $A_i = q_\sigma$ und $A_i = q_{tot}$ am höchsten liegen. Die Resultate von q_σ und q_{tot} unterscheiden sich kaum. Dies ist mit hoher Wahrscheinlichkeit dadurch zu erklären, daß der Cyclohexan-Grundkörper gesättigt ist und somit nur eventuell vorhandene π-Bindungen im Substituenten den Wert der Gesamtladung q_{tot} beeinflussen. Die Werte liegen für die Min/Max-Skalierung über denen der Betragsskalierung. Der Maximalwert für den r_b-Wert ist bei der Min/Max Skalierung mit $s_{max} = 95$, $n = 112$ und $A_i = q_\sigma$ zu finden. Der Wert liegt jedoch nur um 0.001 höher als für q_{tot}. Betrachtet man jedoch die gesamten Experimente, so sind die Ergebnisse für $A_i = q_{tot}$ meistens höher.

Bei beiden Skalierungen werden die Simulationsergebnisse für $s_{max} < 25$ deutlich schlechter. Die r_b-Wert der Experimente mit der Min/Max-Skalierung zeigen sich unabhängig von der Anzahl n der 3D-MoRSE-Codewerte. Im Gegensatz dazu ist es bei der Betragsskalierung notwendig, bei steigendem n auch s_{max} zu erhöhen, um im Bereich von $r_b > 0.84$ zu bleiben.

Simulationsexperimente für Pyridinderivate

Simulationsexperimente mit der Min/Max-Skalierung

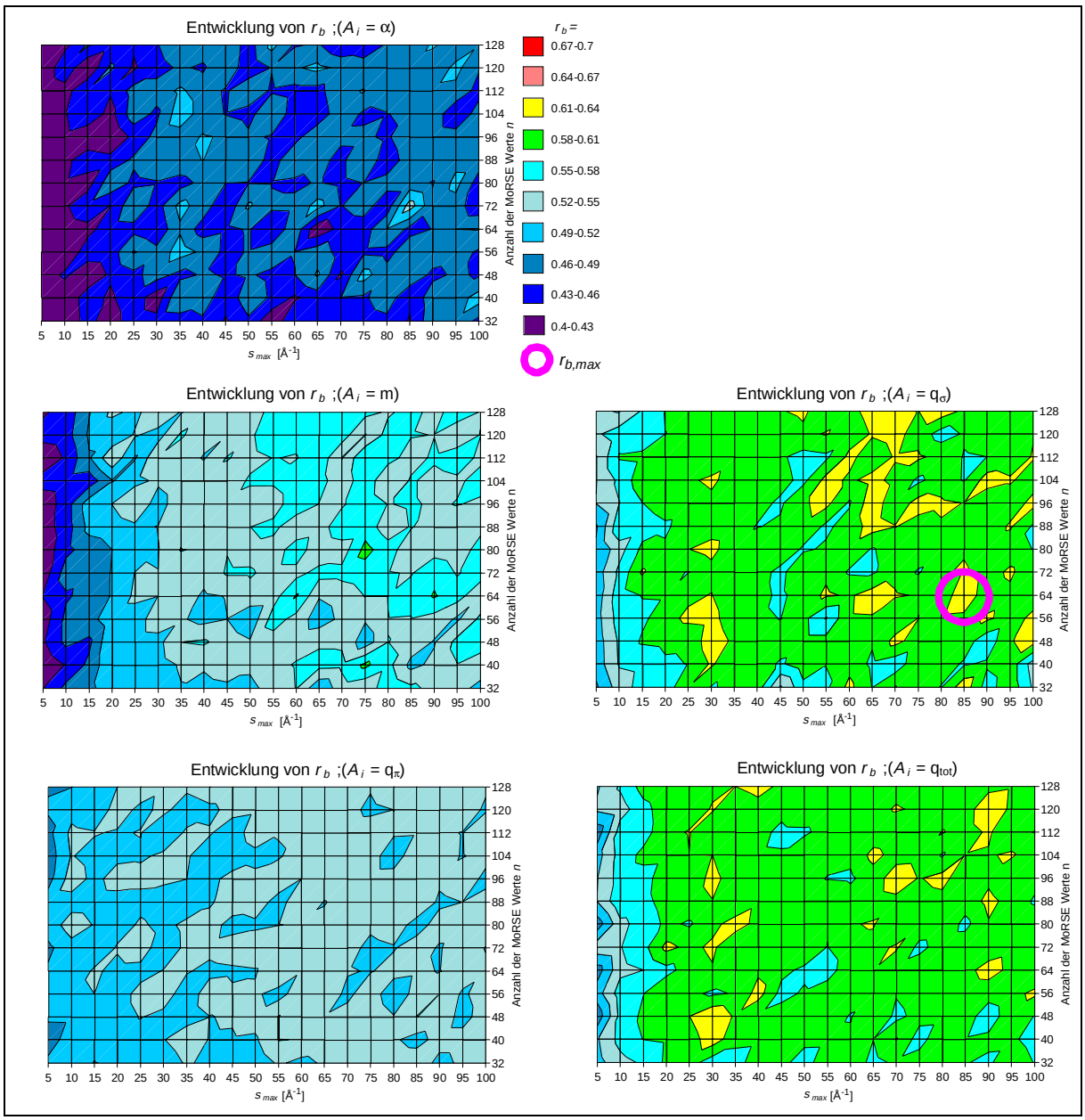


Abb. 2-28: Simulationsergebnisse mit der Min/Max-Skalierung (Pyridinderivate). Die Abbildung zeigt eine Auftragung von s_{max} gegen die Anzahl der 3D-MoRSE-Werte *n*. Die Farben beschreiben die erzielten Simulationsqualitäten anhand des Korrelationskoeffizienten *r* zwischen simulierten und experimentellem Spektren. Der purpurfarbene Kreis markiert den Höchstwert $r_{b,max}$.

Diskussion der Ergebnisse:

Bei den Simulationsexperimenten mit den Min/Max-skalierten Pyridindatensätzen werden mit den Codierungsparametern $A_i = q_\sigma$ und $A_i = q_{tot}$ die besten Ergebnisse erzielt. Bei allen fünf Experimenten ist zu beobachten, daß die Ergebnisse für Werte von s_{max} < 15 - 20 Å^{-1}

schlechter werden. Ein derartiger Grenzwert für die Anzahl der 3D-MoRSE-Codewerte ist dagegen nicht zu beobachten.

Simulationsexperimente mit der Betragsskalierung

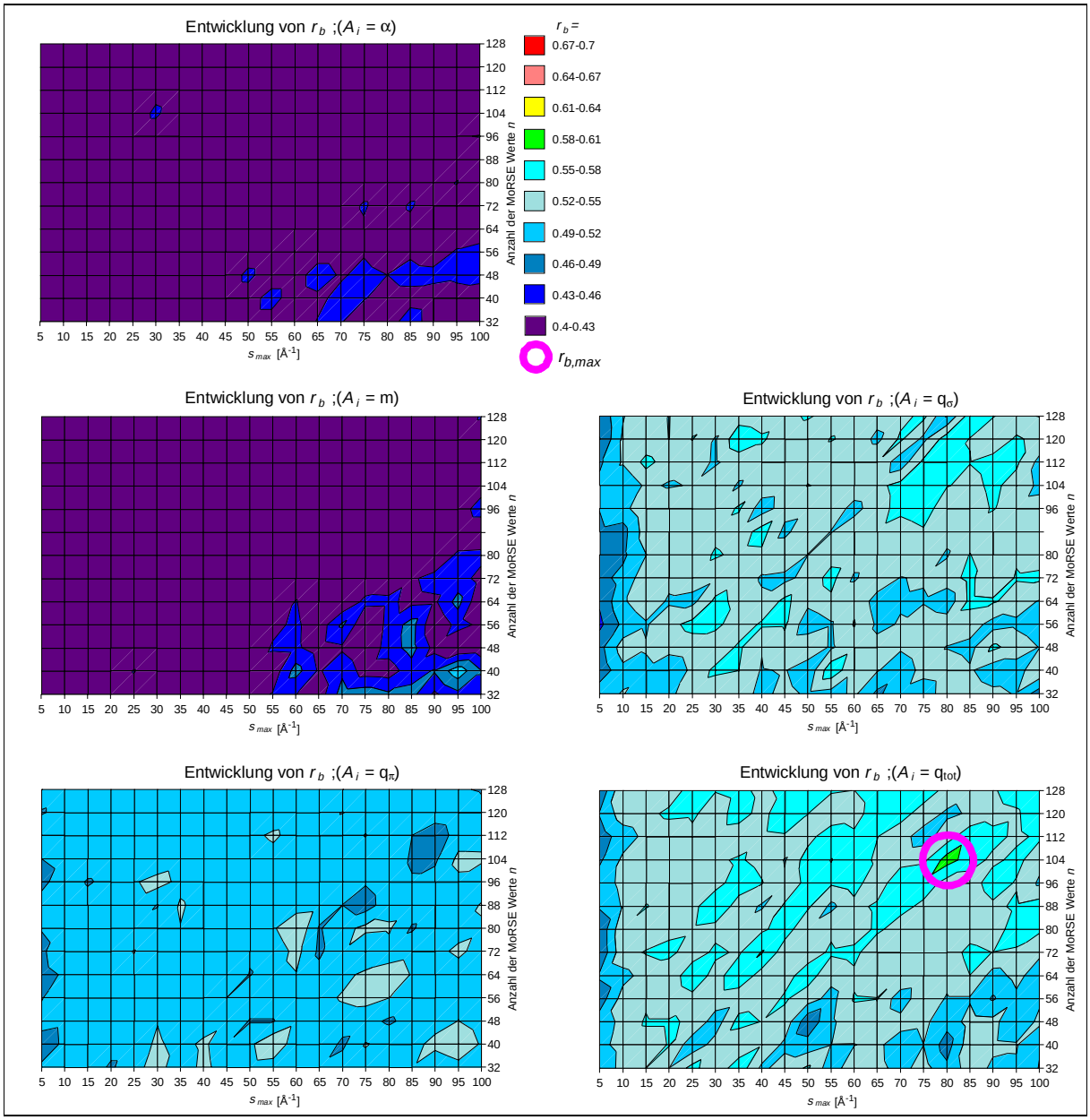


Abb. 2-29: Simulationsergebnisse mit der Betragsskalierung (Pyridinderivate). Die Abbildung zeigt eine Auftragung von s_{max} gegen die Anzahl der 3D-MoRSE-Werte n. Die Farben beschreiben die erzielten Simulationsqualitäten anhand des Korrelationskoeffizienten r zwischen simulierten und experimentellem Spektren. Der purpurfarbene Kreis markiert den Höchstwert $r_{b,max}$.

Diskussion der Ergebnisse mit der Betragsskalierung:

Allgemein fallen die Ergebnisse bei der Betragsskalierung schlechter aus als bei der

Min/Max-Skalierung, wobei wiederum für $A_i = q_\sigma$ und $A_i = q_{tot}$ die besten Ergebnisse erzielt werden. Wie bei den Simulationsexperimenten mit dem Cyclohexandatensatz ist zu beobachten, daß Bereiche gleicher Simulationsqualität diagonal verlaufen. Wird der Wert für s_{max} erhöht, so muß gleichzeitig die Anzahl n der 3D-MoRSE-Werte erhöht werden, damit sich der Wert für den Korrelationskoeffizienten r nicht ändert.

Zusammenfassung der beiden Experimente:

In nachfolgender Tabelle sind die gemittelten r_b-Werte für die beiden Skalierungsmethoden mit den verschiedenen Atomeigenschaften aufgelistet. Tendenziell liegen die r_b-Werte für die Simulationen mit den Pyridinderivaten um etwa 0.2 niedriger als für die Cyclohexanderivate. Dies liegt daran, daß sich die IR-Spektren des Cyclohexandatensatzes mit einem mittleren Korrelationskoeffizienten $\bar{r}_b$ von 0.52 untereinander ähnlicher sind als die IR-Spektren des Pyridindatensatzes ($\bar{r}_b = 0.18$).

Tab. 2-10: Pyridinderivate

Atom-eigenschaft	Min/Max-Skalierung		Betragsskalierung	
	$r_{b,\,max}$	$\bar{r}_b$	$r_{b,\,max}$	$\bar{r}_b$
α	0.535	0.456	0.460	0.392
m	0.590	0.525	0.509	0.397
q_σ	0.636	0.588	0.613	0.529
q_π	0.551	0.552	0.550	0.505
q_{tot}	0.635	0.587	0.594	0.536

Wie bei den Cyclohexanderivaten ergeben die Versuche mit $A_i = q_\sigma$ oder q_{tot} die höchsten r_b-Werte. Ganz analog ist auch zu beobachten, daß Werte für $s_{max} < 25$ zu deutlich schlechteren Simulationsergebnissen führen. Der maximale r_b-Wert findet sich bei der Min/Max-Skalierung mit $s_{max} = 85$, $n = 64$ und $A_i = q_\sigma$. Analog den Versuchen mit den Cyclohexanderivaten ist der Wert nur um 0.001 höher als für q_{tot}.

Simulationsexperimente für Naphthalinderivate

Simulationen für den Min/Max-skalierten Naphthalindatensatz

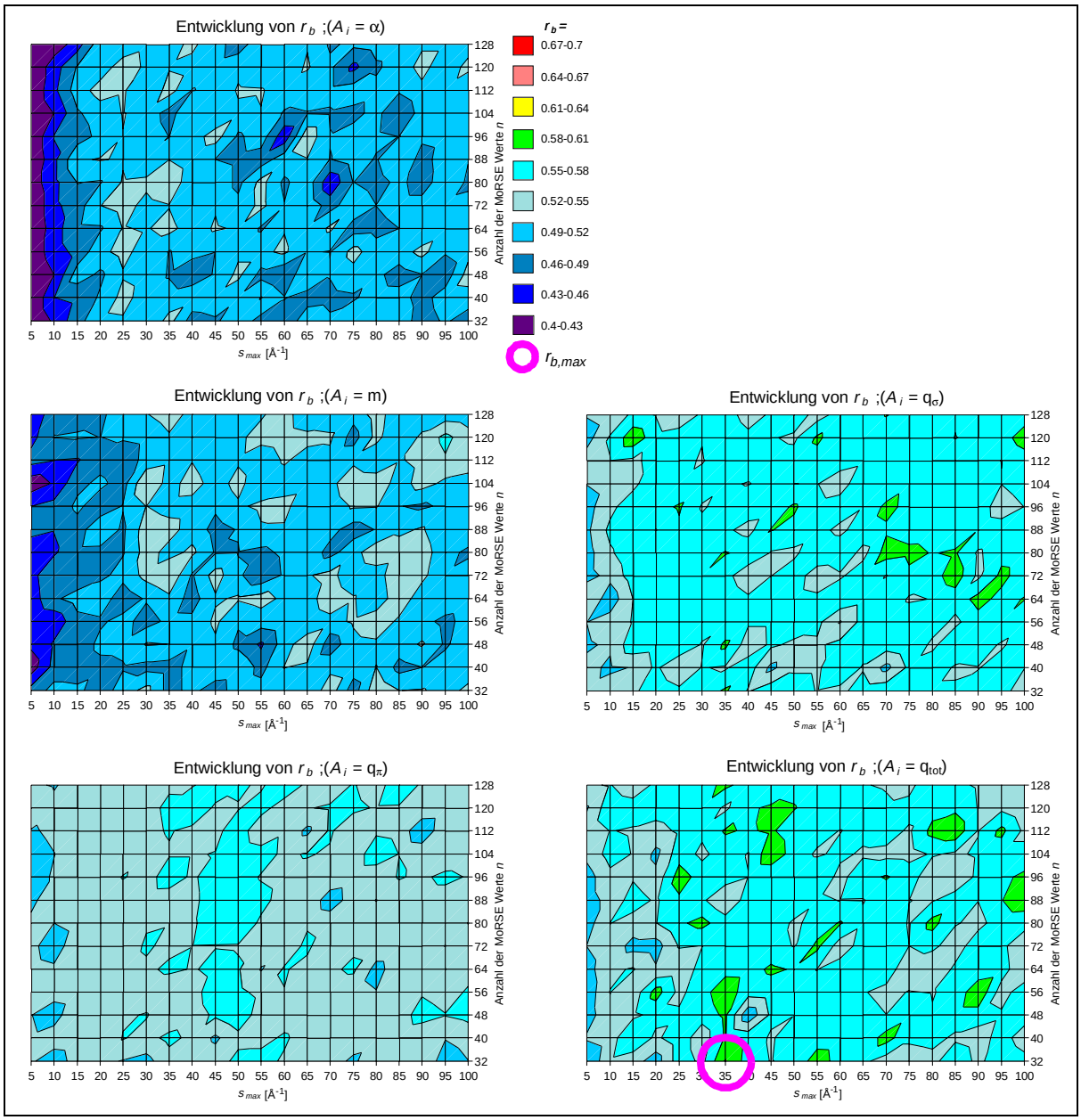


Abb. 2-30: Simulationsergebnisse mit der Min/Max-Skalierung (Naphthalinderivate). Die Abbildung zeigt eine Auftragung von s_{max} gegen die Anzahl der 3D-MoRSE-Werte n. Die Farben beschreiben die erzielten Simulationsqualitäten anhand des Korrelationskoeffizienten r zwischen simulierten und experimentellem Spektren. Der purpurfarbene Kreis markiert den Höchstwert $r_{b,max}$.

Diskussion der Simulationsergebnisse mit der Min/Max-Skalierung:

Entsprechend den obigen Experimenten sind auch hier wieder die besten Simulationsergebnisse bei $A_i = q_\sigma$ und $A_i = q_{tot}$ zu beobachten. Ebenfalls fällt die Simulationsgüte für kleine

s_{max}-Werte. Ein Grenzwert, ab welchem die Ergebnisse deutlich schlechter werden ist am deutlichsten bei den Versuchen mit $A_i = \alpha$ und $A_i = m$ zu beobachten und liegt bei etwa 15 Å^{-1}.

Betragsskalierung auf 1

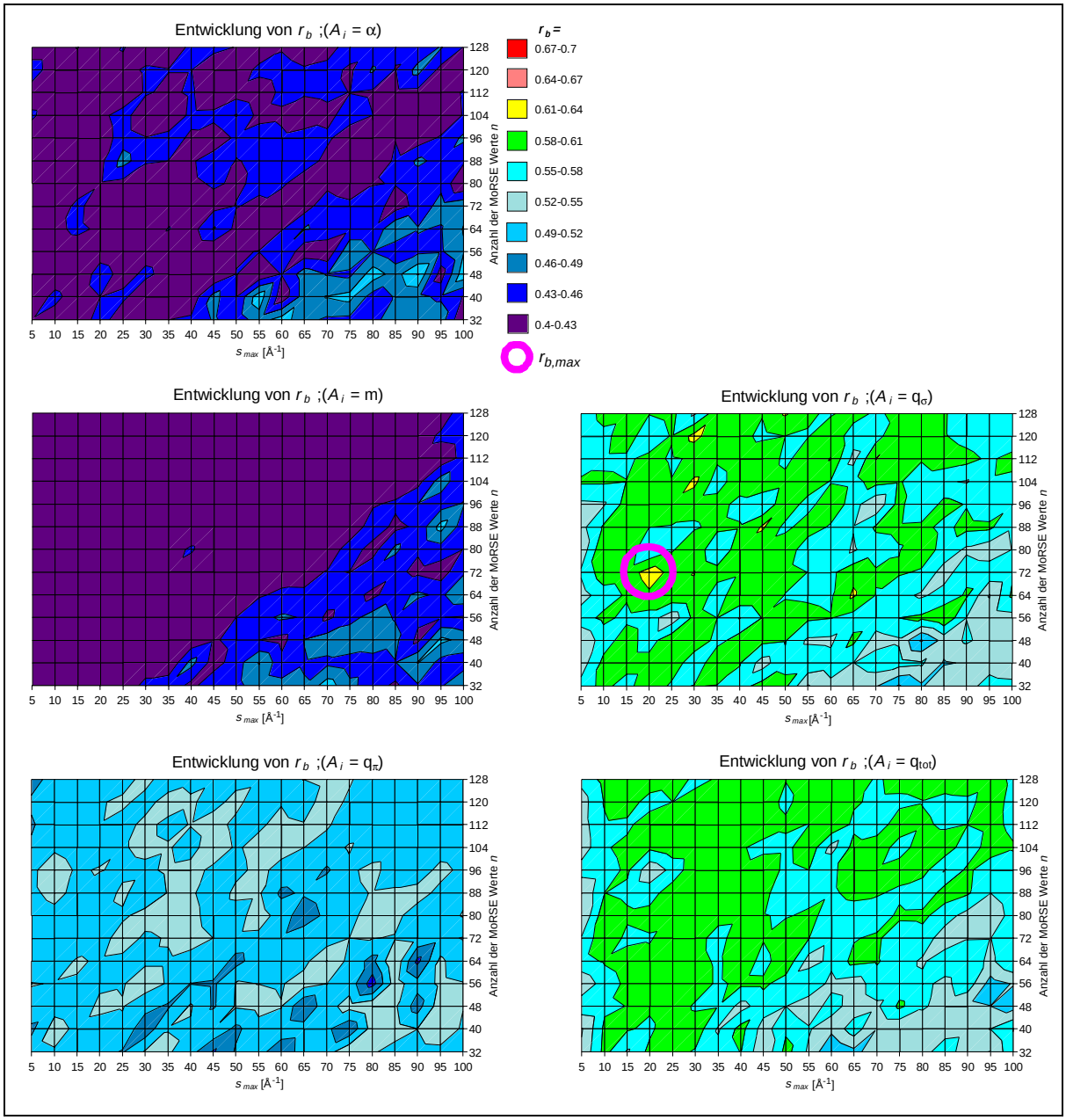


Abb. 2-31: Simulationsergebnisse mit der Betragsskalierung (Naphthalinderivate). Die Abbildung zeigt eine Auftragung von s_{max} gegen die Anzahl der 3D-MoRSE-Werte n. Die Farben beschreiben die erzielten Simulationsqualitäten anhand des Korrelationskoeffizienten r zwischen simulierten und experimentellem Spektren. Der purpurfarbene Kreis markiert den Höchstwert $r_{b,max}$.

Diskussion der Simulationsergebnisse mit der Betragsskalierung:

Anders als bei der Min/Max-Skalierung fallen hier die Ergebnisse für $A_i = q_\pi$ wieder schlechter aus als für $A_i = q_\sigma$ und $A_i = q_{tot}$. Wie bei den betragsskalierten Cyclohexan- und Pyridindatensätzen ist auch hier die Abhängigkeit von s_{max} und der Anzahl n der 3D-MoRSE-Werte zu erkennen.

Zusammenfassung der Ergebnisse für die Naphthalin-Simulationen

In nachfolgender Tabelle sind die gemittelten r_b-Werte für die beiden Skalierungsmethoden mit den verschiedenen Atomeigenschaften aufgelistet.

Tab. 2-11: Naphthalinderivate

Atom-eigenschaft	Min/Max-Skalierung		Betragsskalierung	
	$r_{b,\,max}$	$\bar{r}_b$	$r_{b,\,max}$	$\bar{r}_b$
α	0.549	0.493	0.522	0.429
m	0.558	0.501	0.506	0.408
q_σ	0.601	0.558	0623	0.574
q_π	0.573	0.540	0.551	0.512
q_{tot}	0.612	0.555	0.608	0.572

Bei dieser Versuchsreihe liegen die Werte für die Betragsskalierung über denen der Min/Max Skalierung. Der Maximalwert für den r_b-Wert ist bei der Betragsskalierung mit $s_{max} = 60$, $n = 120$ und $A_i = q_\sigma$ zu finden. Auch hier werden bei beiden Skalierungen die Simulationsergebnisse für $s_{max} < 25$ deutlich schlechter. Ebenso ist es bei der Betragsskalierung notwendig, bei steigendem n auch s_{max} zu erhöhen, um im Bereich von $r_b > 0.58$ zu bleiben.

Zusammenfassung aller Simulationsversuche

Bei allen Versuchen ist zu beobachten, daß sich für $s_{max} < 25$ Å^{-1} die Simulationsergebnisse deutlich verschlechterten. Für $s_{max} \geq 25$ Å^{-1} zeigt sich bei der Min/Max-Skalierung keine Abhängigkeit der Werte des bereichsgewichteten Korrelationskoeffizienten r_b von der Anzahl der 3D-MoRSE-Werte *n*. Diese scheinbare Unabhängigkeit von *n* wird sicherlich durch den Startwert von *n* = 32 verursacht, der eine brauchbare Auflösung des Codes und damit eine genügende Beschreibung der Molekülstruktur gewährleistet.

Die Hyperfläche der r_b-Werte zeigt in einem großen Bereich für $s_{max} > 25$ zufällig gelegene lokale Maxima und Minima. Die Werte in diesem Bereich liegen jedoch alle über dem Niveau der r_b-Werte für $s_{max} < 25$, so daß bei einer Auswahl von Codierungsparametern aus diesem Bereich sinnvolle Simulationsergebnisse zu erwarten sind:

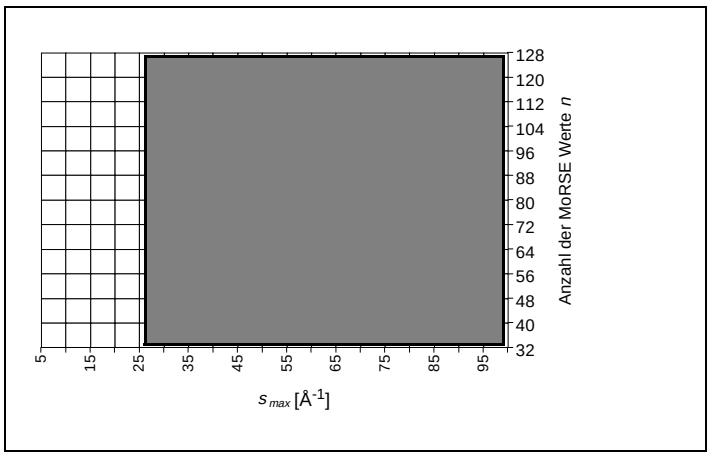


Abb. 2-32: Bereich mit guten Ergebnissen für die Min/Max-Skalierung

Der Grenzwert von 25 für s_{max} findet sich auch bei den Simulationsexperimenten mit der Betragsskalierung. Für $s_{max} \geq 25$ zeigt sich bei diesen Versuchen, daß mit zunehmendem s_{max} auch *n* erhöht werden muß, damit die r_b-Werte nicht wieder absinken. Der maximale Abstand zweier Atome im Molekül r_{max}, der noch durch den Code beschrieben werden kann, ist abhängig von s_{max} und *n* (vgl. Gl. 2-11). Die Grenze, die sich bei den Versuchen mit der Betragsskalierung andeutet, liegt auf der Gerade für $r_{max} = 3$ Å (gestrichelte Linie in Abb. 2-33).

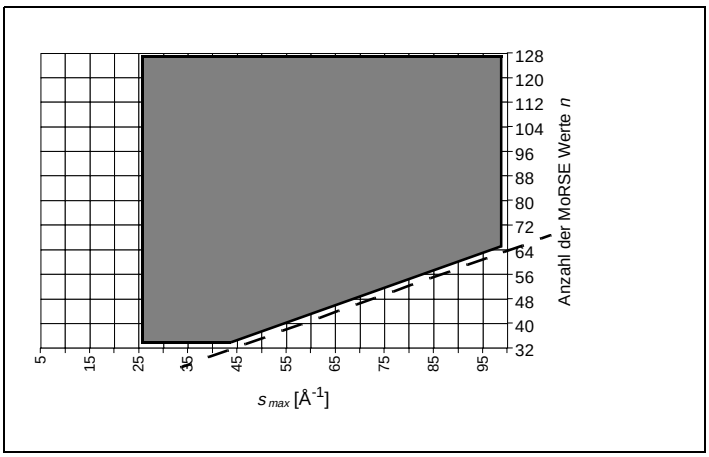


Abb. 2-33: Bereich mit guten Ergebnissen für die Betragsskalierung (graues Feld)

Eine größere Anzahl von Strukturvariablen n führt zu einer längeren Rechenzeit beim Netztraining, wobei der Zeitaufwand etwa quadratisch mit der Zahl der Strukturvariablen zunimmt. Aus diesem Grund sollte die Anzahl der Strukturvariablen so groß wie nötig und so niedrig wie möglich sein. Die Auswahl der Codierungsparameter s_{max} = 30 Å^{-1} und n = 64 erscheint deshalb für sinnvoll.

Da die Methode der Min/Max-Skalierung bei zwei von drei Versuchsreihen bessere Ergebnisse liefert, soll diese auch bei der 3D-MoRSE-Codierung angewendet werden. Bei der Variation der Atomeigenschaft A_i ergaben sich für q_σ und q_{tot} die besten Ergebnisse. Dabei war die Ergebnisse bei der Min/Max-Skalierung für q_{tot} bei den Cyclohexan- und Pyridinsimulationen geringfügig schlechter (Δr_b = 0.001), bei den Naphthalinsimulationen jedoch deutlich besser (Δr_b = 0.011). Daher werden bei den folgenden Simulationsexperimenten die Strukturcodierungen mit $A_i = q_{tot}$ durchgeführt.

2.4.1.2 Radial Code

Wie bei der 3D-MoRSE-Codierung, bietet sich auch bei der Radialcodierung (vgl. Gl. 2-5) die Möglichkeit, durch die Variation verschiedener Codierungsparameter, wie der Atomeigenschaft A_i, den maximal berücksichtigten Atomabstand R_{max} und den Unschärfeparameter B, den Informationsgehalt des Codes zu beeinflussen. Im Abschnitt zur 3D-MoRSE-Codierung konnte bereits gezeigt werden, daß die Gesamtladung q_{tot} eine geeignete Atomeigenschaft für den Codierungsparameter A_i ist. In dem folgenden Abschnitt soll nach geeigneten Werten für den maximal berücksichtigten Atomabstand R_{max} und den Unschärfeparameter B gesucht werden.

Bestimmung eines geeigneten R_{max}:

Bei der Berechnung des Radialcodes $g(R)$ läuft die Variable R von 0 bis R_{max}. R_{max} entspricht anschaulich dem maximalen Atomabstand innerhalb des Moleküls, der noch in die Codierung miteinfließt. Um einen sinnvollen Wert für R_{max} zu ermitteln, wurden folgende Versuche angestellt:

Es wurden für alle Benzole, Pyridine, Naphthaline, Chinoline und Isochinoline der Spec-Info IR-Datenbank die maximalen Atomabstände R_{max} bestimmt.

Tab. 2-12: Untersuchte Datensätze

Verbindungen	Anzahl der Moleküle im Datensatz
Benzolderivate	871
Pyridinderivate	149
Naphthalinderivate	78
Chinolin- und Isochinolinderivate	123

Die Häufigkeiten für das Auftreten der verschiedenen R_{max}-Werte ist in nachfolgender Graphik dargestellt:

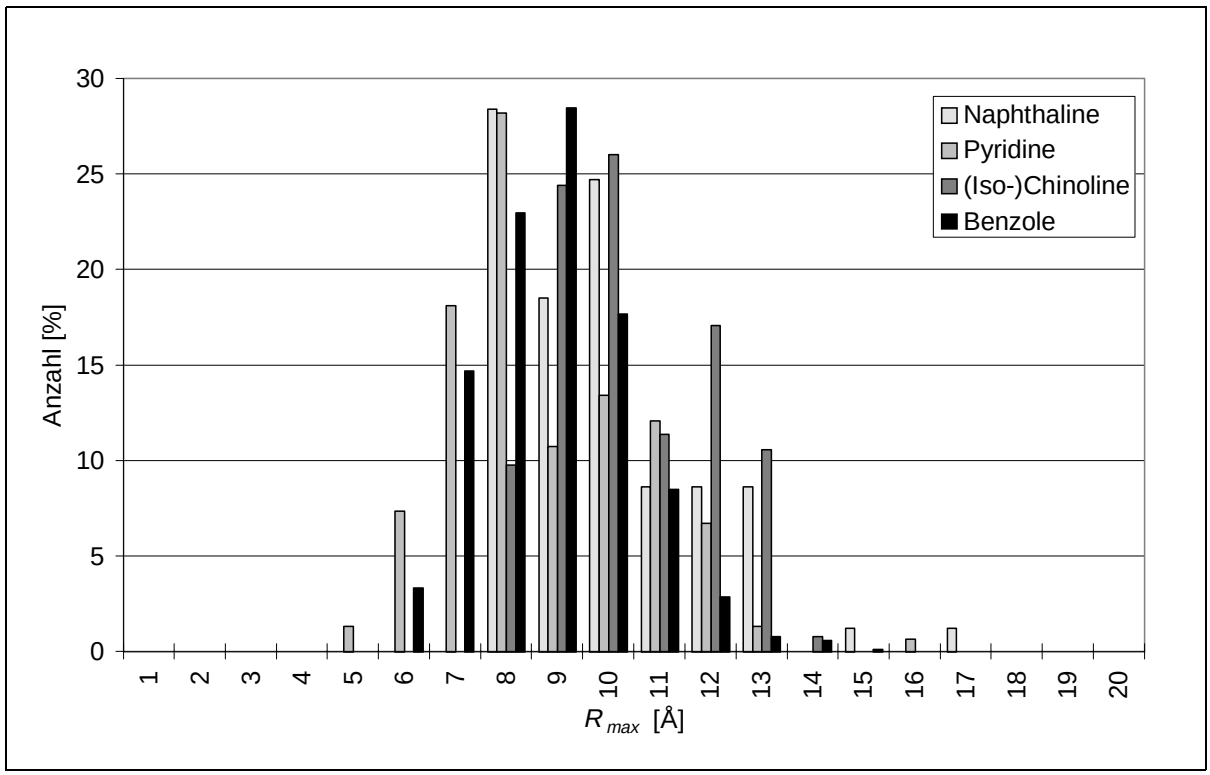


Abb. 2-34: Verteilung von R_{max} für Naphthaline, Pyridine, Chinoline und Isochinoline sowie Benzole der SpecInfo IR-Datenbank

Am häufigsten treten bei diesen Datensätzen R_{max}-Werte von 8-9 Å auf. Ebenso ist zu beobachten, daß die kleineren Abstände häufiger in den Benzol- und Pyridindatensätzen auftreten, während die größeren Atomabstände häufiger bei den Naphthalinen und (Iso)-Chinolinen zu finden sind. Dies ist in Anbetracht der jeweiligen Molekülgerüste nicht weiter verwunderlich, trotzdem hat die Form der verschiedenen Substituenten einen sehr großen Einfluß auf R_{max}, was durch die jeweiligen Maximalwerte bei 9 Å (Benzole) und 11 Å (Pyridine) verdeutlicht wird. Allgemein ist es jedoch schwierig, hier von einer Verteilung zu sprechen, da die Datensätze, mit Ausnahme des Benzoldatensatzes, relativ klein sind (vgl. Tab. 2-12).

Aus diesem Grund wurde diese Untersuchung für die gesamte SpecInfo IR-Datenbank mit 13373 Molekülen durchgeführt. Die Häufigkeiten für das Auftreten der verschiedenen R_{max}-Werte ist in nachfolgender Graphik dargestellt:

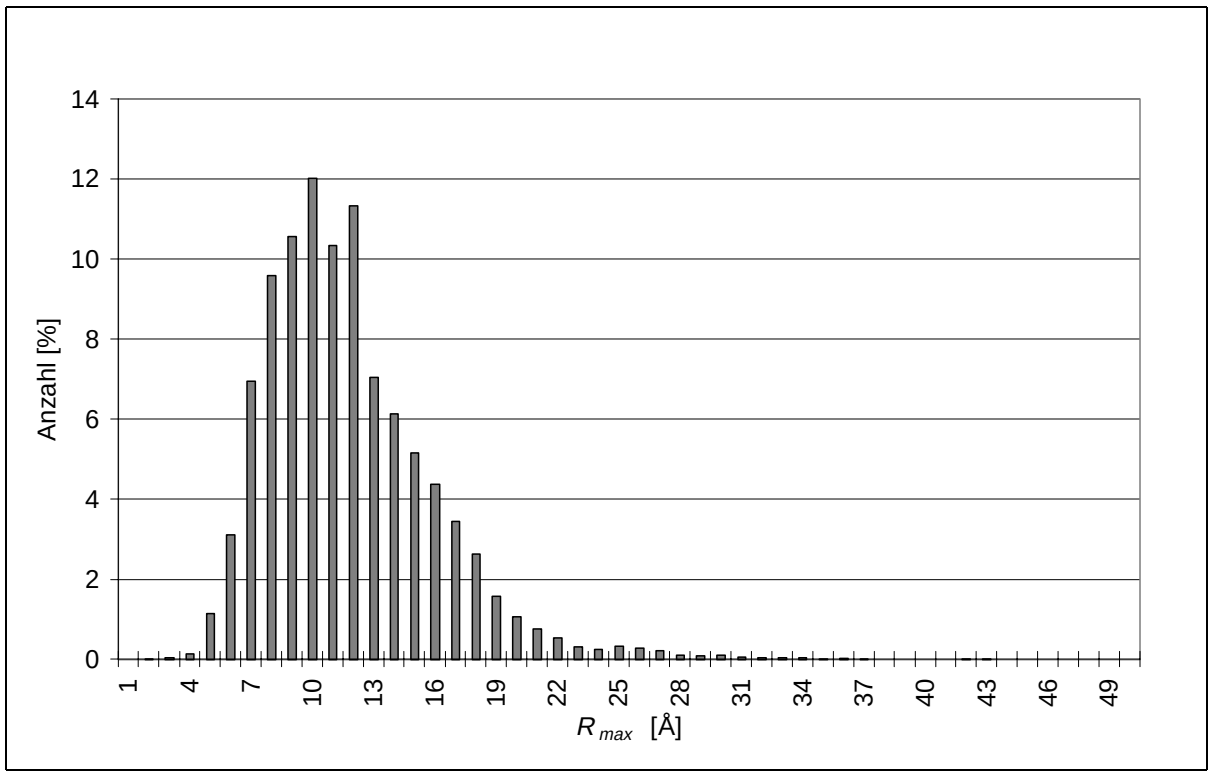


Abb. 2-35: Verteilung von r_{max} für alle Verbindungen der SpecInfo IR-Datenbank

Bei der Untersuchung der gesamten SpecInfo IR-Datenbank sind die Maximalwerte bei 10-12 Å zu beobachten. Prinzipiell sind bei der Ermittlung eines geeigneten R_{max}-Wertes zwei gegenläufige Aspekte zu berücksichtigen: Bei der Wahl eines zu niedrigen Wertes für R_{max} kann es vorkommen, daß ein Atomabstand der größer als R_{max} in die Codierung dieses Moleküls nicht mehr miteinfließt. Andererseits führt eine Erhöhung von R_{max} bei gleichbleibender Anzahl n der Codewerte zu einer schlechteren Auflösung des Strukturcodes. Da jedoch die Nachbarschaftsbeziehungen, also die kürzeren Atomabstände, das Infrarotspektrum am meisten beeinflussen, erscheint es in Anbetracht der Verteilung in Abbildung 2-35 sinnvoll, ein R_{max} von etwa 12 Å zu wählen. Beim Einsatz des Radialcodes zur Strukturaufklärung hatte sich gezeigt, daß eine Anzahl der Radialcodewerte von n = 128 günstig sind. Aus praktischen Erwägungen wurde deshalb ein R_{max}-Wert von 12.8 Å gewählt, um eine Schrittweite von 0.1 Å bei den einzelnen Codewerten zu erreichen.

Bestimmung eines geeigneten Unschärfeparameters *B*

Ein weiterer Parameter bei der Berechnung des Radialcodes (vgl. Gl. 2-5) ist der Unschärfe- oder auch Temperaturparameter *B*. Für einen bestimmtes Atompaar und festgelegten Experimentalbedingungen läßt sich umgekehrt *B* wie folgt berechnen:[55]

$$\frac{1}{B} = (\Delta R) R_{ij}^{\,2} \frac{T}{T_{melt}}$$

Berechnung des Unschärfeparameters *B* [55] (Gl. 2-14)

mit:

ΔR.....Signalauflösung
R_{ij} .....dem Signal entsprechende Spaltbreite
T .....Meßtemperatur
T_{melt}.....Schmelztemperatur der Substanz

Für ein C-H Atompaar (r = 1.09 Å) des Naphthalins (T_{melt} = 353 K, [56]) ergibt sich bei 273 K und einer Signalauflösung von ΔR = 0.1 Å ein *B*-Wert von 10.9 Å^{-2}. Für $B \rightarrow \infty$ (also: $T << T_{melt}$) wird die Radialbasiskurve zu einem Histogramm, wobei die Höhe der einzelnen Säulen proportional zu der Häufigkeit ist, mit welcher die einzelnen Atomabstände in dem Molekül auftreten. Umgekehrt nimmt bei höherer Temperatur die Atombewegung zu, wodurch wiederum der *B*-Wert sinkt und die Signale unschärfer werden. In nachfolgender Abbildung sind die Radialcodes für Ethylbenzol für verschiedene *B*-Parameter (10 , 100 und 1000 Å^{-2}) dargestellt:

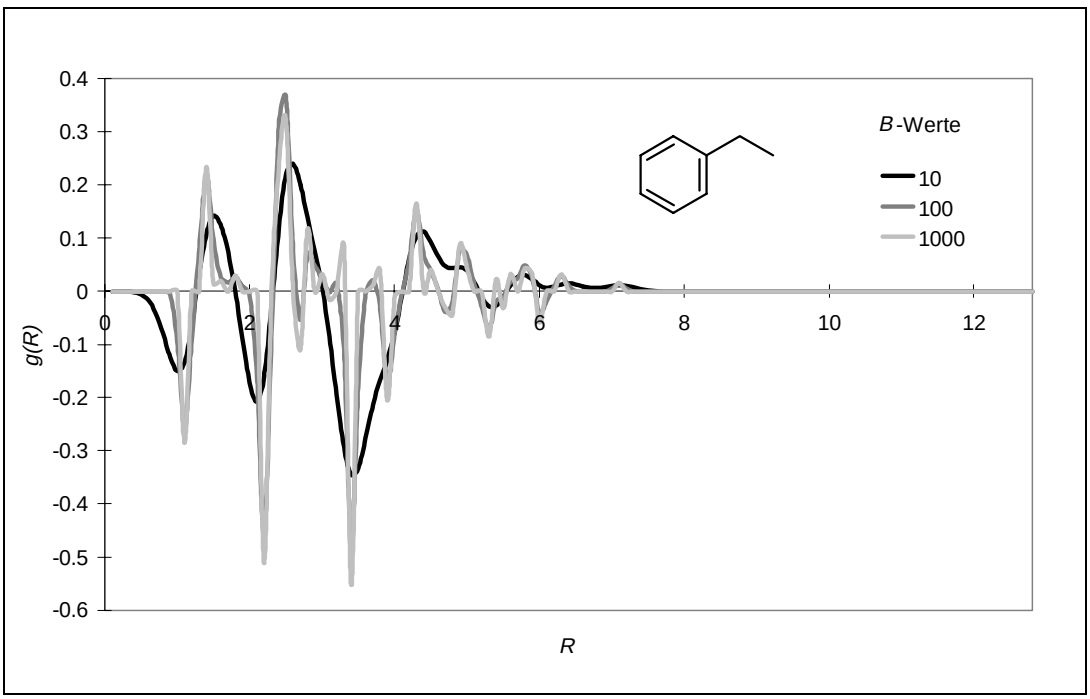


Abb. 2-36: Radialcodes von Ethylbenzol für verschiedene *B*-Parameter

Auch bei der Auswahl eines geeigneten *B*-Parameters gilt es wieder zwei gegenläufige Effekte zu berücksichtigen: Die Auflösung des Codes muß fein genug sein, um Strukturmerkmale wie eine aliphatische oder eine aromatische C-C Bindung unterscheiden zu können. Andererseits sollen die Signale des Codes eine gewisse Unschärfe aufweisen, damit ein Vergleich verschiedener Strukturcodes durch einfache Vergleichsmaße, wie z.B. dem *rms*-Wert möglich ist. Ein Vergleich zweier Histogramme mit dem *rms*-Wert ist nicht sinnvoll, da Codepeaks nur bei absolut identischer Lage als gleich erkannt werden. Dies hat zur Folge, daß Codes von Strukturen mit ähnlichen aber nicht identischen Strukturmerkmalen als zu unähnlich bewertet werden. Aus diesen Gründen erscheint die Auswahl von $B = 100$ Å^{-2} sinnvoll. Für $B = 100$ Å^{-2} weist die Kurve eine ausreichende Feinstruktur auf, ist jedoch noch ausreichend unscharf um einen Vergleich von Bandenmmustern mit dem *rms*-Wert zuzulassen.

2.4.2 Auswahl der Trainingsdatensätze

Es ist bereits wiederholt darauf hingewiesen worden, daß neuronale Netze den Zusammenhang von Struktur und Spektrum selbständig anhand von Beispielen lernen. Welche Daten zum Training verwendet werden, hat einen großen Einfluß auf die Güte der Vorhersage. Folgende Faktoren sind von Bedeutung:

- Reinheit der Probe
- Aufnahmebedingungen
- Phase bzw. Lösungsmittel
- Schichtdicke bzw. Konzentration der Substanz

Diese experimentellen Parameter sind beim Zugriff auf die Datenbank natürlich nicht mehr zu ändern. Prinzipiell wäre es sinnvoll, bei der Vorhersage eines gewünschten Referenzspektrums auch nur solche Datenbankspektren zum Training heranzuziehen, die unter den gleichen Bedingungen aufgenommen worden sind. Dies kann in der Regel aufgrund der mangelnden Verfügbarkeit einer ausreichenden Datenbasis jedoch nicht realisiert werden. Hier kann die Spektrensimulation jedoch als Qualitätstest eingesetzt werden, um bei einer deutlichen Abweichung zwischen simuliertem und experimentellem Spektrum auf einen fehlerhaften Datenbankeintrag hinzuweisen. Darüberhinaus besteht meistens keine Möglichkeit, um bei Spektren minderer Qualität auf andere Datenbanken auszuweichen. Unabhängig davon liefert die Information über die oben aufgeführten experimentellen Parameter möglicherweise einen wichtige Hinweise auf die Ursache für eine schlechte Simulation, wenn aufgrund des vorliegenden Datenmaterials eine bessere Vorhersage zu erwarten gewesen wäre. Letzteres leitet auf

den zweiten wichtigen Aspekt bei der Erstellung eines Trainingsdatensatzes über, nämlich der Auswahl von geeigneten Molekülen aus einer zugrundeliegenden Datenbasis. Die Auswahl der Trainingsmoleküle ist ein entscheidender Schritt der Methode, da diese Struktur-/Spektren-Paare die Wissensquelle für das neuronale Netz darstellen. Verschiedene Methoden zur Erstellung eines Trainingsdatensatzes werden in den nachfolgenden Kapiteln näher erläutert.

2.4.2.1 Globales Netz

Bei diesem Experiment wurde *ein* Netz für alle 9850 ungeladenen H, C, N, O, Hal -Verbindungen der SpecInfo-IR-Datenbank trainiert. Es wurde keine Auswahl im eigentlichen Sinn getroffen, der Datensatz wurde in drei möglichst gleichgroße repräsentative Datensätze unterteilt. Der erste Datensatz diente zum Netztraining. Anhand der Vorhersageergebnisse für den zweiten Datensatzes wurden geeignete Simulationsparameter, wie z.B. die Anzahl der Neuronen, ermittelt. Mit dem dritten Datensatz wurde der eigentliche Simulationstest durchgeführt. Die Aufteilung des Ursprungsdatensatzes wurde mittels eines neuronalen Kohonen-Netzes vorgenommen. Bei diesem Verfahren wird zunächst ein Netz mit dem gesamten, zu teilenden Datensatz trainiert. Von jedem belegten Neuron wird ein Molekül entnommen und in einem Unterdatensatz gespeichert:

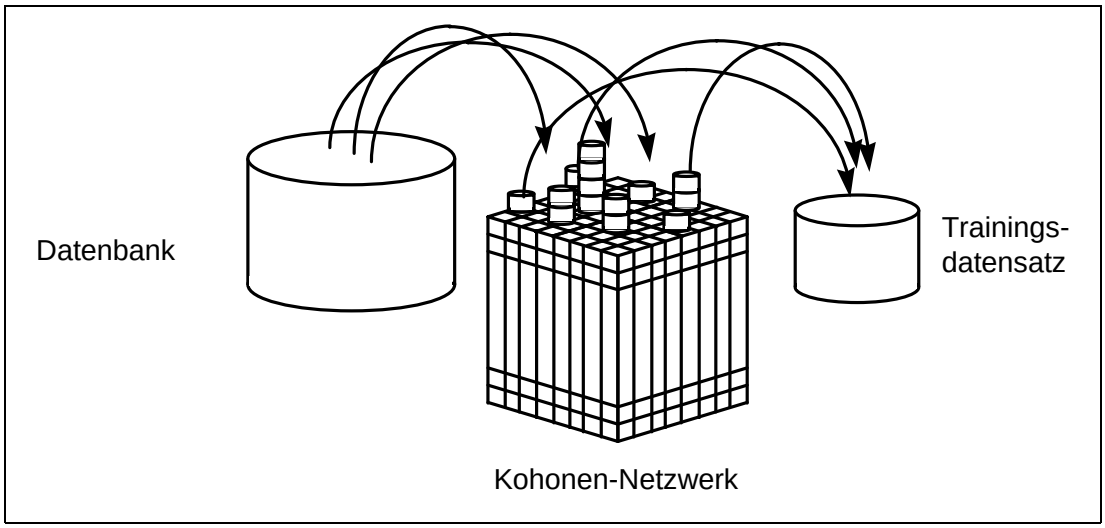


Abb. 2-37: Aufteilung eines Datensatzes mittels eines Kohonen-Netzes

Die verbleibenden Moleküle bilden einen anderen Unterdatensatz. Die Größe der Unterdatensätze läßt sich dabei durch die Anzahl der Neuronen des CPG-Netzes steuern. Je mehr Neuronen das erste Netz enthält, desto mehr Moleküle gelangen in den ersten Unterdatensatz. Da es in der Regel auch unbelegte Neuronen gibt, muß die Zahl der Neuronen etwas größer sein, als die gewünschte Anzahl der Moleküle im Unterdatensatz. Dieses Verfahren kann

mehrfach wiederholt werden, um so zu der gewünschten Anzahl von Teildatensätzen zu gelangen. In dem hier vorgestellten Versuch wurde das Verfahren einmal auf den gesamten Datensatz angewandt und einmal auf den Datensatz mit den Molekülen, die beim ersten Netztraining nicht ausgewählt worden sind. Auf diesem Wege wurde der Gesamtdatensatz, wie bereits erwähnt, in drei Teildatensätze aufgeteilt. Der erste Datensatz mit 3244 Molekülen wurde zum Training verschieden großer neuronaler Netze verwendet. Die Parameter zur Strukturcodierung und zum Netztraining waren wie folgt:

Tab. 2-13: Simulationsparameter

Strukturcodierung	64 3D-MoRSE-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	Aufteilung durch neuronales Netz
Anzahl der Trainingsmoleküle (Datensatz 1)	3244
Anzahl der Testmoleküle (Datensatz 2)	3150
Anzahl der Testmoleküle (Datensatz 3)	3456
Datenbasis	SpecInfo, (ungeladen, H, C, N, O, Hal)
Neuronen	40 x 40, 45 x 45, ..., 80 x 80
Netzwerkform	planar
Training	unüberwacht

Für die 3150 Moleküle des zweiten Datensatzes wurden mit den jeweiligen Netzen die Spektren vorhergesagt und die Korrelationskoeffizienten zwischen allen Paaren aus simulierten und experimentellen Spektren bestimmt. Für jedes Simulationsexperiment wurden 3150 Korrelationskoeffizienten berechnet, deren Mittelwerte $\bar{r}$ in nachfolgender Graphik aufgetragen sind:

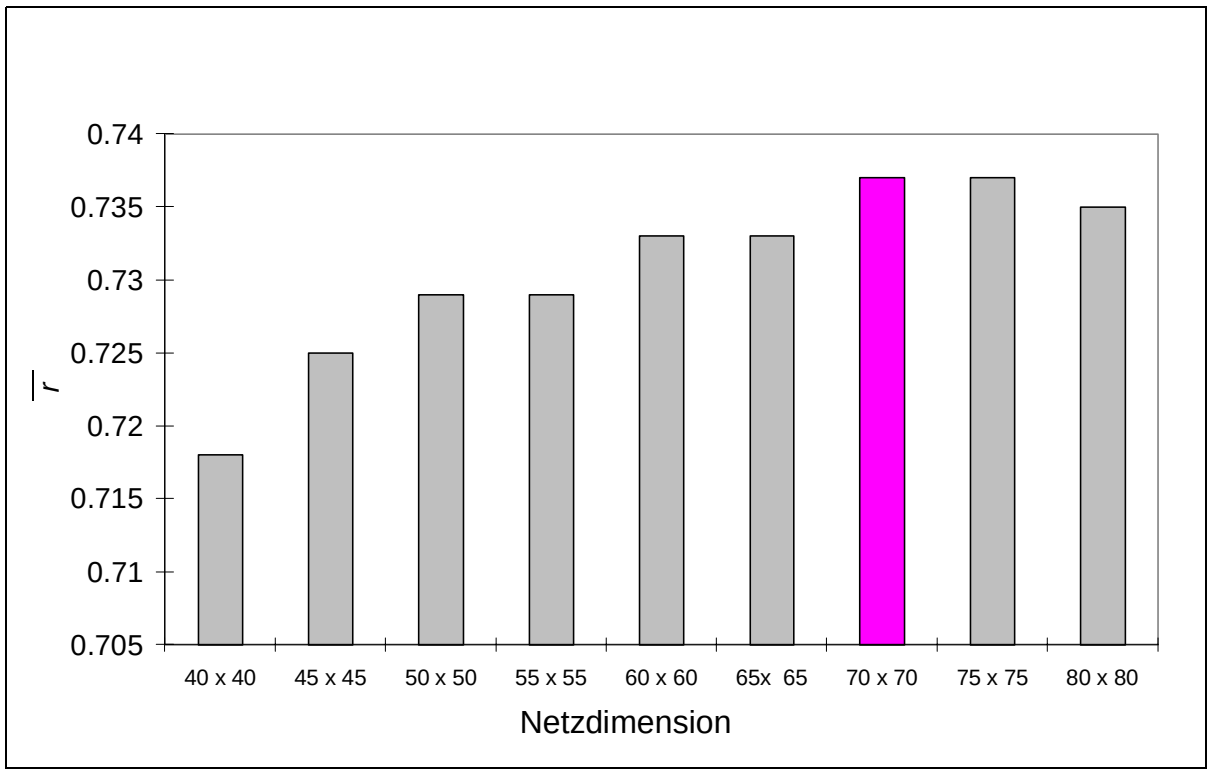


Abb. 2-38: Einfluß der Netzdimension auf die Simulationsgüte

Die Werte für 70 x 70 Neuronen sowie für 75 x 75 sind mit $\bar{r}$ = 0.737 gleich. Wegen der geringeren Rechenzeit wurde jedoch beim nachfolgenden Experiment das kleinere Netz mit 70 x 70 Neuronen gewählt. Mit diesem Netz wurde für die Moleküle von Datensatz 3 eine Simulation durchgeführt. Die Ergebnisse, der mittlere Korrelationskoeffizient und die Verteilung des Korrelationskoeffizienten über alle Simulationen, wurden mit den Ergebnissen der Simulationen für Testdatensatz 2 verglichen. Es wurde überprüft, ob die Ergebnisse vergleichbar sind und ob die Bestrebung zur Auswahl einer geeigneten Netzdimension zwar zu einer guten Vorhersage für den Testdatensatz 2 führen, jedoch bei einem anderen Datensatz versagen oder zumindest schlechtere Ergebnisse liefert. Der mittlere Korrelationskoeffizient für die Simulationen für Testdatensatz 3 fällt jedoch mit $\bar{r}$ = 0.741 sogar geringfügig höher aus, als für Testdatensatz 2. Die entsprechende Verteilung der Korrelationskoeffizienten für die Datensätze 2 und 3 ist in nachfolgender Grafik dargestellt:

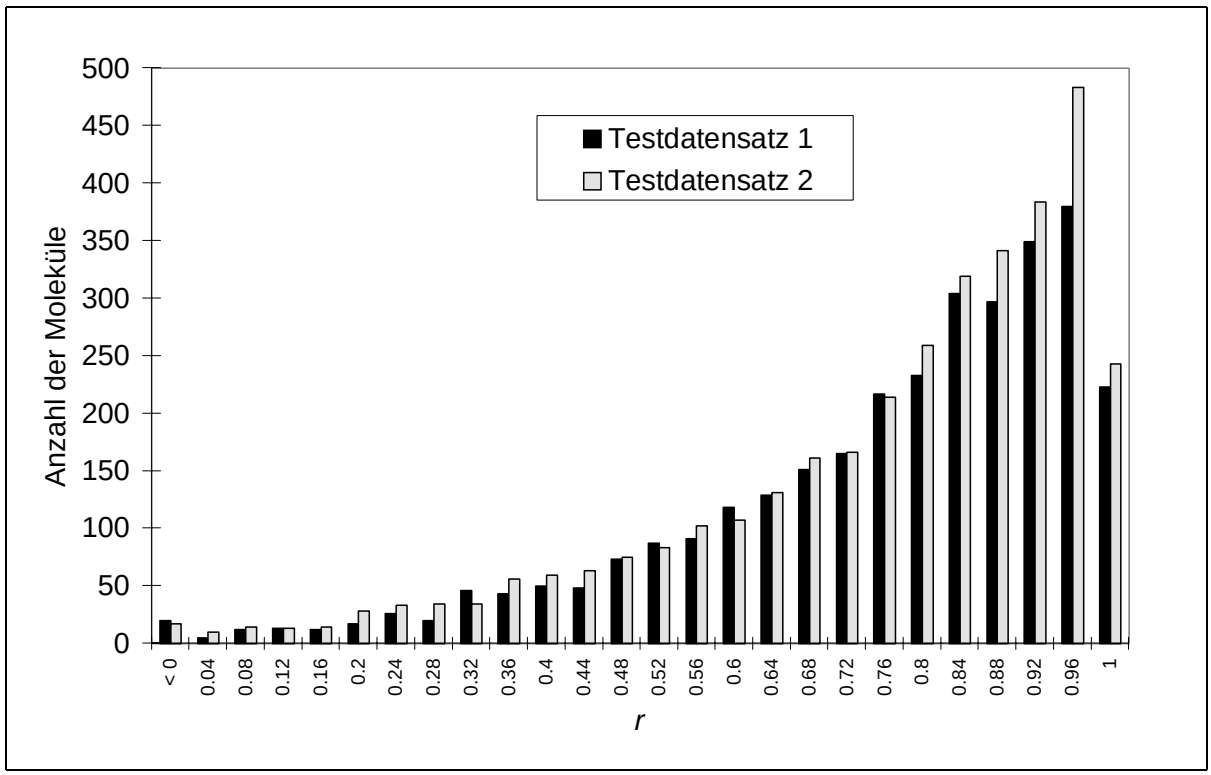


Abb. 2-39: Verteilung der Korrelationskoeffizienten *r* für Datensatz 2 (Testdatensatz1) und Datensatz 3 (Testdatensatz2)

Es ist deutlich zu erkennen, daß der Großteil der Simulationen von hoher Qualität ist. Die Verteilung der Simulationsqualitäten ist in Tabelle 2-14 aufgetragen.

Tab. 2-14: Prozentuale Verteilung der Simulatiosqualitäten

Korrelationskoeffizient *r*	Datensatz 2	Datensatz3
$0.8 \leq r \leq 1$	57%	59%
$0.6 \leq r < 0.8$	25%	23%
$0.4 \leq r < 0.6$	11%	11%

In nachfolgender Abbildung wird das Simulationsergebnis für ein Decalinderivat gezeigt, wobei trotz des komplizierten Molekülgerüsts ein gutes Simulationsergebnis erzielt werden konnte.

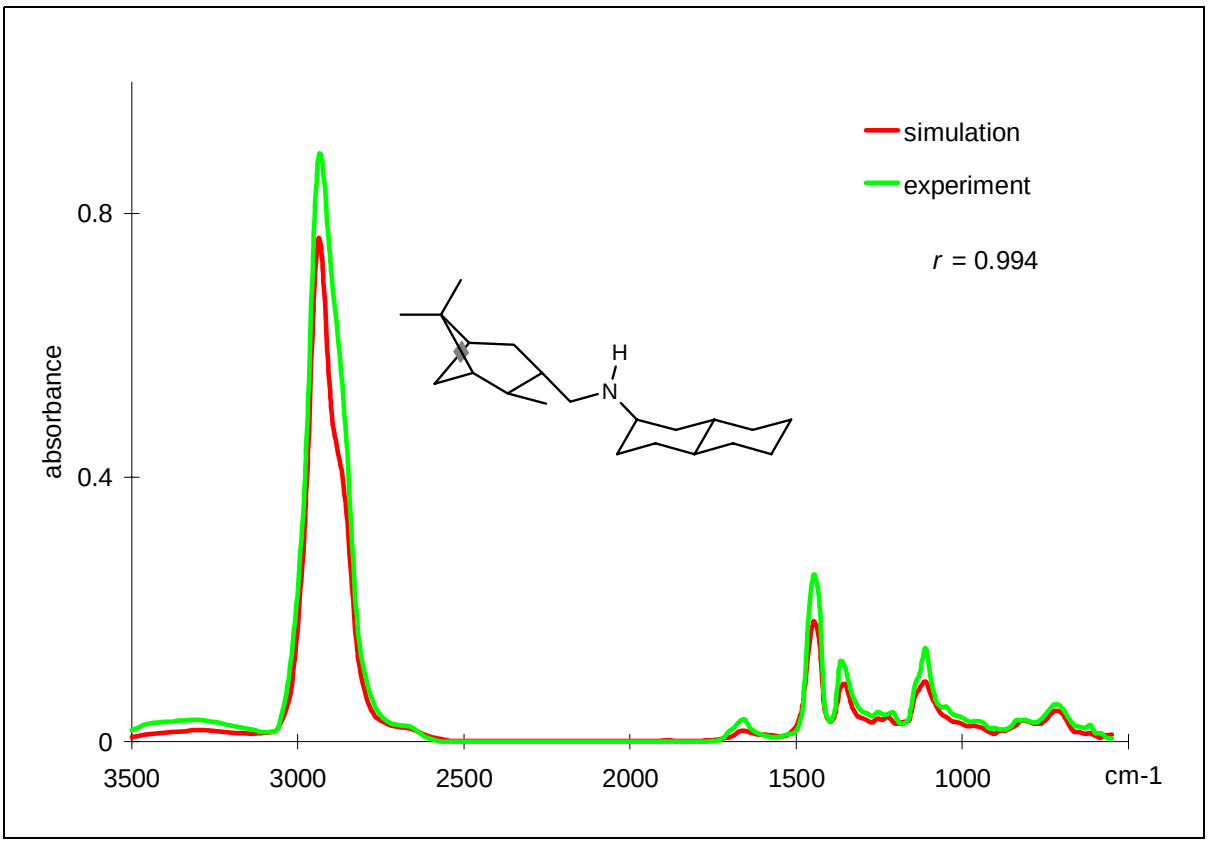


Abb. 2-40: Vergleich von simuliertem und experimentellem Spektrum eines Decalinderivates

Trotz des komplizierten Molekülgerüsts (Verknüpfung eines Decalin und eines Norpinan-Gerüsts) ist die Übereinstimmung von simuliertem und experimentellem Spektrum sehr gut. Dies läßt sich durch die Anwesenheit eines der Anfragestruktur sehr ähnlichen Moleküls im Trainingsdatensatz erklären:

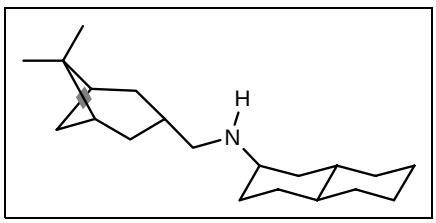


Abb. 2-41: Molekül des Trainingsdatensatzes, das der Anfragestruktur sehr ähnlich ist

Die Experimente wurden auf einer SUN Sparc 10/40 Workstation mit 32 MB Arbeitsspeicher durchgeführt. Der Rechenaufwand der Einzelschritte für Datensatz 1 mit 3244 Molekülen ist in Tabelle aufgeführt.

Tab. 2-15: Rechenaufwand für die Einzelschritte bei der Simulation mit einem globalen Netz

Schritt	Zeitaufwand [min]	Dateigröße [MB]
Ausgangsdatei (Konnektiviätsliste)	-	0.2
3D-Strukturgenerierung (CORINA)	8.0	5.3
berechnung physikochemischer Eigenschaften (PETRA)	11.7	24.9
Netztraining (70 x 70 Neuronen mit je 64 Gewichten)	99.8	2.8
Abfrage des Netzes (Simulation für Datensatz 2)	3.5	-

2.4.2.2 Spezialisierte Netze

Bei diesem Ansatz werden für bestimmte Substanzgruppen, z.B. Pyridin- oder Chinolinderivate, Datensätze ausgewählt und Netze trainiert. Dieselbe Methode wurde bereits bei den Versuchen zur Auswahl geeigneter Codierungsparameter in Kapitel 2.4.1 angewandt. Die Datensätze, die beispielsweise durch einen Substruktursuchalgorithmus aus der SpecInfo-Datenbank ausgewählt wurden, werden in einem zweiten Schritt, ähnlich wie bei obigem Ansatz mit dem globalen Netz, mittels eines CPG-Netzes in einen Trainings- und einen Testdatensatz aufgeteilt.

Im Rahmen dieser Arbeit wurden Simulationsexperimente für Pyridin-, Naphthalin-, Chinolin- und Isochinolinderivate durchgeführt. Von Jan Schuur, einem weiteren Mitarbeiter der Arbeitsgruppe, wurden weiterhin Simulationsexperimente für unterschiedlich substituierte Benzolverbindungen durchgeführt.[48][49] Diese Substanzgruppen sind zahlenmäßig sehr unterschiedlich in der SpecInfo IR-Datenbank repräsentiert. Die Datensätze und ihre Aufteilung in Trainings- und Testdatensätze ist in nachfolgender Tabelle aufgeführt:

Tab. 2-16: Untersuchte Datensätze

Verbindungen	Moleküle im Trainingsdatensatz	Moleküle im Testdatensatz
Benzolderivate	487	381
Pyridinderivate	69	80
Naphthalinderivate	32	46
Chinolin- und Isochinolinderivate	51	72

Bei den Experimenten wurden folgende Simulationsparameter verwendet:

Tab. 2-17: Simulationsparameter

Strukturcodierung	64 3D-MoRSE-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	Aufteilung durch neuronales Netz
Anzahl der Trainingsmoleküle	487, 69, 32, 51
Datenbasis	SpecInfo, (ungeladen, H, C, N, O, Hal)
Neuronen	10 x 10
Netzwerkform	planar
Training	unüberwacht

Für die einzelnen Experimente wurden die Korrelationskoeffizienten zwischen simulierten und experimentellen Spektren berechnet. Die Verteilung für das Auftreten der verschiedenen Korrelationskoeffizienten ist in Graphik 2-42 dargestellt:

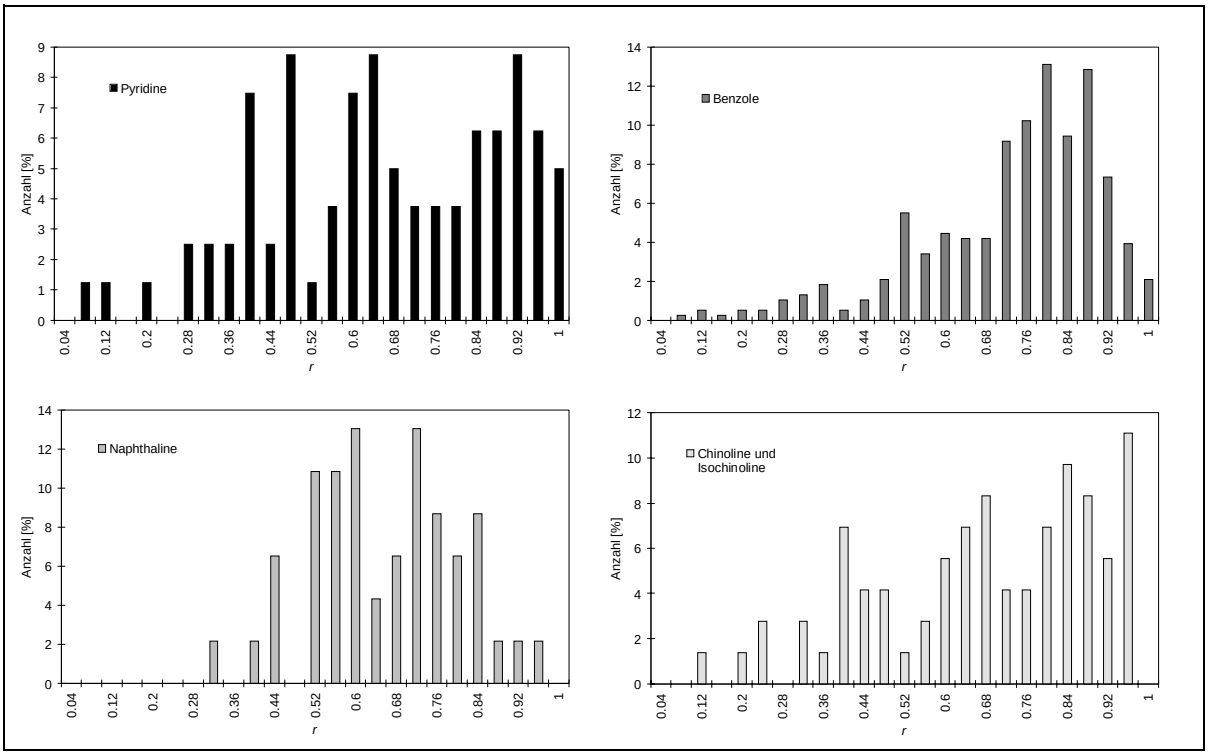


Abb. 2-42: Verteilung der Simulationsqualitäten bei den verschiedenen Datensätzen

Die Verteilungen der *r*-Werte fallen bei den vier Versuchen sehr unterschiedlich aus. Einzig bei den Benzolsimulationen ist eine Häufung bei $r = 0.85$ zu beobachten. Bei den anderen Versuchen sind die Häufungen der *r*-Werte eher zufällig verteilt. Dies läßt sich auf die Tat-

sache zurückführen, daß der Benzoldatensatz auch etwa die fünffache Anzahl an Molekülen enthält, wie die anderen drei Datensätze (vgl. 2-16). Im nachfolgenden sollen je zwei Benzol- und Chinolinsimulationen ausführlicher diskutiert werden.

Abbildung 2-43 zeigt die Benzolsimulation mit dem höchsten Korrelationskoeffizienten *r* von 0.989:

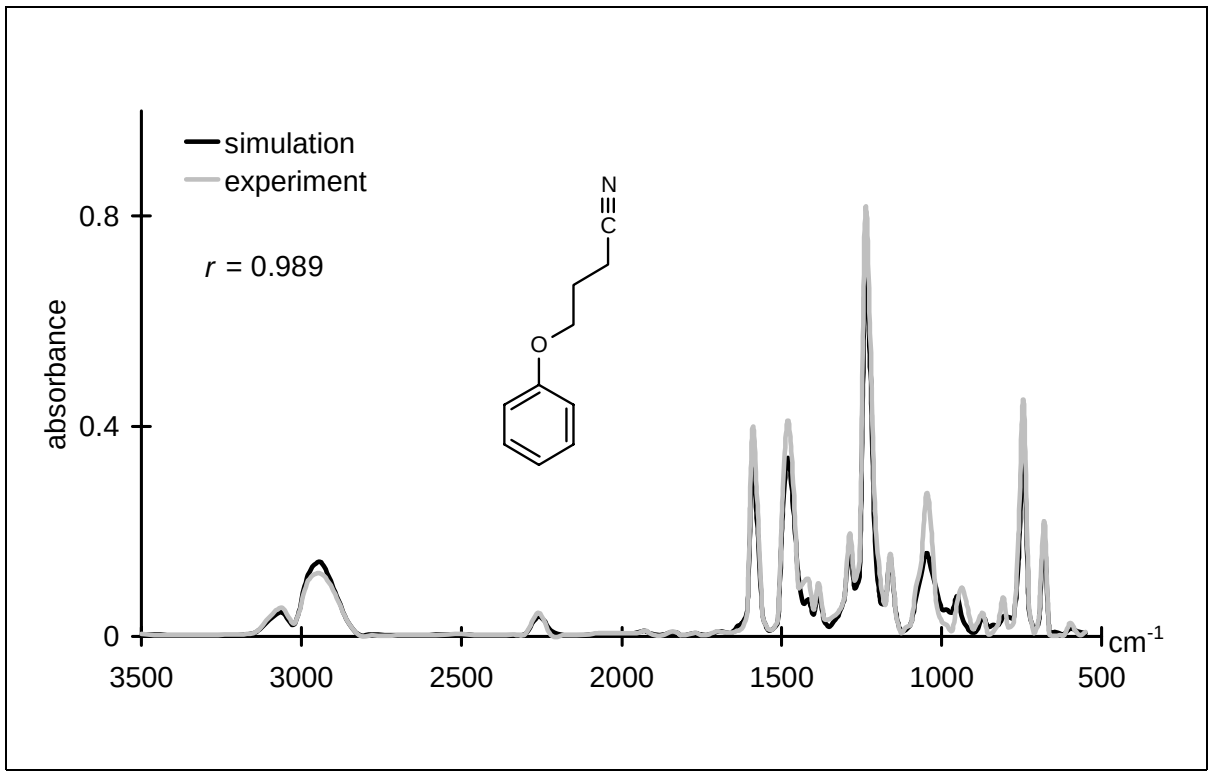


Abb. 2-43: IR-Simulation für 4-Phenoxy-n-butannitril

Simuliertes und experimentelles Spektrum sind sehr ähnlich (vgl. Abb. 2-43). Auch hier liegt der Grund für die gute Simulation wieder darin, daß im Trainingsdatensatz eine der Anfragestruktur sehr ähnliche Verbindung enthalten ist, nämlich das 5-Phenoxy-n-Pentannitril.

Das nachfolgende Beispiel zeigt eine Simulation von geringerer Qualität:

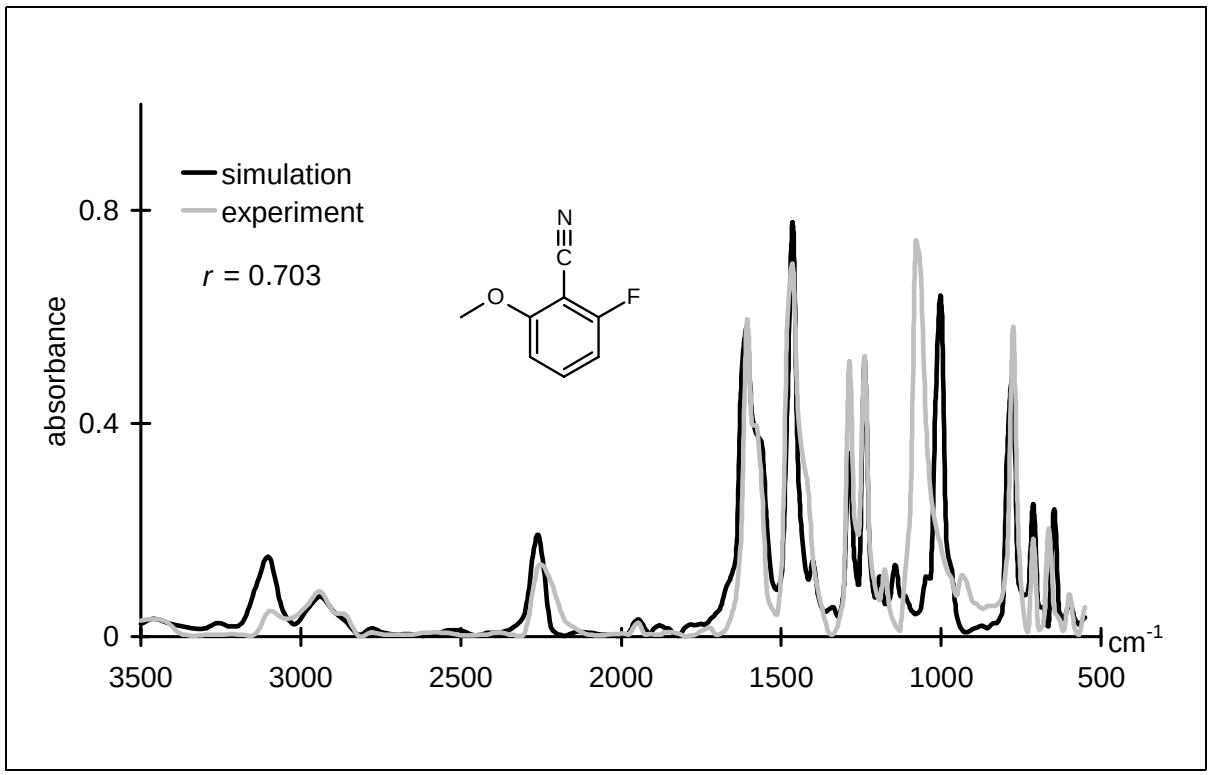


Abb. 2-44: IR-Simulation für 6-Fluor-2-methoxy-benzonitril

Simuliertes und experimentelles Spektrum zeigen über weite Bereiche gute Übereinstimmung. Im Fingerprintbereich ist bei 1084 cm^{-1} die Abweichung eines intensitätsstarken Signals zu beobachten. Dieses auf die Valenzschwingung der Methoxygruppe zurückzuführende Signal erfährt durch die elektronischen Effekte der Nachbargruppen eine hypsochrome Verschiebung. Da im Trainingsdatensatz keine Moleküle mit einer ähnlichen Nachbarschaft enthalten waren, wird die Lage dieses speziellen Signals falsch vorhergesagt. Nachfolgende Abbildung zeigt das simulierte Spektrum des Chinolin-Testdatensatzes mit dem höchsten Korrelationskoeffizienten:

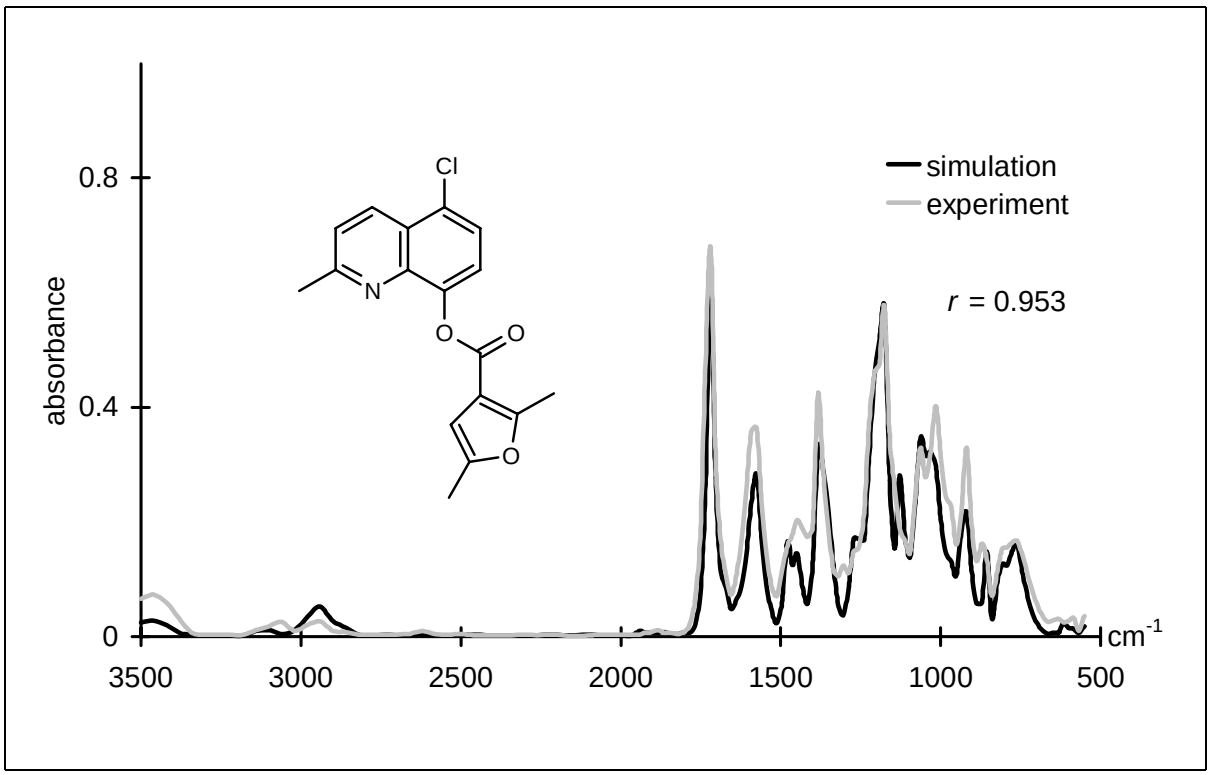


Abb. 2-45: IR-Simulation für 5-Chlor-2-methyl-8-chinolinylester-2,5-dimethyl-3-furancarboxylsäure

Trotz der relativ komplexen Molekülstruktur, ist die Vorhersage des Infrarotspektrums von hoher Qualität. Auch hier waren im Trainingsdatensatz zwei Moleküle mit dem identischen Molekülgerüst enthalten, von denen sich eines nur durch das Fehlen der Methylgruppe vom Anfragemolekül unterscheidet, was sich im Spektrum jedoch kaum bemerkbar macht:

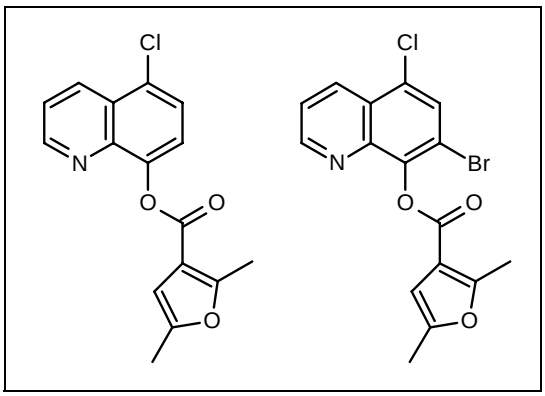


Abb. 2-46: Zwei der Anfragestruktur sehr ähnlichen Moleküle, die im Trainingsdatensatz enthalten waren

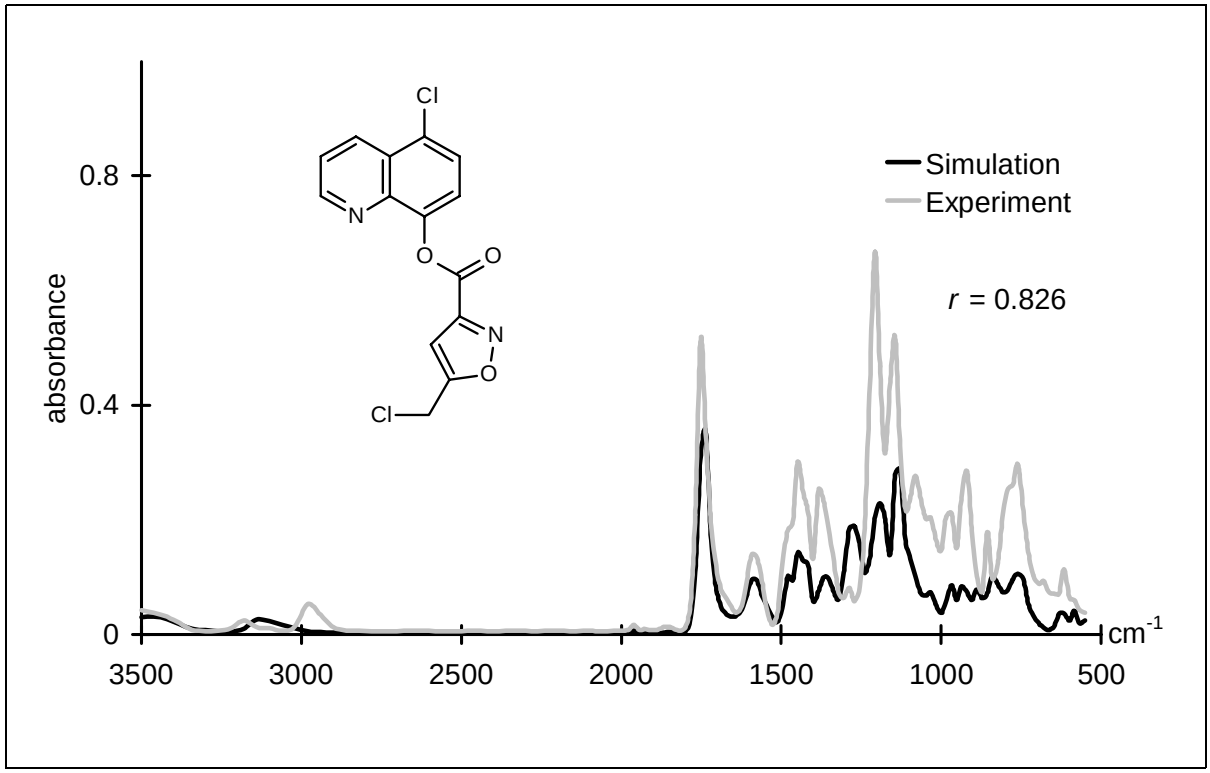


Abb. 2-47: IR-Simulation für 5-Chlor-8-chinolinylester-5-chlormethyl-3-isoxazolcarboxylsaäure

Bei der Simulation in Abbildung 2-47 sind bereits mehr Abweichungen zu beobachten, als in dem Beispiel davor. Dennoch sind einige Signalmuster des experimentellen Spektrums im simulierten Spektrum wiederzufinden. Auch hier sind einige sehr ähnliche Strukturen im Trainingsdatensatz enthalten:

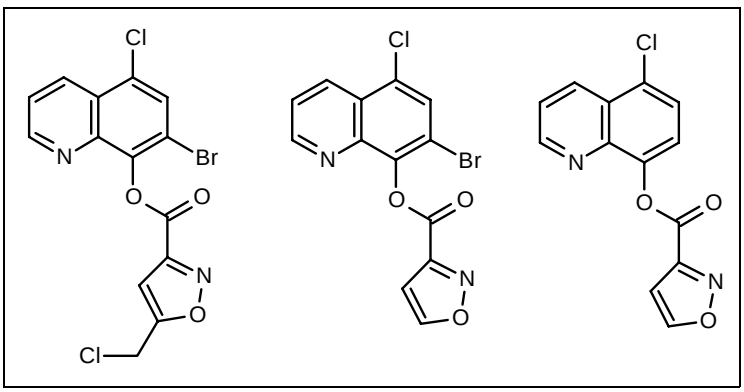


Abb. 2-48: Drei Moleküle des Trainingsdatensatzes mit einem identischen Grundgerüst

Die strukturellen Unterschiede, die sie gegenüber der Anfragestruktur aufweisen, führen jedoch bereits zu deutlich unterschiedlichen Spektren.

Der Ansatz dedizierter Netze für bestimmte Substanzgruppen läßt sich sinnvoll anwenden, wenn ein eng umrissener Substanzbereich spektroskopiert wird und in diesem Bereich auch bereits eine ausreichende Menge an Datenmaterial zur Verfügung steht. Er hat gegenüber dem Ansatz des globalen Netzes den Vorteil, daß nicht der unnötige Datenbankballast mitgezogen wird, der einerseits größere Anforderungen an Hardwarekapazitäten stellt (vgl. Tab. 2-15), und andererseits auch die Simulationsergebnisse verrauschen und negativ beeinflussen kann. Ein Problem dieses Ansatzes kann jedoch die Einteilung bzw. Suche nach Substanzklassen sein. In Fällen größerer Molekülgerüste, die dann möglicherweise auch noch mehrfach substituiert sind, läßt sich die Zuordnung zu einer bestimmten Substanzklasse nicht eindeutig treffen. So ist nicht klar zu festzulegen ob nachfolgende Verbindung den Chinolinen, den Furanen oder den Estern zugeordnet werden soll. Die sinnvollste Lösung wäre es, das Molekül in alle drei Substanzgruppen aufzunehmen.

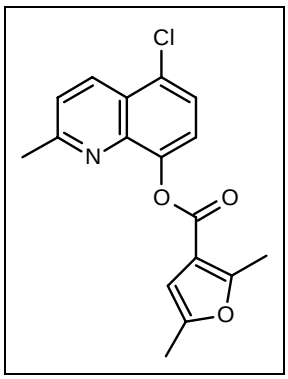


Abb. 2-49: Zuordnungsproblematik

2.4.2.3 Anfragestrukturorientierter Ansatz

Der anfragestrukturorientierte Ansatz kann als konsequente Fortführung der beiden obigen Ansätze betrachtet werden. Während beim globalen Ansatz ein Netz für die gesamte organische Chemie bzw. für den Ausschnitt der in der SpecInfo IR-Datenbank enthalten ist trainiert wurde, wurde beim zweiten Ansatz bereits eine Spezialisierung der Netze angestrebt, um so eine bessere Anpassung an die jeweilige Testsubstanz und damit eine hochwertigere Simulation zu erreichen. Beim anfragestrukturorientierten Ansatz wird die Spezialisierung noch einen Schritt weitergetrieben, indem für jedes einzelne Anfragemolekül ein Trainingsdatensatz ausgewählt wird. Als Trainingsdatensatz werden die 50 Moleküle der SpecInfo IR-Datenbank ausgewählt, deren Strukturcodes dem Strukturcode der Anfragestruktur am ähnlichsten sind (vgl. Abb. 2-50).

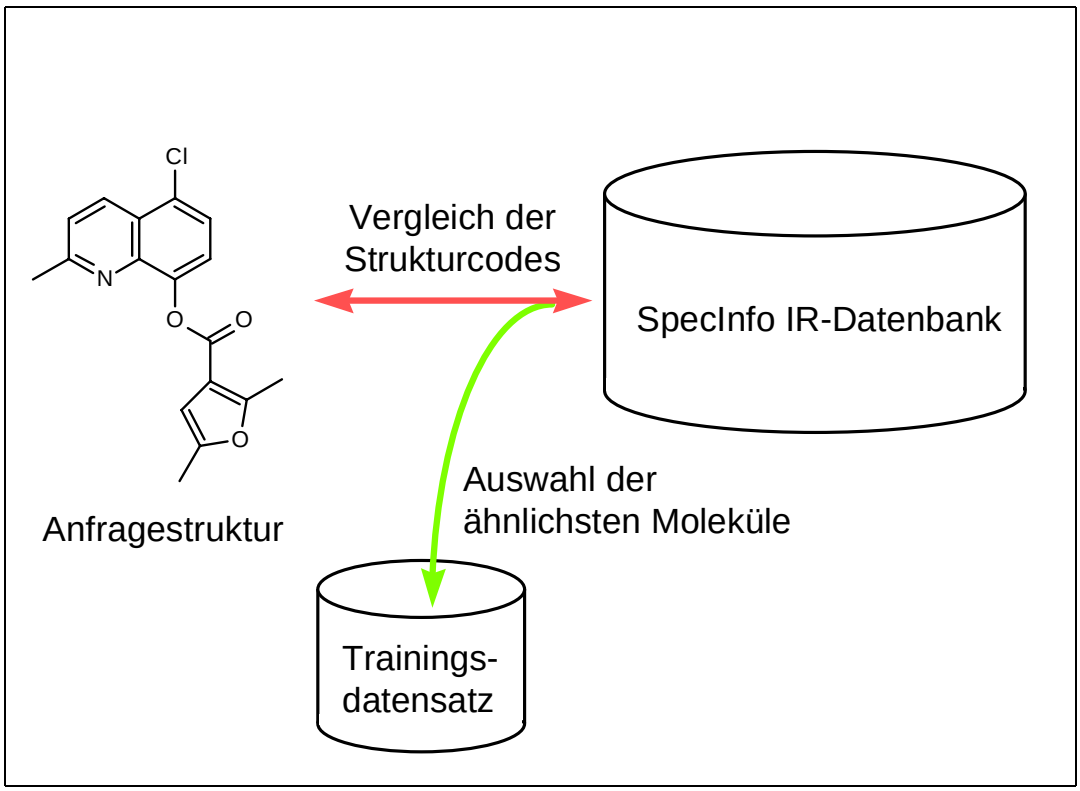


Abb. 2-50: Trainingsdatensatzauswahl beim anfragestrukturorientierten Ansatz

Als Vergleichsmaß für die Strukturcodes wird der *rms*-Wert verwendet.

Im ersten Beispiel für den Ansatz mit anfragestrukturorientierter Trainingsdatensatzauswahl wurde das IR-Spektrum für Citronellal simuliert. Bei den Experimenten wurden folgende Simulationsparameter verwendet:

Tab. 2-18: Simulationsparameter

Strukturcodierung	64 3D-MoRSE-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfragestrukturorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	9850 ungeladene Verbindungen der SpecInfo-Datenbank mit den Elementen: H, C, N, O, Hal
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

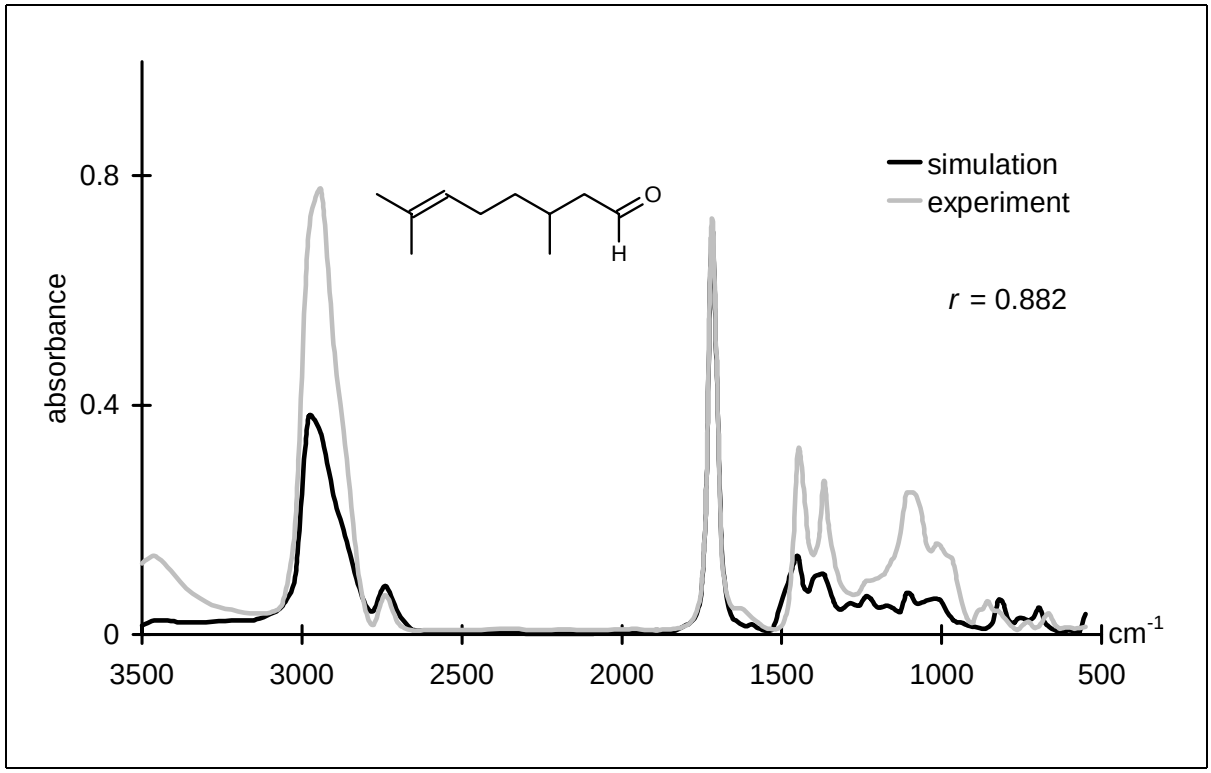


Abb. 2-51: IR-Spektrensimulation für Citronellal

Das simulierte Spektrum ist dem experimentellen Spektrum sehr ähnlich. Das Signal für den Carbonylpeak ist in simuliertem und experimentellem Spektrum nahezu deckungsgleich. Bei den übrigen Signalen sind die Absorbanzwerte des simulierten Spektrums niedriger als die des experimentellen Spektrums. Die Signalmuster und -positionen des experimentellen Spektrums sind jedoch im simulierten Spektrum sehr gut wiedergegeben. Beim Netzwerktypus des Counterpropagation-Netzwerks bieten sich sehr gute Möglichkeiten zur Analyse des Netzes. Von besonderem Interesse ist dabei, welche Moleküle beim Netztraining auf welche Neuronen gefallen sind. Hier gilt es in erster Linie das Gewinnerneuron, also dasjenige Neuron aus welchem das simulierte Spektrum entnommen wurde, und dessen Nachbarschaft zu untersuchen, da die Moleküle auf diesen Neuronen den größten Einfluß auf das simulierte Spektrum haben. Das Netz war bei diesem Experiment nicht planar, sondern toroidal. Dies bedeutet, daß die jeweils gegenüberliegenden Kanten miteinander verbunden sind. Die Draufsicht auf das Netz sowie ein Ausschnitt von 3 x 3 Neuronen sind in nachfolgender Abbildung dargestellt:

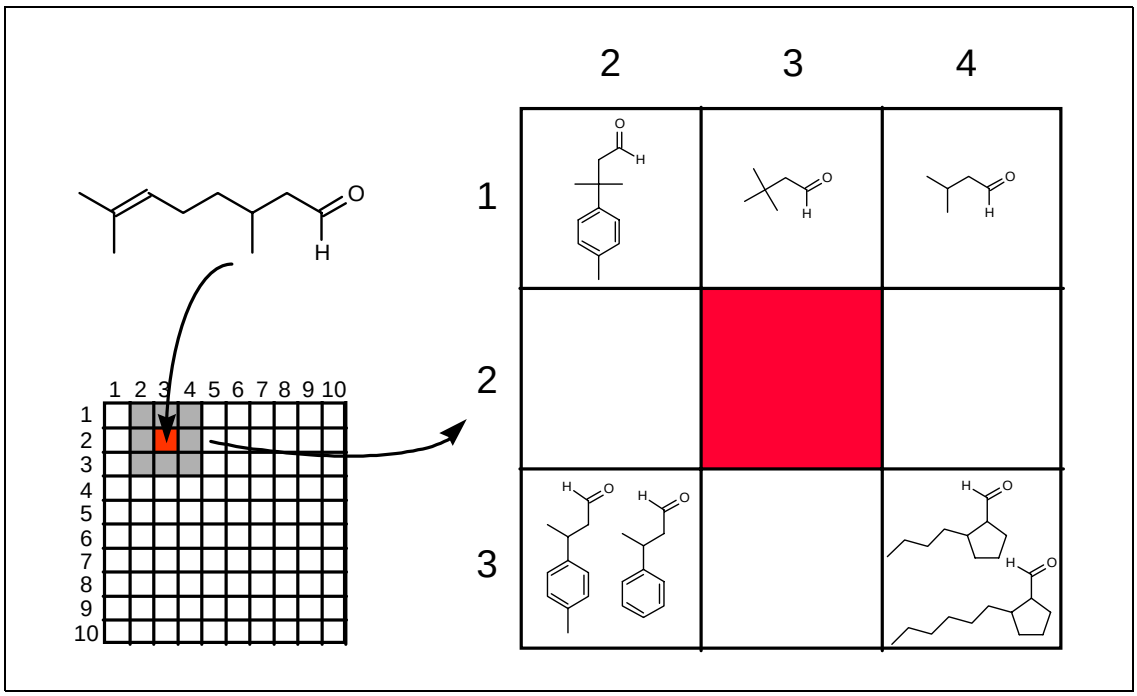


Abb. 2-52: Abbildung des neuronalen Netzes zur Simulation des Spektrums von Citronellal. Das Gewinnerneuron ist dunkelgrau schraffiert.

Dem Gewinnerneuron ist in diesem Fall während des Netztrainings kein Molekül zugeordnet worden. Dieses Verhalten ist sehr oft zu beobachten. Es bedeutet, daß der Strukturcode des Gewinnerneurons, der ein Interpolationspunkt der Nachbarneuronen darstellt, dem Anfragestrukturcode ähnlicher ist als der Strukturcode von Neuronen, die im Training mit Molekülen belegt worden sind. Das ist ein wichtiger Aspekt, da hier deutlich wird, daß das Netz aus vorhandenen Daten neue Information entwickelt hat, um der Anfrage so gut wie möglich gerecht zu werden. Es fällt auf, daß die dargestellten Trainingsmoleküle verschiedene Strukturmerkmale der Anfragestruktur aufweisen. Das Netz war somit in der Lage aus den verschiedenen Strukturmerkmalen und den entsprechenden Spektrensignalen ein entsprechendes Spektrum für die Anfragestruktur zusammenzusetzen. Auch wenn zwei auftretende Strukturmerkmale, nämlich der Phenylring der Moleküle auf Neuron (3,2) und der Fünfring der Moleküle auf Neuron (3,4), nicht in der Anfragestruktur zu finden sind, so sind die diesen Gruppen entsprechenden Schwingungen den olefinischen Gerüstschwingungen des Anfragemoleküls sehr ähnlich. Dies kann als Hinweis gesehen werden, daß die infrarotrelevante Information erkannt und im Netz abgebildet wird.

Bei der Erklärung des Counterpropagation-(CPG)-Netzes wurde bereits erwähnt, daß der obere Block des neuronalen Netzes die Strukturcodes und der untere Teil die Infrarotspektren enthält. Jede Schicht des Spektrenblocks entspricht dabei einer Wellenzahl:

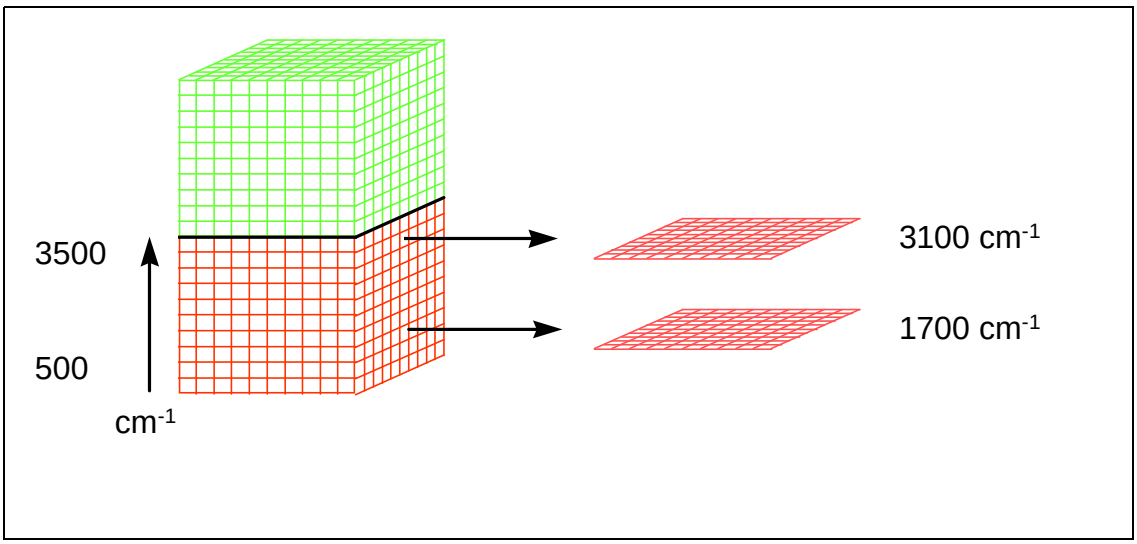


Abb. 2-53: Jede Schicht des Spektrenblocks (unterer Teil) entspricht einer Wellenzahl.

Durch die Analyse einzelner Schichten kann festgestellt werden, welche Neuronen und damit welche Trainingsmoleküle bei verschiedenen Wellenzahlen das simulierte Spektrum beeinflußt haben. So ist das Zustandekommen einer guten Simulation oder einer Abweichung leicht nachzuvollziehen. Bei dem Simulationsexperiment für Citronellal wurden die Wellenzahlen 1720 (ν C=O), 1448 (ν C=C olefinisch) und 1112 cm^{-1} (ν C-Me) untersucht, da an diesen Positionen im experimentellen Spektrum deutliche Signale zu beobachten sind. In allen drei Fällen sind an den entsprechenden Stellen auch im simulierten Spektrum Signale zu finden, die mit Ausnahme des Carbonylpeaks jedoch deutlich geringere Intensitäten aufweisen. Die folgende Abbildung zeigt wieder den Ausschnitt mit 3 x 3 Neuronen wie in Abbildung 2-52, wobei die Farben die Absorbanzwerte bei dieser Wellenzahl anzeigen.

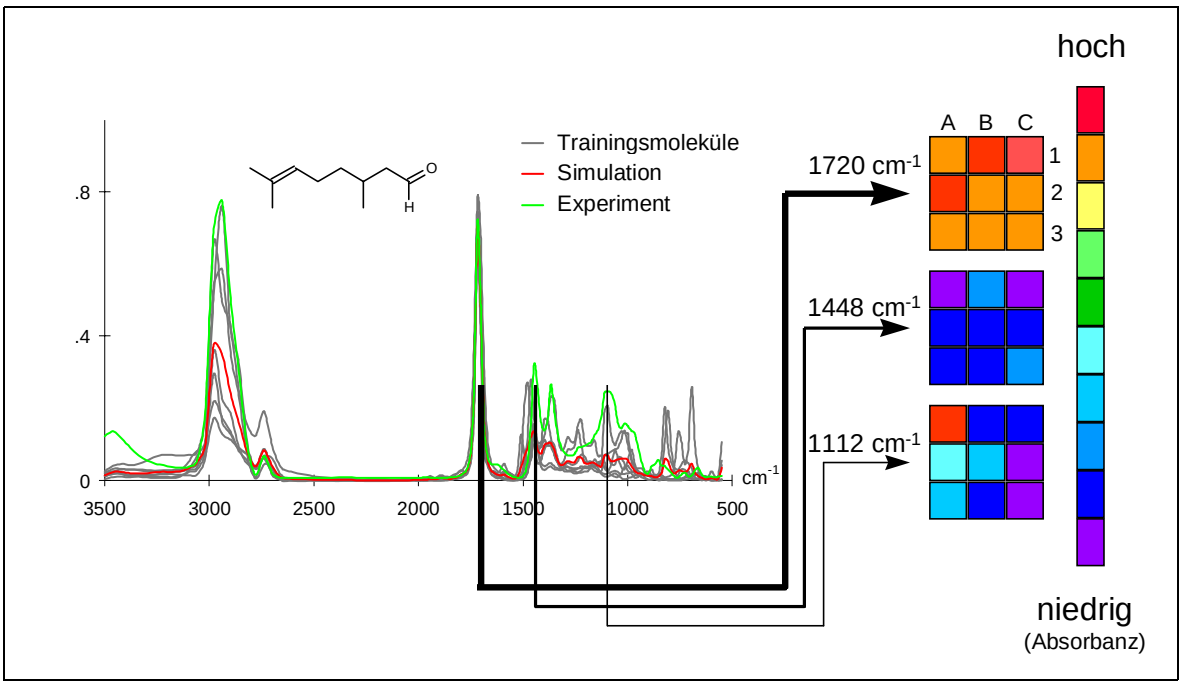


Abb. 2-54: Untersuchung verschiedener Wellenzahlschichten bei der Citronellal-Simulation

In der Spektrenschicht von 1720 cm^{-1} (ν C=O) zeigen alle Umgebungsneuronen hohe bis sehr hohe Werte der Absorbanz. Die beiden Neuronen mit sehr hohen Werten in der oberen Reihe (1,B und 1,C) sind im Training mit Molekülen belegt worden (vgl. Abb. 2-52). Das Neuron in der ersten Spalte der zweiten Reihe (2,A) hat ebenfalls einen sehr hohen Wert, ist jedoch im Training nicht mit einem Molekül belegt worden. Der hohe Absorbanzwert dieses Neurons stellt einen Interpolationspunkt zwischen dessen Umgebungsneuronen dar. Hier sind auch Neuronen, die im Bezug auf das Gewinnerneuron in der zweiten Nachbarschaftssphäre liegen und in dieser Abbildung nicht mehr dargestellt sind, beteiligt. Das Signal im simulierten Spektrum bei 1448 cm^{-1} (ν C=C olefinisch) gibt zwar das Bandenmuster des experimentellen Spektrums sehr gut wieder, liegt jedoch mit seinem Absorbanzwert deutlich unter dem des experimentellen Spektrums. Die größten Beiträge zeigen hier das mittlere Neuron in der ersten Reihe (1,B) sowie das Neuron in der rechten Spalte der letzten Zeile (3,C), welches im Training mit den Cyclopentylen belegt worden ist. Allgemein sind die Absorbanzwerte der Umgebungsneuronen des Gewinnerneurons jedoch alle relativ ähnlich. Bei der Untersuchung der Schicht von 1112 cm^{-1} (ν C-Me) zeigt hingegen das Neuron der linken Spalte der ersten Reihe (1,A) einen deutlich höheren Wert als alle anderen und bewirkt so zumindest ein kleines Signal im simulierten Spektrum. Dieses Neuron wurde im Training mit einem Molekül belegt, das ebenso wie die Anfragestruktur zwei geminale Methylgruppen trägt (vgl. Abb. 2-52) und deshalb auch die ähnlichste ν C-Me Schwingung zeigt.

Beim nächten Experiment wurde die anfragestrukturorientierte Spektrensimulation für Cyanazin durchgeführt. Die folgende Abbildung zeigt den Vergleich von simuliertem und experimentellem Spektrum:

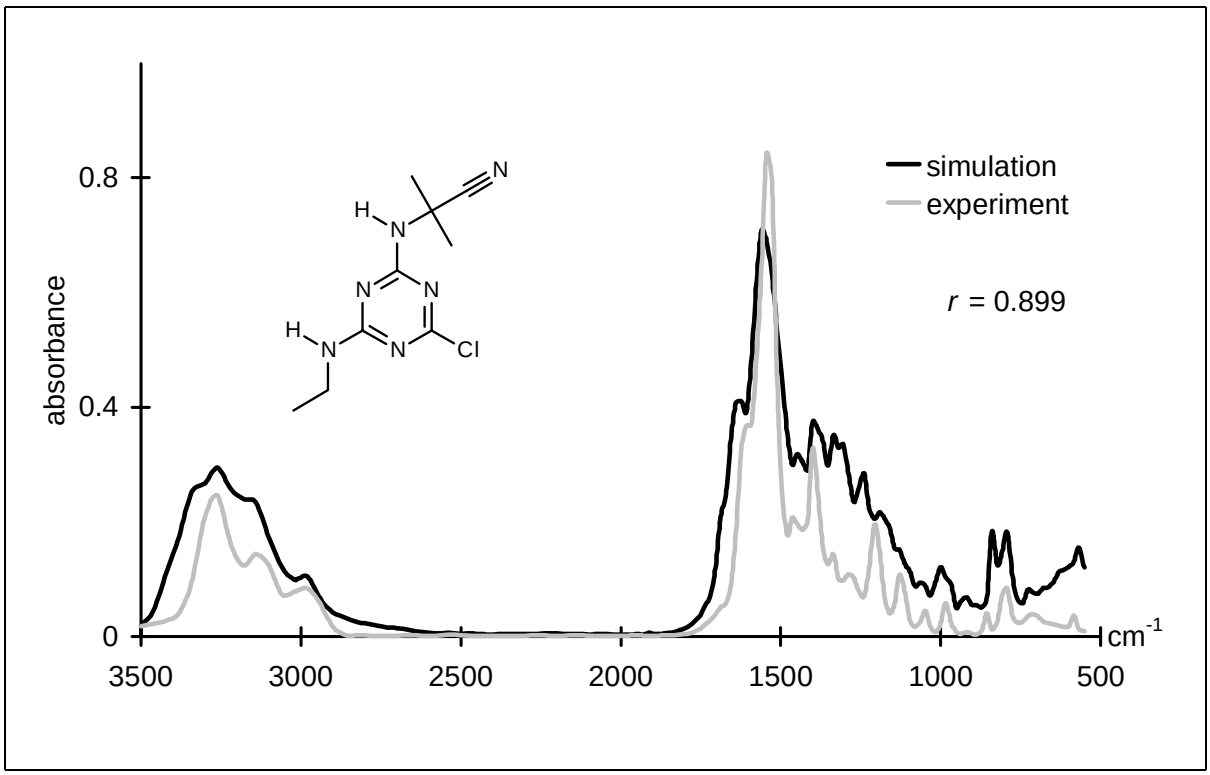


Abb. 2-55: Vergleich von simuliertem und experimentellem Spektrum von Cyanazin

Das simulierte Spektrum gibt bis auf wenige Ausnahmen im Fingerprint-Bereich die Signalmuster des experimentellen Spektrums sehr gut wieder. Sogar die Signale im Bereich oberhalb 3000 cm^{-1}, welche sehr stark von den Aufnahmebedingungen abhängen, sind im simulierten Spektrum wiederzufinden. Dies wird durch eine hohe Zahl von Triazinverbindungen, die unter gleichen Bedingungen vermessen worden sind, bewirkt. Ebenso sind die Signale der ν C=N Schwingung zwischen 1500 und 1600 cm^{-1} sehr gut wiedergegeben. Im Bereich der ν C-N Schwingungen zwischen 1000 und 1400 cm^{-1} sind jedoch einige Abweichungen zu beobachten. Auch das Dublett bei etwa 700 cm^{-1}, wie es bei substituierten Aromaten häufig auftritt, ist im simulierten Spektrum deutlich zu erkennen. Weiterhin fällt auf, daß das experimentelle Spektrum zwischen 2210 und 2260 cm^{-1} keinen Nitrilpeak zeigt, obwohl eine entsprechende Funktionalität im Molekül enthalten ist. Bei capto-dativen Systemen, wie hier eines vorliegt, kann es mitunter vorkommen, daß das Nitrilsignal nicht im Spektrum zu finden ist. Bemerkenswert ist, daß im simulierten Spektrum auch kein Nitrilsignal vorhergesagt wird. Die Moleküle, die im Training auf die Nachbarneuronen des Gewinnerneurons gefallen sind,

sind in nachfolgender Abbildung dargestellt:

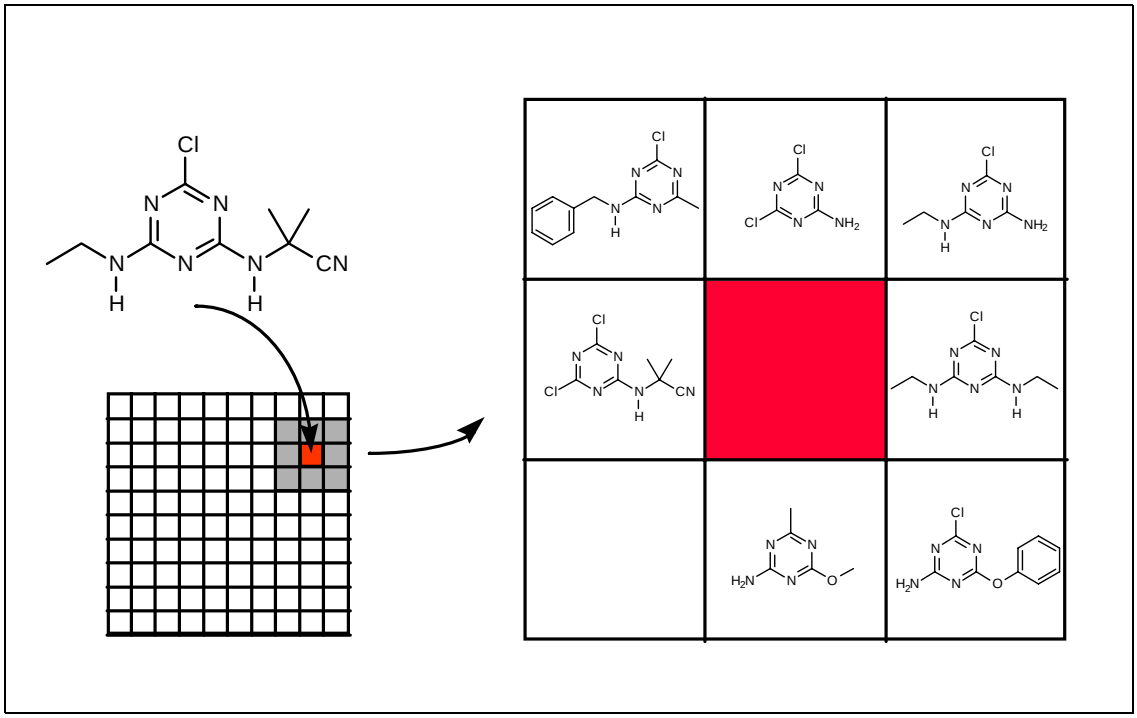


Abb. 2-56: Abbildung des neuronalen Netzes zur Simulation des Spektrums von Cyanazin. Das Gewinnerneuron ist dunkelgrau schraffiert.

Auch bei diesem Beispiel ist das Gewinnerneuron im Training mit keinem Molekül belegt worden. Alle Moleküle, die in der ersten Nachbarschaftssphäre des Gewinnerneurons zu liegen kommen, besitzen einen Triazin-Grundkörper. Die Moleküle sind weiterhin 1,3,5-substituiert, wobei die Substituenten jedoch sehr unterschiedlich sind. So sind beispielsweise auf zwei Neuronen in der dritten Reihe im Training ethersubstituierte Moleküle gefallen. Diese Verbindungen scheinen zunächst sehr wenig mit der Anfragestruktur gemeinsam zu haben. Betrachtet man diesen Sachverhalt jedoch aus der Perspektive der IR-Spektroskopie, so stellt man fest, daß die entsprechenden Schwingungen für ν C-O und ν C-N sehr ähnlich sind und bei etwa 1000-1260 cm^{-1} bzw. 1000-1400 cm^{-1} liegen. Die Auswahl dieser Verbindungen zum Netztraining kann somit als sinnvoll erachtet werden.

Auch hier sollen wieder verschiedene Wellenzahlschichten näher analysiert werden, nämlich 1560 cm^{-1} (ν C=N), 1400 cm^{-1} (ν C-N) und 1192 cm^{-1}(ν C-N). In nachfolgender Abbildung ist wiederum der 3 x 3 Neuronen-Ausschnitt des Netzes dargestellt (vgl. Abb. 2-56), in dessen Zentrum das Gewinnerneuron ist. Die Farbe der Neuronen beschreibt die Höhe des Absorbanzwertes des Neurons bei dieser Wellenzahl.

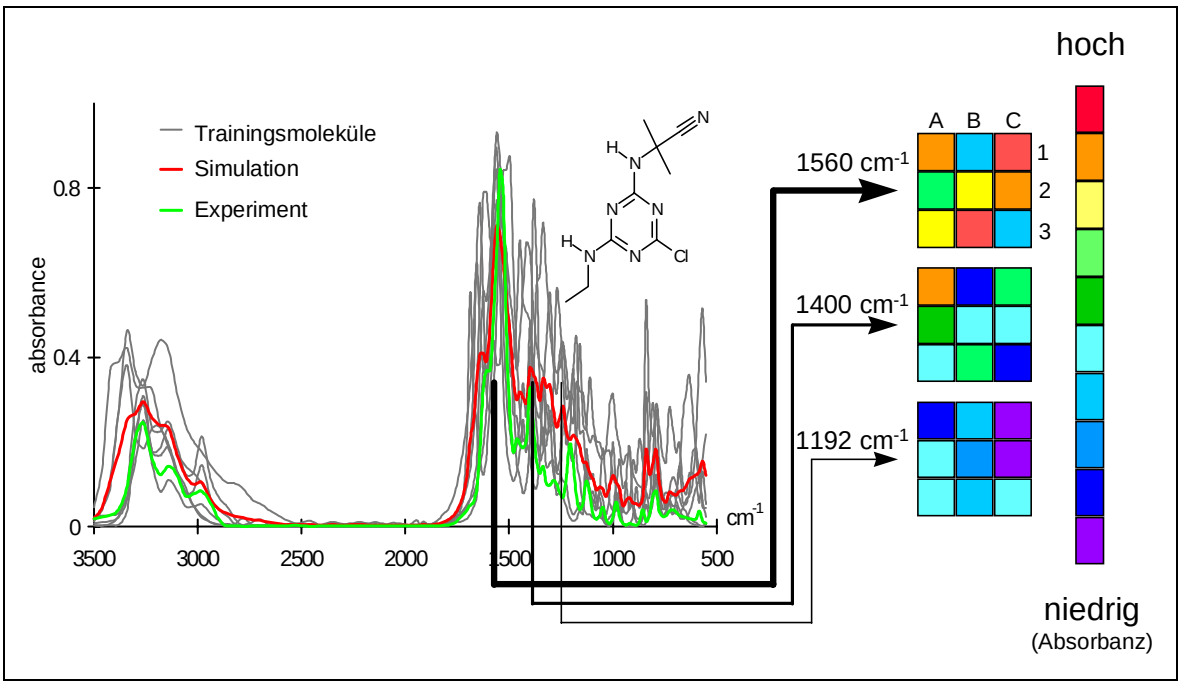


Abb. 2-57: Untersuchung verschiedener Wellenzahlschichten bei der Cyanazin-Simulation

Die Kurven der überlagerten Trainingsspektren zeigt eine viel größere Vielfalt als bei dem Simulationsexperiment für Citronellal (vgl. Abb. 2-54). Ebenso zeigen die Absorbanzwerte der Umgebungsneuronen bei den Wellenzahlen 1560 cm^{-1} und 1400 cm^{-1} größere Differenzen untereinander und zum Gewinnerneuron als beim vorherigen Beispiel. Wie bereits erwähnt wurde, tragen alle Moleküle, die im Training auf Neuronen der ersten Nachbarschaftssphäre gefallen sind, einen Triazin Grundkörper. Somit werden die dazugehörigen Spektren auch zwischen 1520 und 1690 cm^{-1} die entsprechenden Signale für die ν C=N Schwingung zeigen. Die Lage dieser Signale hängt jedoch von den jeweiligen Substituenten und deren elektronischer Natur ab. Wird nun die der Wellenzahl von 1560 cm^{-1} entsprechende Schicht untersucht, so kann festgestellt werden, bei welchen Trainingsmolekülen die ν C=N Schwingung an gleicher oder ähnlicher Stelle wie beim Anfragemolekül zu beobachten ist und bei welchen nicht. Hier fällt auf, daß die beiden Neuronen (3,B und 3,C) mit ethersubstituierten Verbindungen (vgl. Abb. 2-56) sehr unterschiedliche Absorbanzwerte besitzen. Während das Neuron mit der methoxysubstituierten Verbindung (3,B) einen sehr hohen Absorbanzwert hat und damit dem Wert des experimentellen Spektrums sehr ähnlich kommt, ist auf dem Neuron mit dem Phenoxymolekül (3,C) ein relativ niedriger Wert zu finden. Bei 1192 cm^{-1} ist im experimentellen Spektrum kein Signal zu beobachten, wohl aber im simulierten, was als Fehler anzusehen ist. Die Neuronen, welche die niedrigsten Absorbanzwerte besitzen sind die beiden Neuronen der ersten und zweiten Reihe der rechten Spalte (1,C und 2,C). Die Moleküle auf diesen Neuronen tragen beide einen Ethylaminosubstituenten und zeigen somit ein ähnliches ν C-N

Schwingungsverhalten, wie das Anfragemolekül. Das Molekül mit zwei Chlor- und einem Aminosubstituenten auf dem mittleren Neuron der ersten Reihe (1,B), zeigt bei den Wellenzahlen 1560 und 1400 cm^{-1} sehr niedrige Absorbanzwerte, bei der Wellenzahl 1192 cm^{-1}, wo das experimentelle Spektrum kein Signal zeigt, einen verhältnismäßig hohen Absorbanzwert. Das Neuron hat damit einen negativen Einfluß auf das Simulationsergebnis. Die Auswahl des Moleküls dieses Neurons in den Trainingsdatensatz muß als ungünstig erachtet werden. Wie anfangs beschrieben, werden bei dem Ansatz mit anfragestrukturorientierter Trainingsdatensatzauswahl die 50 ähnlichsten Moleküle aus der Basisdatenbank in den Trainingsdatensatz aufgenommen. Dies werden für den Fall, daß die Anfragestruktur gut durch ähnliche Moleküle in der Datenbank repräsentiert ist, auch wirklich 50 ähnliche Verbindungen sein. Für den Fall, daß die Verbindung schlecht repräsentiert ist, werden es jedoch die 50 “am wenigsten unähnlichen” Moleküle sein. Diese weisen nur eine geringe Ähnlichkeit mit der Anfragestruktur auf und führen somit zu einer qualitativ minderwertigen Spektrenvorhersage. Ein einfacher Grenzwert für den *rms*-Wert bei der Trainingsdatensatzauswahl, ab dem ein Molekül als zu unähnlich angesehen wird und nicht mehr in den Trainingsdatensatz aufgenommen wird, ist keine Lösung des Problems, da der *rms*-Wert zwischen den Strukturcodes größerer Moleküle, die jedoch verhältnismäßig ähnlich sind, höher ausfallen kann als bei unähnlichen kleinen Molekülen. Eine solche Abwägung muß also momentan noch durch den Experimentierenden vorgenommen werden. Dies kann durch eine Untersuchung des Netzes und der Trainingsmoleküle geschehen, wie es für die beiden Beispielmoleküle in diesem Abschnitt demonstriert wurde.

2.4.2.4 Diskussion und Vergleich der Auswahlmethoden

Der Ansatz zur Simulation mit dem globalen Netz, also einem Netz, das mit allen zur Verfügung stehenden Daten trainiert wurde, hat den Vorteil, daß für ein beliebiges Anfragemolekül schnell eine Simulation durchgeführt werden kann. Der Nachteil ist, daß die großen Datenmengen, sowohl die Datensätze zum Training des neuronalen Netzes als auch die Dateien bei der Speicherung der Gewichte des neuronalen Netzes, schnell die Hardwareresourcen des Rechners erschöpfen können (vgl. Tab. 2-15). Eine Untersuchung des Netzwerks zur Analyse eines Simulationsexperiments, so wie es im vorigen Kapitel gezeigt wurde, wird damit sehr umständlich. Ein weiterer Nachteil dieses Verfahrens ist, daß ein Training mit allen Daten ohne eine Vorauswahl zu einer Verschlechterung der Simulationsergebnisse führen kann, da während des Trainings aufgrund der großen Anzahl an Trainingsmolekülen ein unähnliches Molekül in die Nähe des späteren Gewinnerneurons gelegt wird. Diesem Problem, der Verrauschung des Ergebnisses, wird im Ansatz mit den spezialisierten Netzen begegnet, indem eine Vorauswahl nach bestimmten strukturellen Gesichtspunkten getroffen wird. Hier besteht der gleiche Vorteil wie beim Ansatz mit dem globalen Netz nämlich, daß das Netz

bereits mit den ausgewählten Daten trainiert werden kann und die anschließende Abfrage des Netzes, also der eigentliche Simulationsschritt, dann sehr schnell geht. Die Datenmengen sind hier wesentlich geringer als bei obigem Ansatz. Ein Nachteil dieses Ansatzes ist, daß durch die Vorauswahl der Trainingsmoleküle anhand bestimmter Strukturmerkmale möglicherweise Moleküle nicht in den Trainingsdatensatz gelangen, die jedoch für die Simulation günstig gewesen wären. Aus diesem Grund erscheint es sinnvoll, für jede Anfragestruktur einen eigenen Trainingsdatensatz auszuwählen. Das Auswahlkriterium ist, wie oben beschrieben, der rms-Wert zwischen dem Anfragestrukturcode und allen Strukturcodes der Moleküle eines Basisdatensatzes, z.B. der gesamten SpecInfo-Datenbank. Dieser Ansatz erhöht die Flexibilität des Systems, um auch bei schwer einzuordnenden Strukturen noch die günstigste Auswahl an Trainingsmolekülen zu treffen. Die Ähnlichkeit der Strukturcodes der Moleküle des Trainingsdatensatzes mit dem Strukturcode der Anfragestruktur kann zudem möglicherweise als Maß für die Vorhersage der Simulationsqualität dienen. Auf diesen Aspekt soll jedoch erst in Kapitel 2.5 näher eingegangen werden. Ein Nachteil dieses Ansatzes ist, daß erst beim Abschicken einer Anfragestruktur an das Simulationssystem ein Trainingsdatensatz ausgewählt und ein Netz trainiert wird. Der Simulationsvorgang dauert hier also etwas länger, ist jedoch mit insgesamt etwa 1.5 min (SGI ORIGIN 200 mit 256 MB Arbeitsspeicher) durchaus akzeptabel.

Nach diesen prinzipiellen Überlegungen sollen nun die verschiedenen Ansätze anhand eines Testdatensatzes durchgeführt und die Simulationsergebnisse miteinander verglichen werden. Der Testdatensatz wurde aus den Triazinverbindungen der SpecInfo IR-Datenbank zusammengestellt (vgl. Anhang A.2).

Simulation mit einem globalen Netz

Bei diesem Experiment wurde ein Netz mit der gesamten SpecInfo IR-Datenbank ohne die Triazinverbindungen trainiert. Die Simulationsparameter sind in nachfolgender Tabelle aufgeführt:

Tab. 2-19: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	SpecInfo gesamt, ohne Triazine
Anzahl der Trainingsmoleküle	13327
Datenbasis	SpecInfo gesamt
Neuronen	163 x 163
Netzwerkform	toroidal
Training	unüberwacht

Der mittlere Korrelationskoeffizient $\bar{r}$ zwischen den simulierten und den experimentellen Spektren beträgt 0.731.

Simulation mit einem spezialisierten Netz

Bei der Simulation mit spezialisierten Netzen wurde das jeweilige Anfragemolekül aus dem Triazindatensatz herausgenommen und das Netz mit den verbleibenden 45 Triazinen trainiert. Die Simulationsparameter sind in nachfolgender Tabelle aufgeführt:

Tab. 2-20: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	Triazine
Anzahl der Trainingsmoleküle	46-1
Datenbasis	SpecInfo gesamt
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Der mittlere Korrelationskoeffizient $\bar{r}$ zwischen den simulierten und den experimentellen Spektren beträgt bei diesem Experiment 0.785.

Simulation mit anfragestrukturorientierter Trainingsdatensatzauswahl

Wie bereits in Kapitel 2.4.2.3 beschrieben, wurde bei diesem Ansatz für jedes Anfragemolekül ein eigener Trainingsdatensatz ausgewählt. Dazu wurden die 50 Moleküle der SpecInfo IR-Datenbank ausgewählt, deren Strukturcodes dem Strukturcode der Anfragestruktur am ähnlichsten sind. Vergleichkriterium der Strukturcodes war der *rms*-Wert.

Tab. 2-21: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfragestrukturorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo gesamt
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Der mittlere Korrelationskoeffizient $\bar{r}$ zwischen den simulierten und den experimentellen Spektren beträgt 0.833.

Beim Vergleich der mittleren Korrelationskoeffizienten ist zu beobachten, daß bei der Simulation mit der anfragestrukturorientierten Trainingsdatensatzauswahl die besten Ergebnisse erzielt wurden. Die graphische Darstellung der Verteilung der mittleren Korrelationskoeffizienten verdeutlicht dieses Ergebnis:

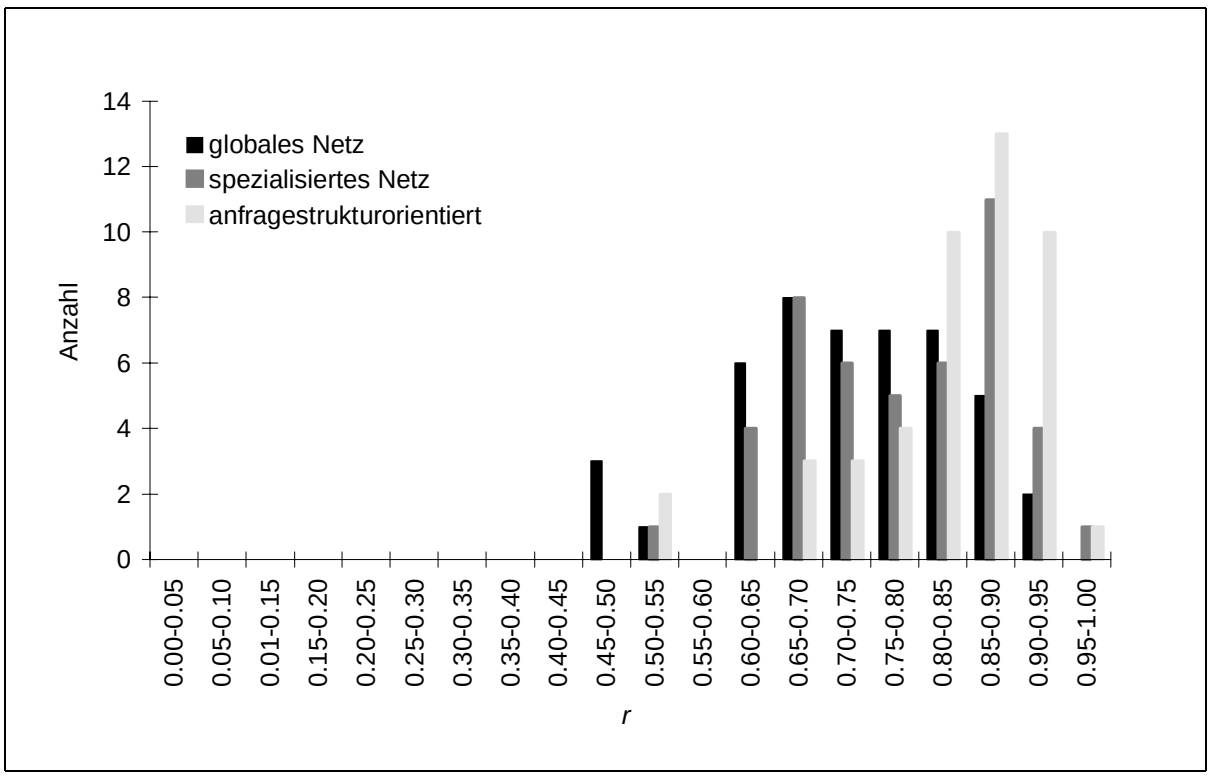


Abb. 2-58: Verteilung der mittleren Korrelationskoeffizienten $\bar{r}$ für die verschiedenen Simulationsansätze

Der anfragestrukturorientierte Ansatz schneidet bei obigem Vergleich am besten ab. Sicherlich ist dieser Ansatz aufgrund seiner Flexibilität bei den meisten Fragestellungen sinnvoll einzusetzen. Trotzdem kann in manchen Situationen, z.B. bei einer wenig umfangreichen Datenbasis, der Einsatz eines spezialisierten Netzes sinnvoll sein. Der Ansatz mit dem globalen Netz ist aufgrund der beschriebenen Nachteile und dem schlechten Abschneiden im Vergleich mit den beiden anderen Ansätzen (vgl. Abb. 2-58), als ungünstig zu bewerten.

2.5 Vorhersage der Simulationsqualität

In obigen Kapiteln wurde bereits ausführlich über Vergleichsmaße berichtet. Speziell den Vergleichsmaßen für Infrarotspektren kommt eine wesentliche Bedeutung bei der Beurteilung von Simulationsergebnissen zu. All diese Versuche hatten natürlich zur Voraussetzung, daß ein experimentelles Spektrum vorhanden war, welches mit dem simulierten Spektrum verglichen werden konnte. Bei einem Einsatz in der Praxis ist dies nun eben nicht der Fall. Die Methode findet ja gerade dann Anwendung, wenn kein experimentelles Spektrum verfügbar ist. Eine Größe, die es erlaubt die Qualität der Simulation vorherzusagen, wäre von allergrößtem Nutzen. Auch hier ist das Ergebnis in einem Vorhandensein von Ähnlichkeiten zu suchen. Als Ausgangsparameter für die Simulation stehen die Ähnlichkeit des Strukturcodes der Anfragestruktur mit den Molekülen des Trainingsdatensatzes bzw. dem Gewinner-Neuron zur Verfügung. Bei den nachfolgenden Versuchen wurde für alle ungeladenen H, C, N, O, Hal-Verbindungen (9850 Moleküle) der SpecInfo IR-Datenbank anfrageorientierte Simulationen durchgeführt und die drei nachfolgenden Größen bestimmt und gespeichert:

- *rms*-Wert zwischen dem Anfragestrukturcode und dem ähnlichsten Molekül des Trainingsdatensatzes (Experiment a)
- mittlerer *rms*-Wert zwischen dem Anfragestrukturcode und allen Molekülen des Trainingsdatensatzes (Experiment b)
- *rms*-Wert zwischen dem Anfragestrukturcode und dem Gewinner-Neuron (Experiment c)

Die Simulationsparameter sind in Tabelle 2-22 aufgeführt.

Tab. 2-22: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfragestrukturorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo, ungeladene H, C, N, O, Hal Verbindungen (9850 Moleküle)
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Diese Untersuchungen wurden erst durch den Einsatz eines SGI ORIGIN 200 Rechenservers sowie eine Geschwindigkeitsoptimierung des Programms zur Trainingsdatensatzauswahl möglich. Die Simulation für ein Molekül dauerte jedoch immer noch ca. 1 min 45 s, wobei sich eine Gesamtrechenzeit von etwa 12 Tagen ergab. Zur Auswertung der Ergebnisse wurden die drei *rms*-Werte gegen den Korrelationskoeffizienten zwischen dem jeweiligen simulierten und experimentellen Spektrum aufgetragen:

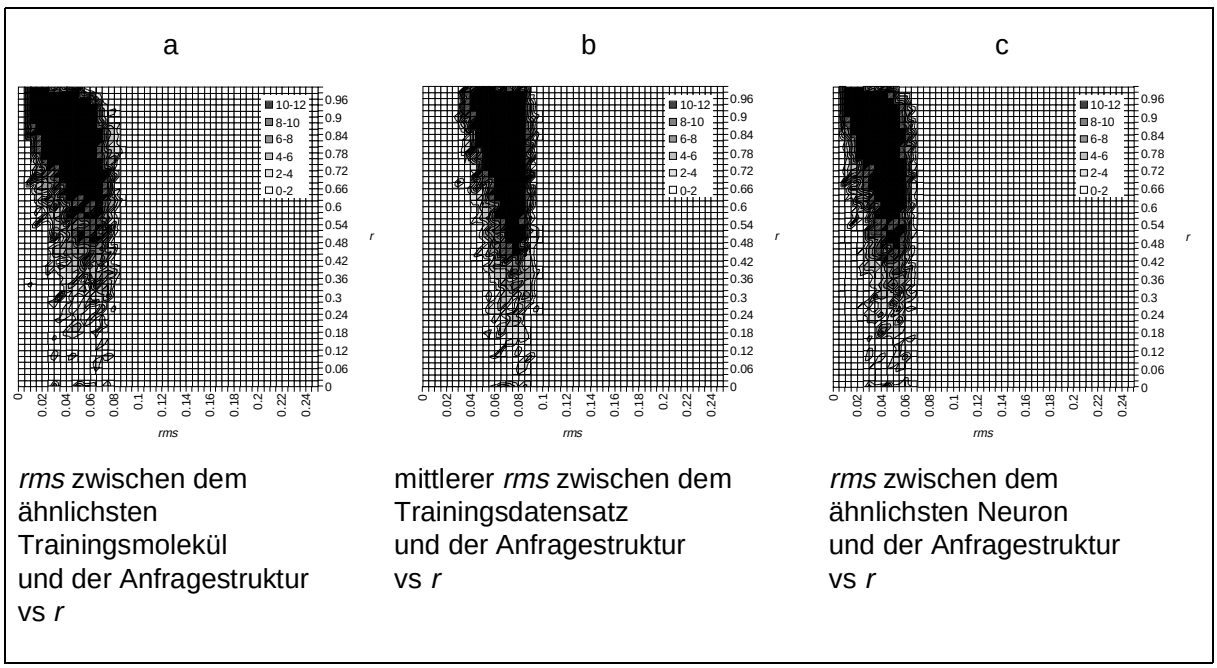


Abb. 2-59: Auftragung der *rms*-Werte (zwischen Strukturcodes) vs Korrelationskoeffizient *r* (zwischen Spektren). Die Grautöne stellen die *z*-Achse dar und beschreiben die Häufigkeit mit der bestimmte Kombinationen aus den jeweiligen *rms*- und *r*-Werten auftreten.

Bei allen drei Experimenten ist eine Häufung der Punkte im Bereich geringer *rms*-Werte und hoher Korrelationskoeffizienten zu beobachten (linker oberer Bereich der Graphiken). Dies bedeutet, daß bei sehr geringen *rms*-Werten, also sehr ähnlichen Strukturen, auch sehr gute Simulationsergebnisse zu erwarten sind. Das ist als sehr positiv zu bewerten, da es darauf hinweist, daß eine große Strukturcodeähnlichkeit mit einer hohen Ähnlichkeit der entsprechenden Spektren einhergeht. Von großem Interesse sind jedoch auch die Bereiche abnehmender struktureller Ähnlichkeit. Würde bei abnehmender Ähnlichkeit der Strukturen, also einem steigenden *rms*-Wert, auch die Simulationsqualität sinken, so wäre eine einfache Vorhersage der Simulationsqualität ausgehend von den Strukturcodeähnlichkeiten möglich. So ideal und einfach wie eben beschrieben ist es in keinem der drei Experimente der Fall. So kommt es für höhere *rms*-Werte zu der Situation, daß die Ähnlichkeit der vorhergesagten Spektren in einen *r*-Bereich von 0.30 bis 0.98 fällt. Die Simulationsergebnisse werden somit bei steigendem *rms*-

Wert nicht zwingendermaßen schlechter. Die Höhe des jeweiligen Korrelationskoeffizienten *r* ist jedoch zufällig. Als ungünstig ist das Verhalten zu beurteilen, wenn aufgrund eines niedrigen *rms*-Werts eine hohe Strukturähnlichkeit angezeigt wird, das vorhergesagte Infrarotspektrum jedoch sehr unähnlich ist (linker unterer Bereich des Graphen). Beim Vergleich der Experimente a-c (vgl. Abb. 2-59) fällt auf, daß dieser Bereich bei Experiment a die geringste Dichte an Punkten enthält. In nachfolgender Graphik ist dies noch deutlicher zu erkennen, da hier der Wertebereich der *z*-Skala, die quasi senkrecht auf der Blattebene steht und durch die verschiedenen Grautöne ausgedrückt wird, höher gewählt wurde:

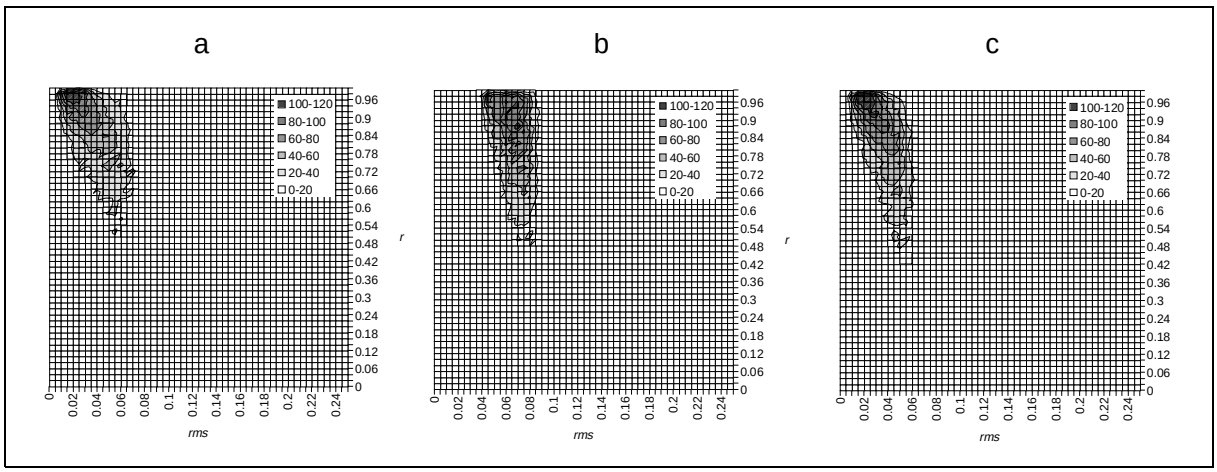


Abb. 2-60: Auftragung der *rms*-Werte (zwischen Strukturcodes) vs Korrelationskoeffizient *r* (zwischen Spektren). Die Grautöne stellen die *z*-Achse dar und beschreiben die Häufigkeit mit der bestimmte Kombinationen aus den jeweiligen *rms*- und *r*-Werten auftreten. Der Wertebereich der *z*-Achse wurde hier höher gewählt als bei Abbildung 2-60.

Der *rms*-Wert zwischen dem Strukturcode des Anfragemoleküls und dem ähnlichsten Molekül des Trainingsdatensatzes ermöglicht also in begrenztem Umfang eine Vorhersage der Simulationsqualität. Aus dem Verlauf bzw. der Lage der Graphen soll nun eine Art Grenzwert für die Mindestqualität der Simulation bestimmt werden. Die graphische Ermittlung dieses Grenzwerts erfolgt derart, daß für einen bestimmten Korrelationskoeffizienten *r* waagerecht eine Linie gezogen wird, bis diese den linken Rand des Graphen erreicht. Von diesem linken Rand wird eine senkrechte Linie nach unten gezogen, um den entsprechenden *rms*-Wert zu ermitteln. Welche *rms*-Werte vorliegen müssen, damit Simulationen mit $r > 0.9$, $r > 0.8$, $r > 0.7$ sowie $r > 0.6$ zu erwarten sind, soll nun anhand der Graphik für Experiment a (vgl. Abb. 2-61) ermittelt werden.

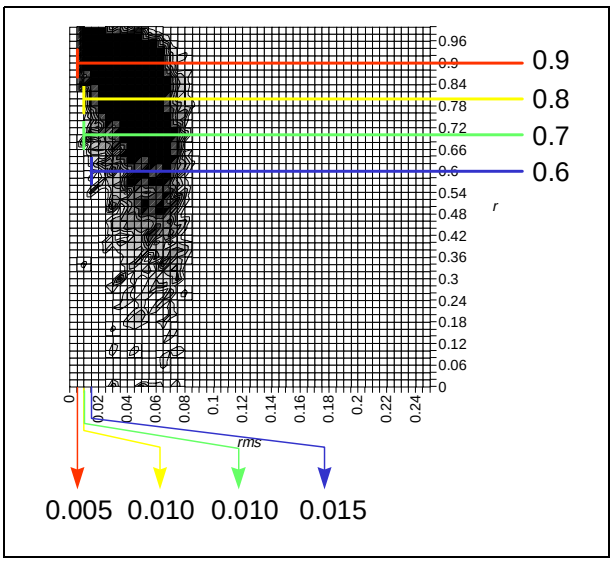


Abb. 2-61: Graphisch ermittelte *rms*-Grenzwerte für bestimmte Vorhersagequalitäten

Für die vier Werte des Korrelationskoeffizienten *r* wurden anhand der Graphik vier *rms*-Werte bestimmt:

$r \geq 0.9 \rightarrow rms \leq 0.005$

($r \geq 0.8 \rightarrow rms \leq 0.010$) nicht sinnvoll, da Grenzlinie des Graphen hier senkrecht verläuft

$r \geq 0.7 \rightarrow rms \leq 0.010$

$r \geq 0.6 \rightarrow rms \leq 0.015$

Liegt nun beispielsweise zwischen Anfragestrukturcode und dem Strukturcode des ähnlichsten Moleküls des Testdatensatzes ein *rms*-Wert von 0.005 vor, so wird aufgrund des vorhandenen Datenmaterials der Korrelationskoeffizient *r* zwischen simuliertem und experimentellem Spektrum zwischen 0.9 und 1 liegen. Der Grenzwert für $r \geq 0.8$ ist jedoch nicht sinnvoll, da die Begrenzungslinie des Graphen in diesem Bereich parallel zur *r*-Achse verläuft und sich aus diesem Grund für $r \geq 0.7$ scheinbar ebenfalls ein Grenzwert von $rms \leq 0.010$ ergibt. Wie die Beispiele in den folgenden Kapiteln zeigen werden, sind diese Werte jedoch als untere Grenzwerte für eine Mindestqualität der Simulation zu sehen. In den meisten Fällen wird ein besseres Simulationsergebnis erzielt werden, als es anhand obiger Werte abgeschätzt worden wäre.

2.6 Test der Methode anhand eines repräsentativen Datensatzes

Ziel dieser Untersuchung war es das Vorhersageverhalten der Methode anhand eines Datensatzes repräsentativer Moleküle zu untersuchen. Die Moleküle des Testdatensatzes wurden von Herrn Dipl.-Chem. Thomas aus dem Arbeitskreis von Prof. Salzer am Institut für Analytische Chemie der TU Dresden unter dem Gesichtspunkt ausgewählt, daß sie einen möglichst großen Bereich an struktureller Vielfalt in der organischen Chemie abdecken sollten. Zusätzlich sollten die Verbindungen eindeutig einer Substanzklasse zuzuordnen sein und keine zu komplexe Struktur aufweisen. Die mit dresden_1 bis dresden_16 bezeichneten Verbindungen sind in nachfolgender Abbildung dargestellt:

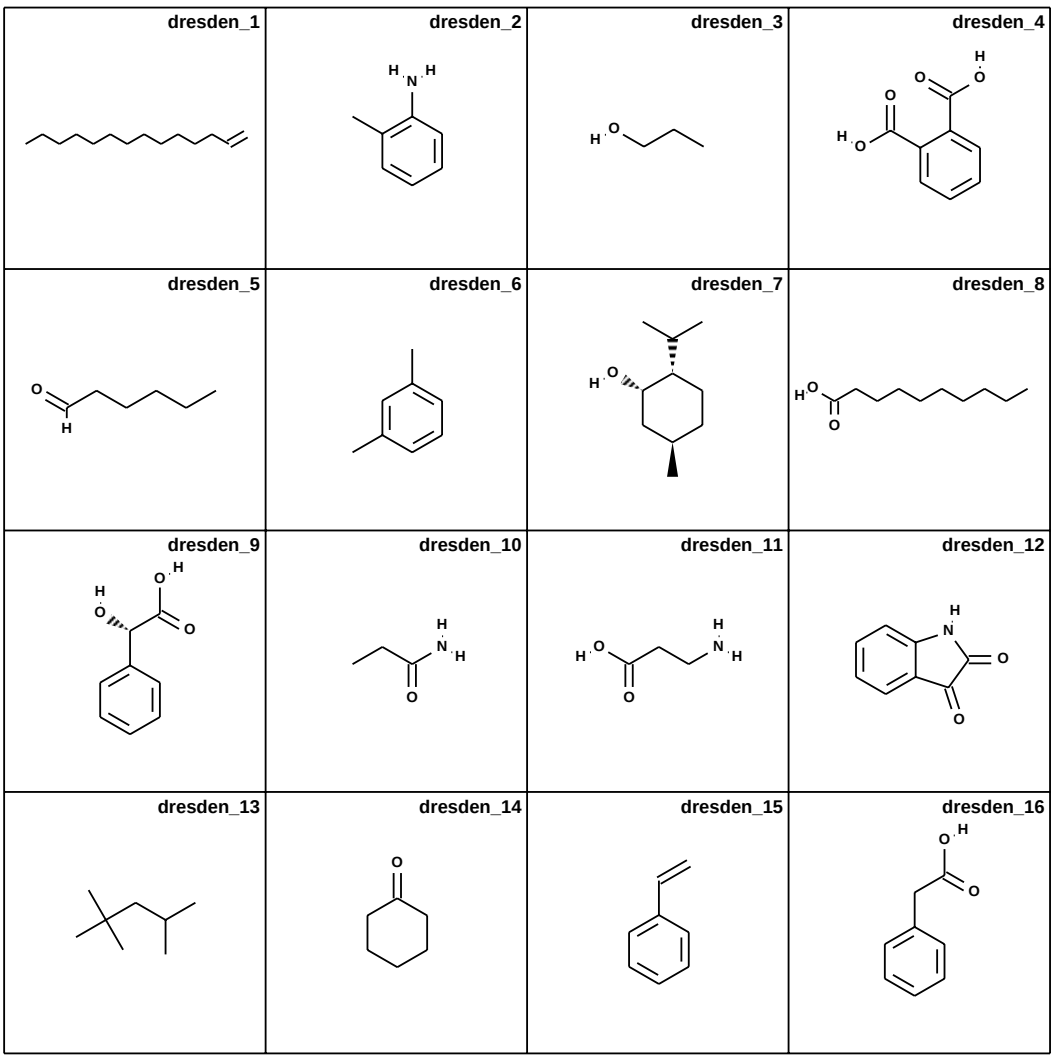


Abb. 2-62: Ausgewählte Testmoleküle

Für die ausgewählten Moleküle wurden die Infrarotspektren vorhergesagt. Die Simulationsparameter sind in nachfolgender Tabelle aufgeführt:

Tab. 2-23: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfragestrukturorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo gesamt
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Die Verbindungen 1, 2, 3, 6, 7, 8, 9, 13, 14, 15 und 16 sind in der SpecInfo IR-Datenbank enthalten. Bei der Auswahl der Trainingsdatensätze wurden jedoch die Datenbankeinträge nicht in die jeweiligen Trainingsdatensätze aufgenommen. Zur Bestimmung der Simulationsqualität wurden die simulierten Spektren mit den experimentellen Spektren verglichen. Als Vergleichsmaß wurde der Korrelationskoeffizienten *r* berechnet. Ebenso wie die Auswahl der Verbindungen wurden die experimentellen Spektren ebenfalls im Arbeitskreis von Prof. Salzer aufgenommen. Die Spektren wurden mit einem Nicolet 5PC Spektrometer mit einer Auflösung von 4 cm^{-1} aufgezeichnet, wobei die Substanzen mittels GC-MS auf ihre Reinheit überprüft wurden. Die Aufnahme fester Proben erfolgte als KBr-Preßling, flüssige Proben als Film zwischen KBr-Scheiben und wachsartige Substanzen als Schmelzfilm.

Die Ergebnisse des Vergleichs von simulierten und experimentellen Spektren sind in Abbildung 2-63 abgebildet. Eine Sammlung aller simulierten Spektren befinden sich in Anhang A.3.

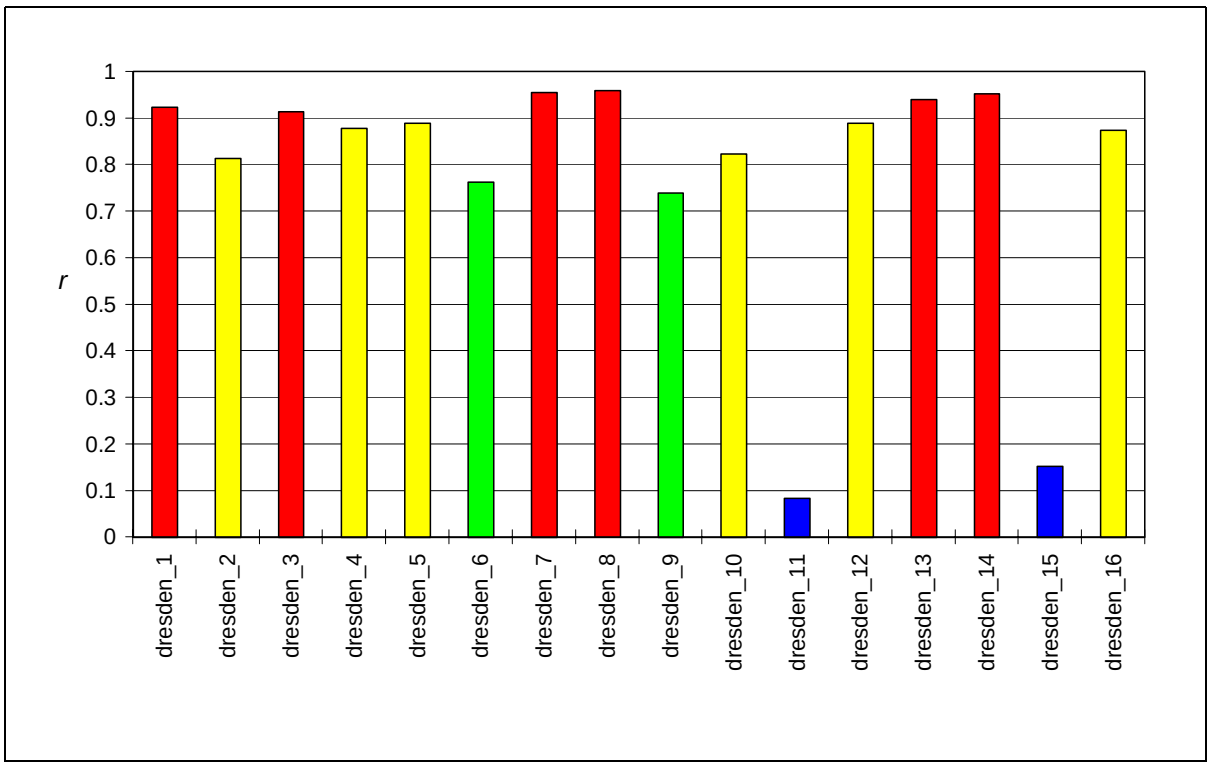


Abb. 2-63: Darstellung der Simulationsqualitäten

Nahezu alle Simulationen sind von guter Qualität ($r > 0.7$), einige sind sogar von sehr guter Qualität ($r > 0.9$). Desweiteren fällt auf, daß die Simulationen für die Verbindungen 11 und 15 von sehr schlechter Qualität sind ($r_{11} = 0.083$ und $r_{15} = 0.153$). Für die verschiedenen Experimente wurde untersucht, wie groß die Ähnlichkeit zwischen dem Anfragestrukturcode und dem Strukturcode des ähnlichsten Moleküls im Testdatensatz war (rms_{min}). Die Ergebnisse sind in nachfolgender Graphik dargestellt, wobei die Höhe der Balken die Ähnlichkeiten der Strukturcodes beschreibt. Die Farben zeigen die Güte der Simulation an.

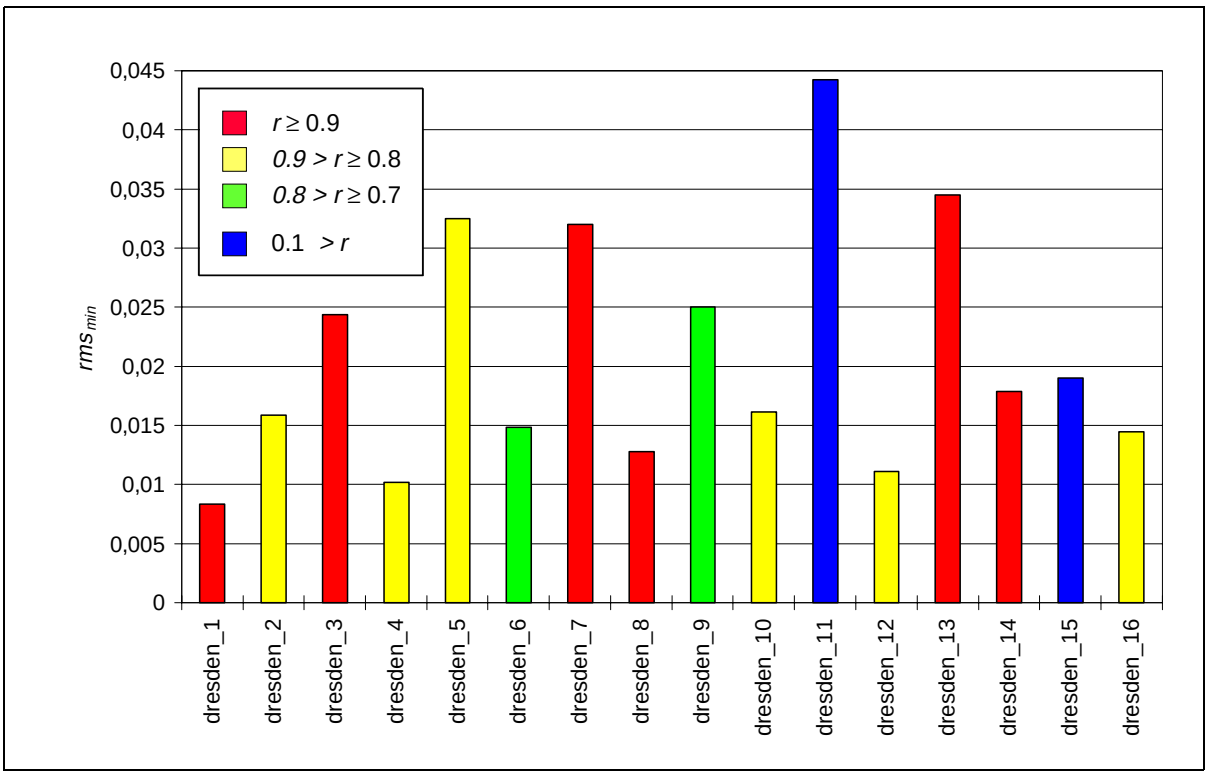


Abb. 2-64: Auftragung des *rms*-Wertes zwischen dem Anfragestrukturcode und dem ähnlichsten Molekül des Trainingsdatensatzes gegen den Korrelationskoeffizienten zwischen simuliertem und experimentellem Spektrum

Bei der Auswertung der Graphik in Abbildung 2-64 ist zunächst festzustellen, daß die in Kapitel 2.5 ermittelten Werte für die Vorhersage der Simulationsgüte anhand der *rms*-Werte zwischen dem Anfragestrukturcode und dem ähnlichsten Molekül des Trainingsdatensatzes als untere Grenzwerte gesehen werden müssen. In den meisten Fällen sind auch bei geringeren Ähnlichkeiten wesentlich bessere Simulationsgüten zu erwarten. Es zeigt sich, daß zehn Verbindungen (1, 4, 12, 8, 16, 6, 2, 10, 14, 15) relativ hohe Ähnlichkeiten mit dem ähnlichsten Molekül des Trainingsdatensatzes aufweisen und so den Schluß zulassen könnten, daß für diese Verbindungen gute Simulationen zu erwarten sind.

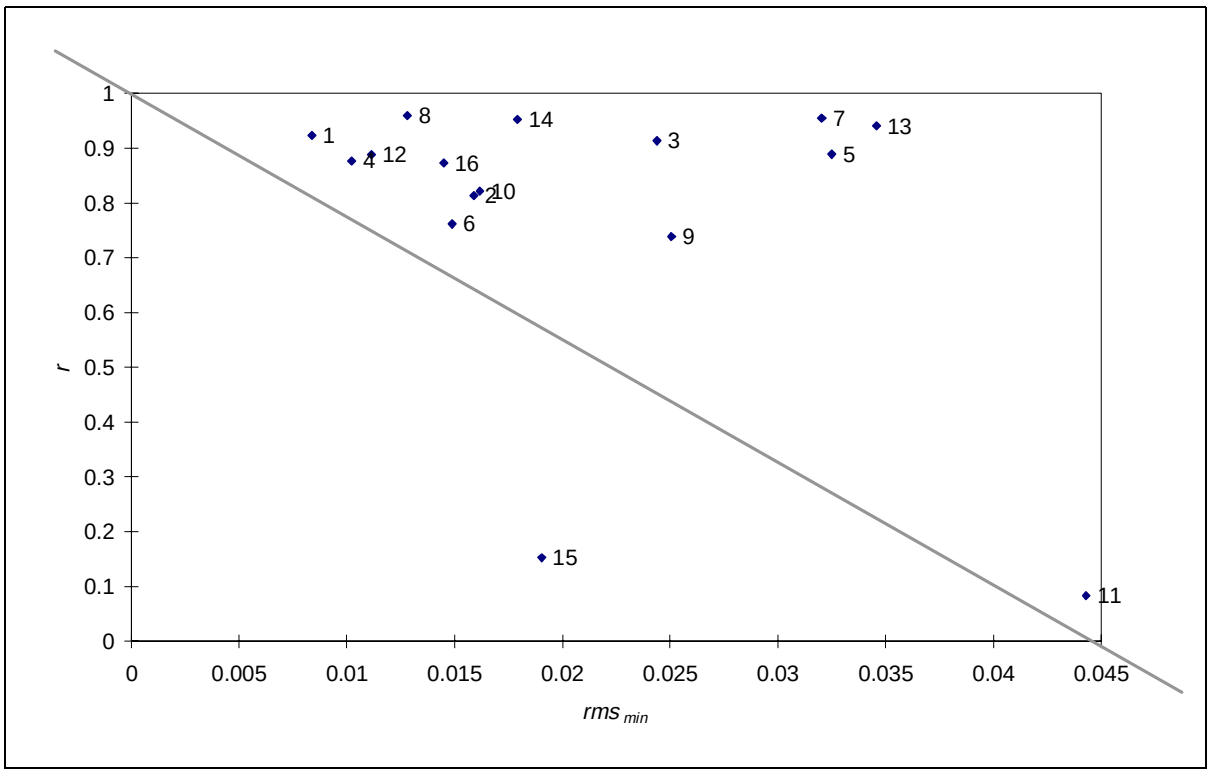


Abb. 2-65: Auftragung des *rms*-Wertes zwischen dem Anfragestrukturcode und dem ähnlichsten Molekül des Trainingsdatensatzes gegen den Korrelationskoeffizienten zwischen simuliertem und experimentellem Spektrum

Es fällt auf, daß die Punkte für die Verbindungen 11 und 15 deutlich abseits der Punkte für die übrigen Verbindungen liegen. In diesem Zusammenhang ist zu untersuchen, ob sich die schlechten Simulationsergebnisse für die Verbindungen 11 und 15 aufgrund dieser geringen Ähnlichkeiten, zumindest in Relation zu den anderen Verbindungen des Testdatensatzes, vorhersagen hätten lassen, d.h., ob eine geringe Ähnlichkeit zwischen dem Strukturcode des Anfragemoleküls und dem ähnlichsten Molekül des Testdatensatzes mit einem schlechten Simulationsergebnis einhergeht, wie es durch die diagonal gezogene Linie angedeutet werden soll. Dies ist für Verbindung 11 (β-Alanin) der Fall: Der *rms*-Wert von 0.044 zwischen dem Anfragestrukturcode und dem ähnlichsten Molekül des Testdatensatzes ist der höchste, der in dem gesamten Experiment aufgetreten ist. Der entsprechende Korrelationskoeffizient *r* zwischen simuliertem und experimentellem Spektrum beträgt 0.083 und ist damit der niedrigste des gesamten Experiments.

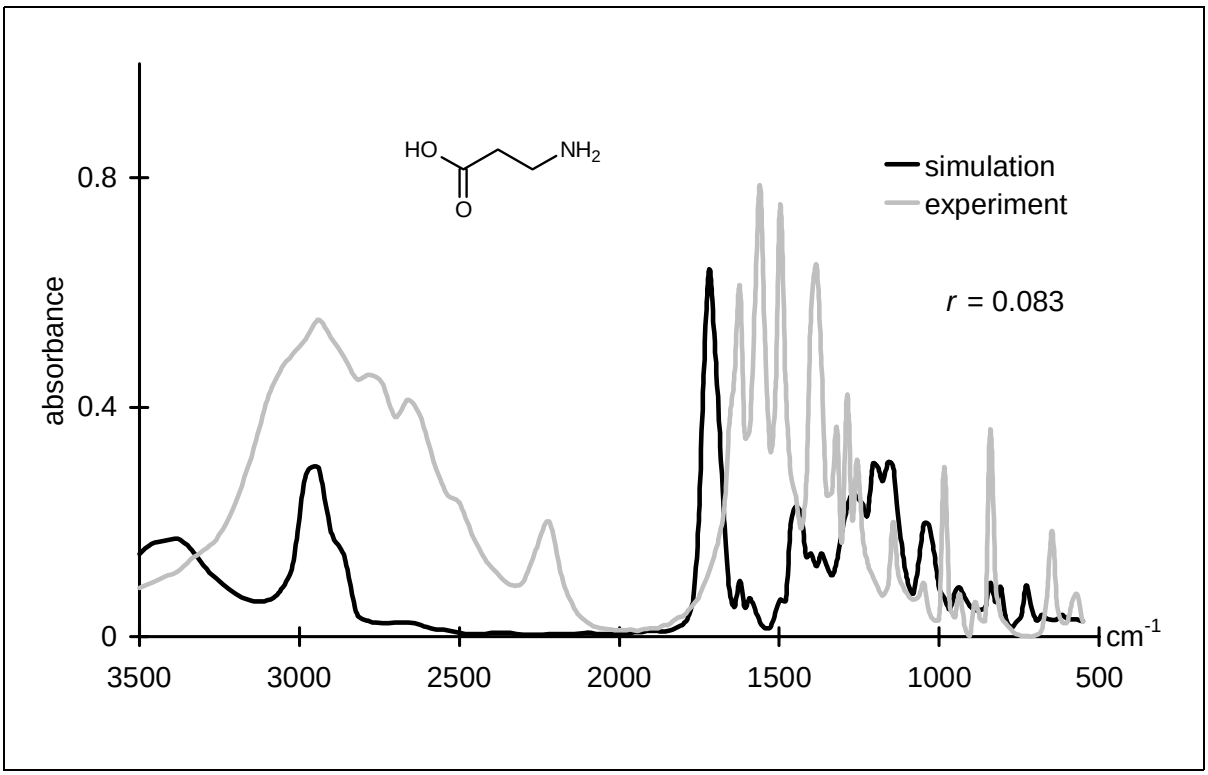


Abb. 2-66: Simulationsergebnis für Verbindung 11, β-Alanin
(schlechtestes Ergebnis des gesamten Experiments)

Tatsächlich sind experimentelles und simuliertes Spektrum sehr unterschiedlich. Dies fällt besonders bei der Lage der Carbonylbande im simulierten Spektrum bei 1700 cm^{-1} auf. Dies ist die typische Lage für Carbonylbanden wie sie bei Aldehyden, Ketonen, Carbonsäuren, Estern und Amiden zu finden sind. Aminosäuren sind diesbezüglich jedoch ein Sonderfall. In der zwitterionischen Form, wie β-Alanin sicherlich überwiegend vorliegen wird, ist die entsprechende Carbonylbande bathochrom verschoben (1605 - 1555 cm^{-1}).

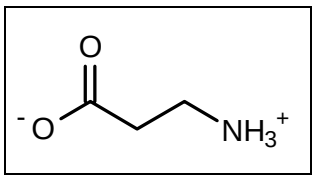


Abb. 2-67: β-Alanin in der zwitterionischen Form

Ein Aminosäurenspektrum läßt sich somit nicht aus den oben aufgezählten Verbindungsklassen interpolieren. In der ersten Nachbarschaftssphäre (vgl. Abb. 2-68) des Gewinnerneurons sind jedoch nur solche Verbindungen zu finden, was die schlechte Qualität der Simulation erklärt.

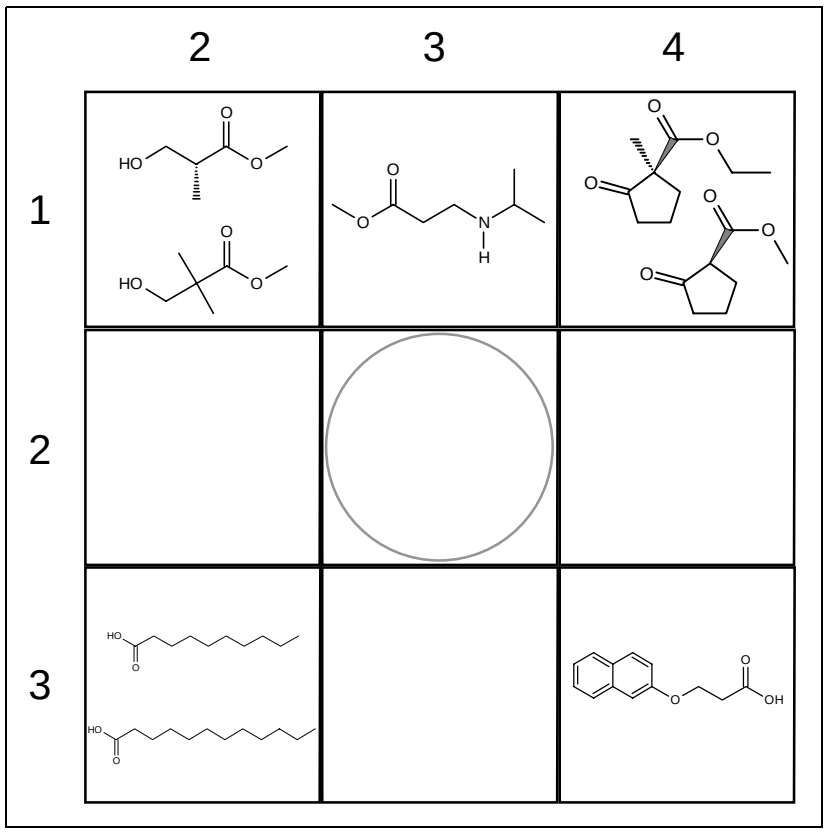


Abb. 2-68: Ausschnitt des neuronalen Netzes mit der Zuordnung der Trainingsmoleküle (β-Alanin). Das Gewinnerneuron ist mit einem Kreis markiert.

In einem weiteren Experiment sollte untersucht werden, inwieweit sich das Simulationsergebnis ändert, wenn die zwitterionische Form des β-Alanins eingegeben (vgl. Abb. 2-67) wird. Der Korrelationskoeffizient *r* zwischen simuliertem und experimentellem Spektrum fällt mit $r = 0.444$ zwar immer noch sehr niedrig aus, ist jedoch deutlich höher als beim ersten Experiment ($r = 0.083$). Auch beim Vergleich zwischen simuliertem und experimentellem Spektrum sind bereits mehr Übereinstimmungen zu beobachten als im ersten Fall.

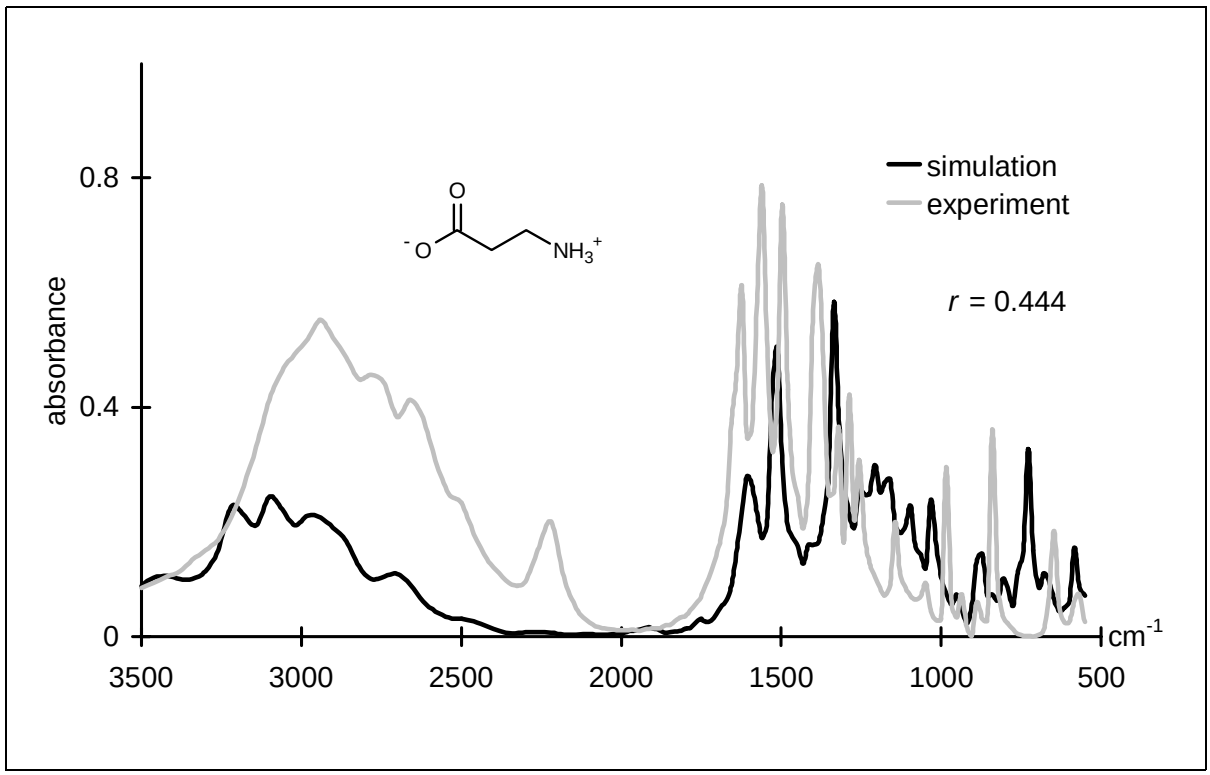


Abb. 2-69: Simulationsergebnis für Verbindung 11, β-Alanin in der zwitterionischen Form

Besonders das Bandenmuster um 3000 cm^{-1} läßt darauf schließen, daß bei diesem Experiment Aminosäuren im Trainingsdatensatz enthalten waren. Eine Analyse des neuronalen Netzes bestätigt diese Vermutung:

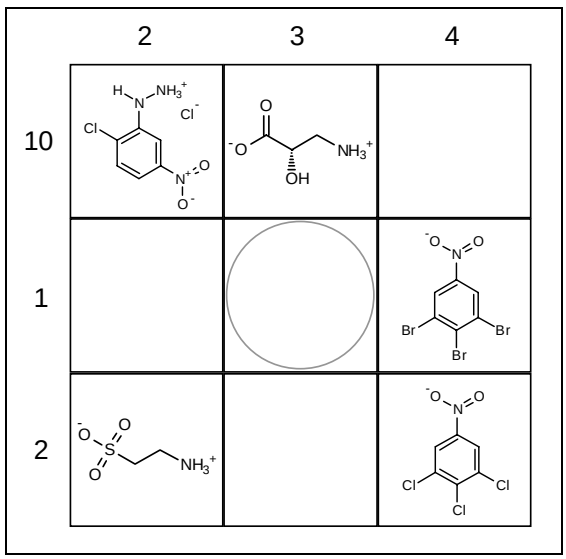


Abb. 2-70: Ausschnitt des neuronalen Netzes mit der Zuordnung der Trainingsmoleküle (β-Alanin in der zwitterionischen Form). Das Gewinnerneuron ist mit einem Kreis markiert.

Durch die Eingabe des β-Alanins in der zwitterionischen Form konnte eine Verbesserung des Simulationsergebnisses erreicht werden. Dies zeigt, wie wichtig es für ein Simulationsergebnis ist, daß die Anfragestruktur in der richtigen Form eingegeben wird. Das bedeutet daß es notwendig ist, die Struktur in einer chemisch sinnvollen Form einzugeben, bzw. in der Form in der vergleichbare Strukturen in der Datenbank zu finden sind. Beispielsweise konnte bei anderen Experimenten beobachtet werden, daß Nitroverbindungen in der Datenbank ebenfalls in der zwitterionischen Form vorliegen. Wird eine Anfragestruktur mit einem Nitrosubstituenten in der Neutralform eingegeben, so werden wenige oder keine ähnlichen Nitrosubstituierten Moleküle in den Trainingsdatensatz aufgenommen.

Bei Verbindung 15, dem Styrol, ist die Ausgangslage zunächst etwas anders als bei der ersten β-Alanin-Simulation. Der rms_{min}-Wert von 0.019 zwischen dem Anfragestrukturcode und dem Strukturcode des ähnlichsten Moleküls des Trainingsdatensatzes deutet eine hohe strukturelle Ähnlichkeit an und läßt so ein gutes Simulationsergebnis erwarten. Das Simulationsergebnis ist jedoch das zweitschlechteste. Beim Vergleich von simuliertem und experimentellem Spektrum fallen sofort zwei Bereiche mit deutlichen Abweichungen auf. Zunächst ist im simulierten Spektrum unterhalb von 3000 cm^{-1} eine Bande zu finden, die eindeutig aliphatischen ν-CH Schwingungen zuzuordnen ist. Eine solche Bande fehlt natürlich im experimentellen Spektrum. Desweiteren fallen Abweichungen im Fingerprintbereich auf (vgl. Abb. 2-71).

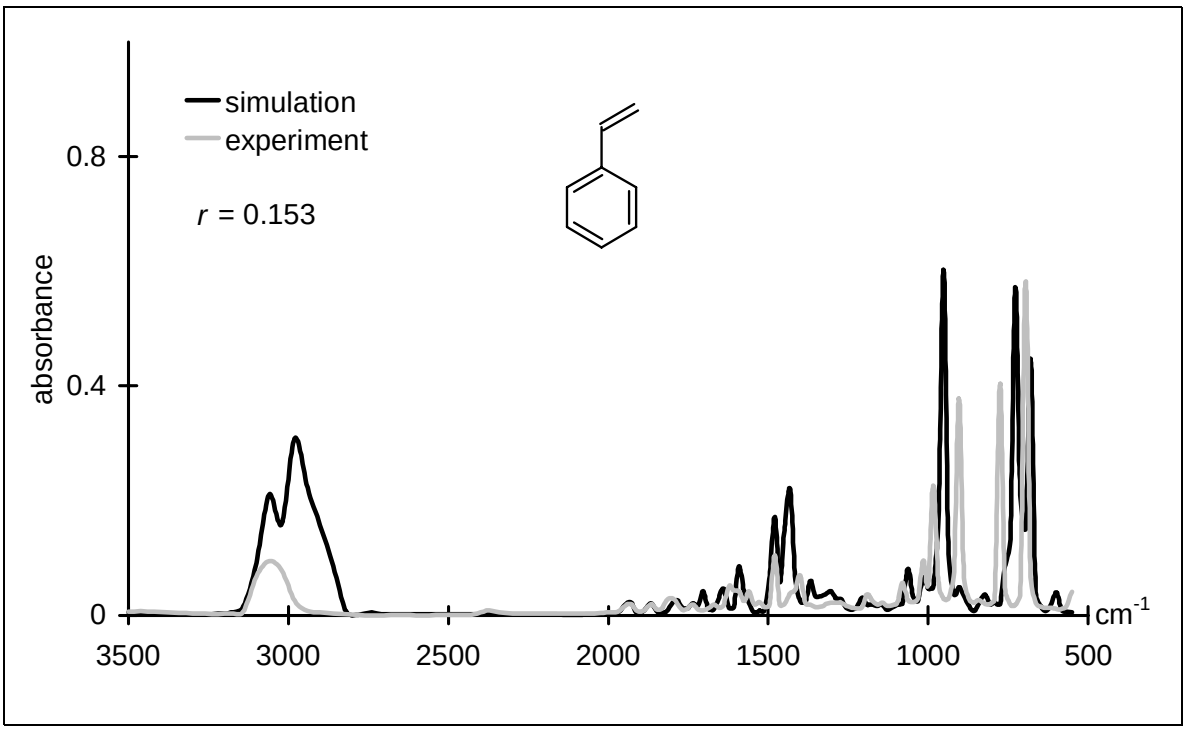


Abb. 2-71: Simulationsergebnis für Verbindung 11, Styrol
(zweitschlechtestes Ergebnis des gesamten Experiments)

Zur genaueren Analyse des Simulationsergebnisses soll auch hier die Zuordnung der Trainingsmoleküle im neuronalen Netz näher betrachtet werden.

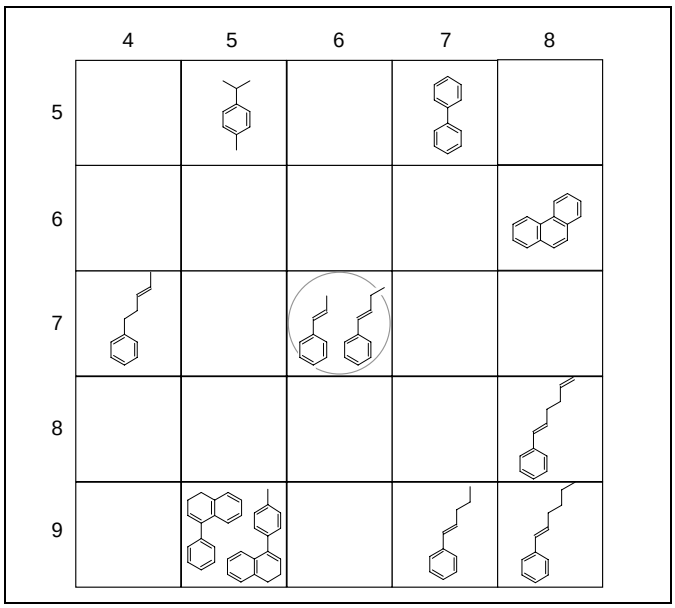


Abb. 2-72: Ausschnitt des neuronalen Netzes mit der Zuordnung der Trainingsmoleküle (Styrol). Das Gewinnerneuron ist mit einem Kreis markiert.

Auch hier ist das Zustandekommen der Abweichungen sehr leicht nachzuvollziehen. Das Gewinnerneuron, aus dem das simulierte Spektrum stammt, wurde mit einem Kreis markiert. Diesem Neuron sind im Training zwei Moleküle zugeordnet worden, nämlich trans-1-Phenyl-1-propen und trans-1-Phenyl-1-buten. Beide Moleküle weisen strukturell gesehen eine sehr große Ähnlichkeit mit dem Anfragemolekül auf. Aus IR-spektroskopischer Sicht ist die Ähnlichkeit deutlich geringer, da die Trainingsmoleküle aliphatische Merkmale in die Simulation miteinbringen, die in der Anfragestruktur nicht vorhanden sind.

Als letztes Beispiel soll das Simulationsexperiment mit der höchsten Übereinstimmung zwischen simuliertem und experimentellem Spektrum näher untersucht werden.

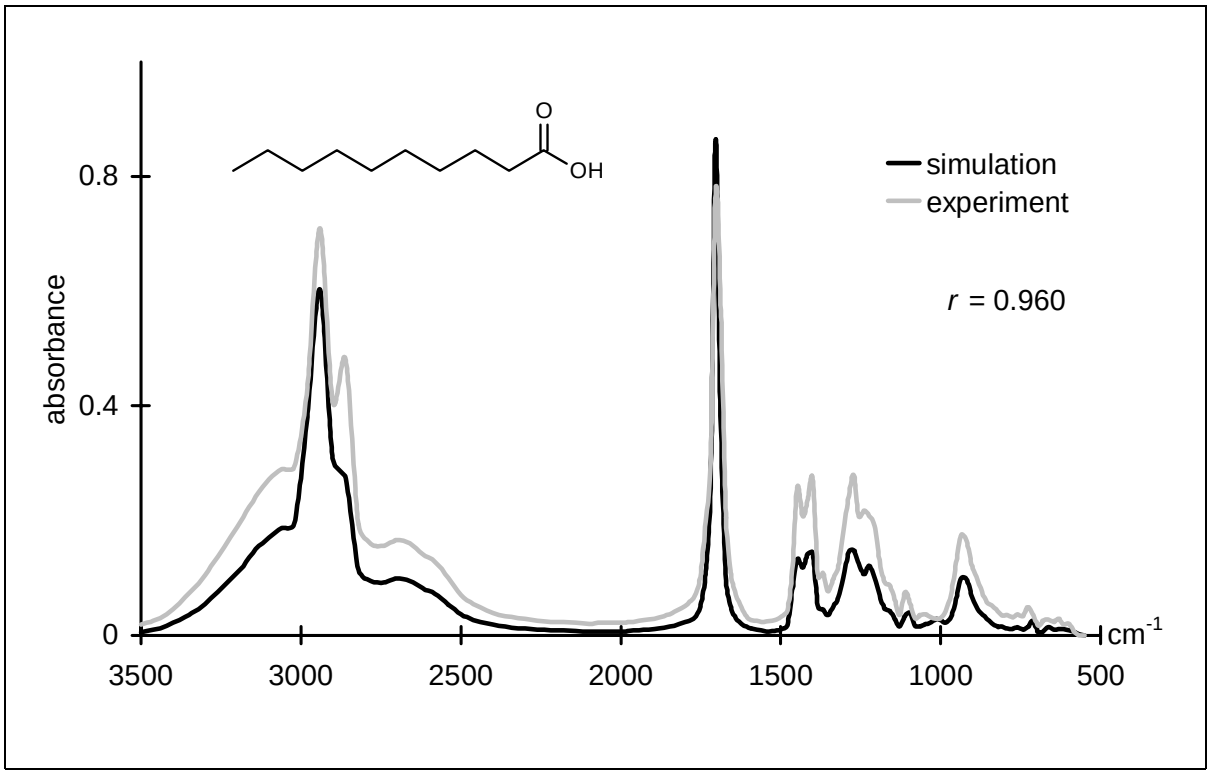


Abb. 2-73: Simulationsergebnis für Verbindung 8, Decansäure
(bestes Ergebnis des gesamten Experiments)

Auch hier soll die Zuordnung der Trainingsmoleküle näher betrachtet werden.

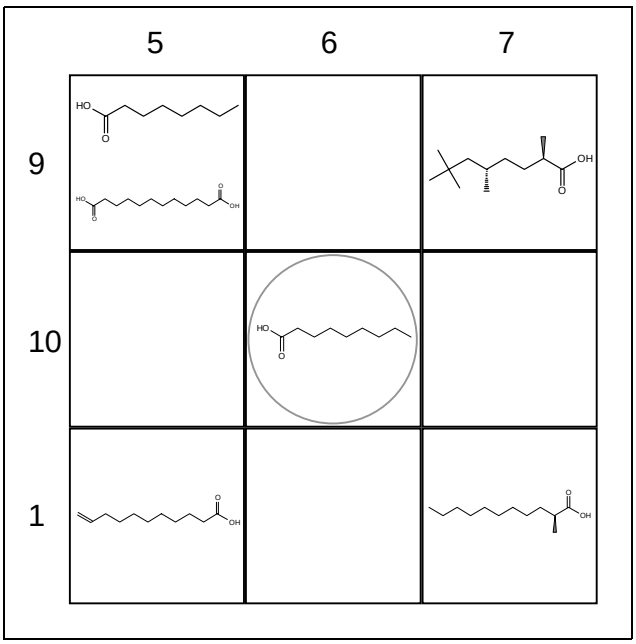


Abb. 2-74: Ausschnitt des neuronalen Netzes mit der Zuordnung der Trainingsmoleküle zur Simulation von Decansäure (vgl. Abb. 2-73)

Dieses Beispiel kann nahezu als Idealfall eines Netztrainings betrachtet werden. Die Nonansäure, die sowohl aus struktureller als auch aus IR-spektroskopischer Sicht eine sehr hohe Ähnlichkeit mit der Anfragestruktur aufweist, ist im Training jenem Neuron zugeordnet worden, welches im anschließenden Simulationsschritt als Gewinnerneuron bestimmt wurde.

Abschließend kann bemerkt werden, daß nahezu für alle Verbindungen des Testdatensatzes gute bis sehr gute Simulationsergebnisse erzielt werden konnten. Die beiden Moleküle, für welche sehr geringe Übereinstimmungen zwischen Simulation und Experiment zu beobachten waren, müssen bezüglich ihrer Struktur-Spektrenkorrelation als Sonderfälle angesehen werden. Eine Verbesserung der Spektrenvorhersagequalität für derartige Sonderfälle könnte möglicherweise über eine Erweiterung der Codierungsmethode, beispielsweise durch die Kombination mit einer Methode zur Erkennung spezieller Strukturmerkmale, erreicht werden (vgl. Kap. 5.1).

2.7 Vorhersage von Spektren ohne Datenreduktion

Bei allen bisherigen, im Rahmen dieser Arbeit vorgestellten Simulationen waren zum Training des neuronalen Netzes datenreduzierte Spektren mit 128 Stützstellen verwendet worden. Somit waren die simulierten Spektren ebenfalls datenreduziert. Die Datenreduktion wurde dabei durch eine Reduzierung der Koeffizienten bei einer Hadamardtransformation erreicht (vgl. Kap. 2.1.1.1). Bei den nachfolgenden Simulationsexperimenten wurden Vollspektren, wie sie direkt aus dem Spektrometer ausgegeben werden, zum Training verwendet. Als Basisdatensatz diente ein Datensatz mit 81 Molekülen, der freundlicherweise vom Arbeitskreis von Prof. Salzer am Institut für Analytische Chemie der TU Dresden zur Verfügung gestellt wurde. Die Moleküle sind von 1 bis 83 numeriert, wobei die Einträge No. 38 und No. 81 fehlen. Der Datensatz ist in Anhang A.4 dargestellt. Die Simulationsparameter sind in nachfolgender Tabelle aufgeführt:

Tab. 2-24: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Spektrenbeschreibung	1868 Absorbanzwerte von 4000 - 500 cm^{-1} Datenpunktabstand 2 cm^{-1}
Trainingsdatensatzauswahl	leave-one-out Verfahren Bei dieser Methode wird jeweils ein Molekül aus dem Datensatz entnommen. Für dieses Molekül wird die Spektrensimulation durchgeführt und mit den restlichen Molekülen wird das Netz trainiert.
Anzahl der Trainingsmoleküle	80
Datenbasis	Datensatz mit 81 Molekülen (vgl. Anhang A.4)
Neuronen	12 x 12
Netzwerkform	toroidal
Training	unüberwacht

Die Verteilung der Simulationsqualitäten ist in Abbildung 2-75 dargestellt. Die Abbildungen der Vergleiche von simulierten und experimentellen Spektren befinden sich in Anhang A.5.

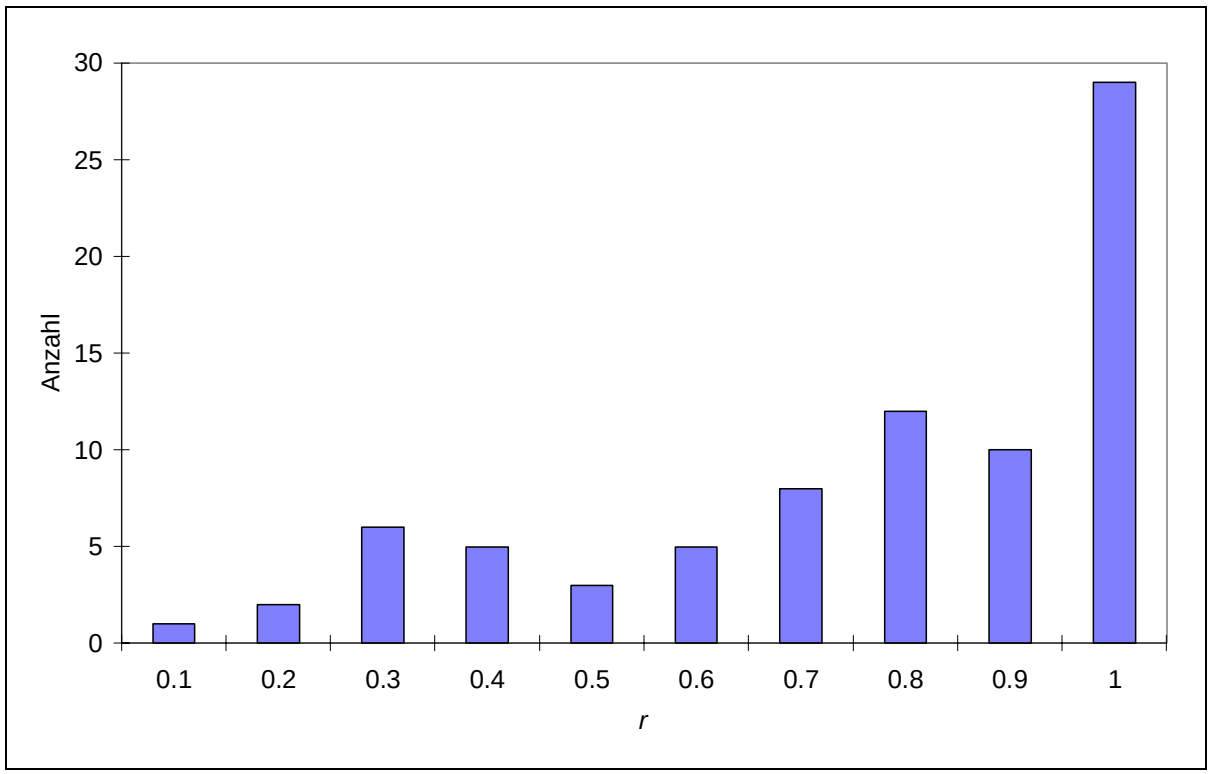


Abb. 2-75: Verteilung der Simulationsqualitäten

Es ist deutlich zu erkennen, daß das Intervall von $0.9 < r \leq 1$ mit 28 Simulationen (35%) am stärksten besetzt ist. 46 Simulationen erreichen einen Korrelationskoeffizient von $r > 0.7$. Dieses Experiment unterscheidet sich von den vorhergehenden dadurch, daß der zugrundeliegende Basisdatensatz wenig umfangreich, sehr heterogen und unausgewogen ist. Um zu untersuchen, welche Ähnlichkeiten zwischen den einzelnen Molekülen bestehen wurde der *rms*-Wert zwischen dem Strukturcode der Anfragestruktur und dem ähnlichsten Molekül des Trainingsdatensatzes bestimmt. Das Ergebnis ist in Abbildung Abbildung 2-76 dargestellt. Im Verlauf der Punkte sind immer wieder Sprünge zu beobachten. An solchen Stellen endet jeweils eine Reihe von Molekülen mit ähnlichen Strukturmerkmalen.

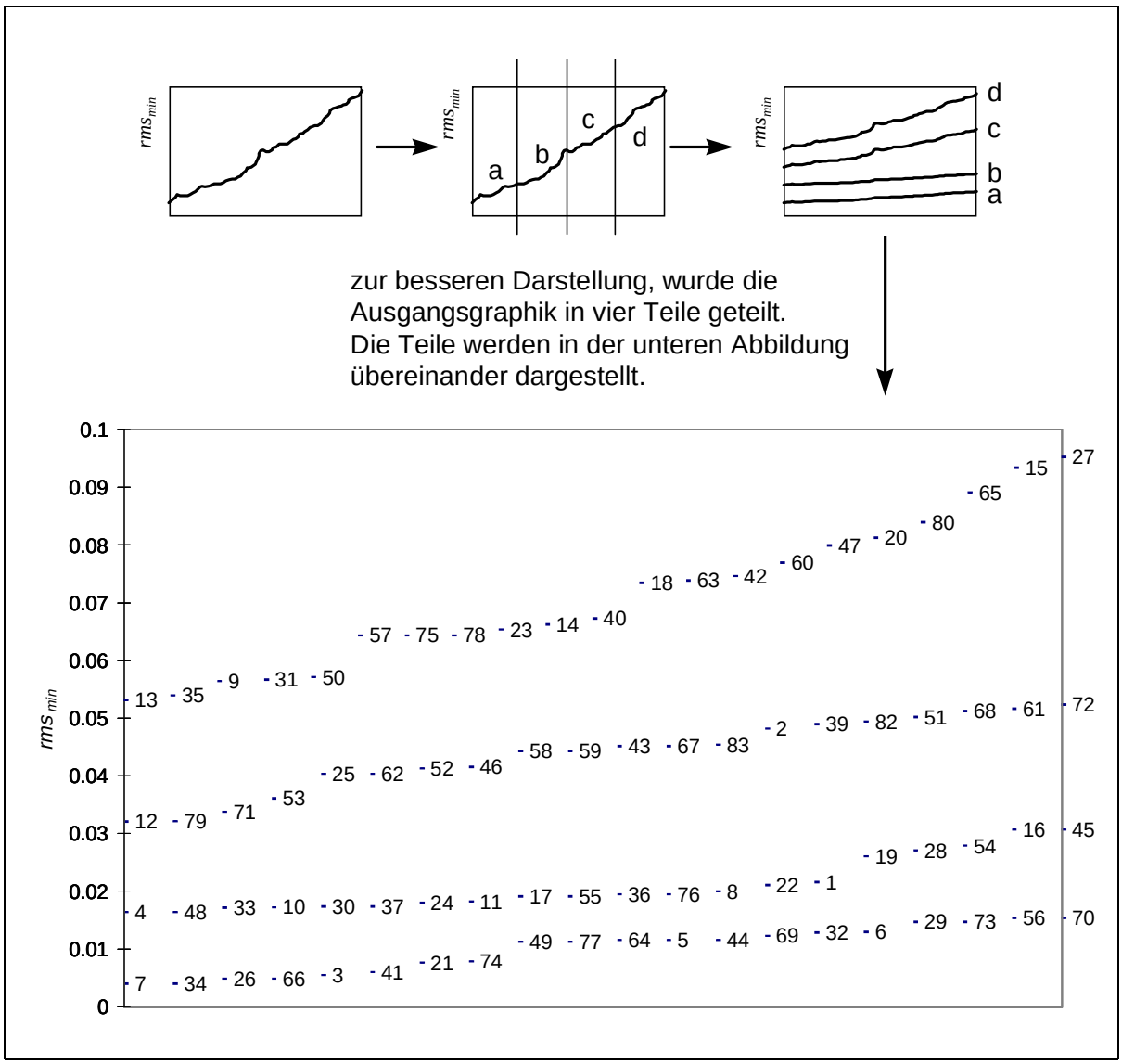


Abb. 2-76: *rms*-Werte zwischen dem Strukturcode der jeweiligen Anfragestruktur und dem ähnlichsten Molekül des Trainingsdatensatzes

In Abbildung 2-77 ist das Beispiel mit dem höchsten Korrelationskoeffizienten von r = 0.996 zwischen simuliertem und experimentellem Spektrum dargestellt. In den Bandenmustern, einschließlich der feinen Banden im Fingerprintbereich, ist praktisch keine Abweichung zwischen simuliertem und experimentellem Spektrum zu erkennen.

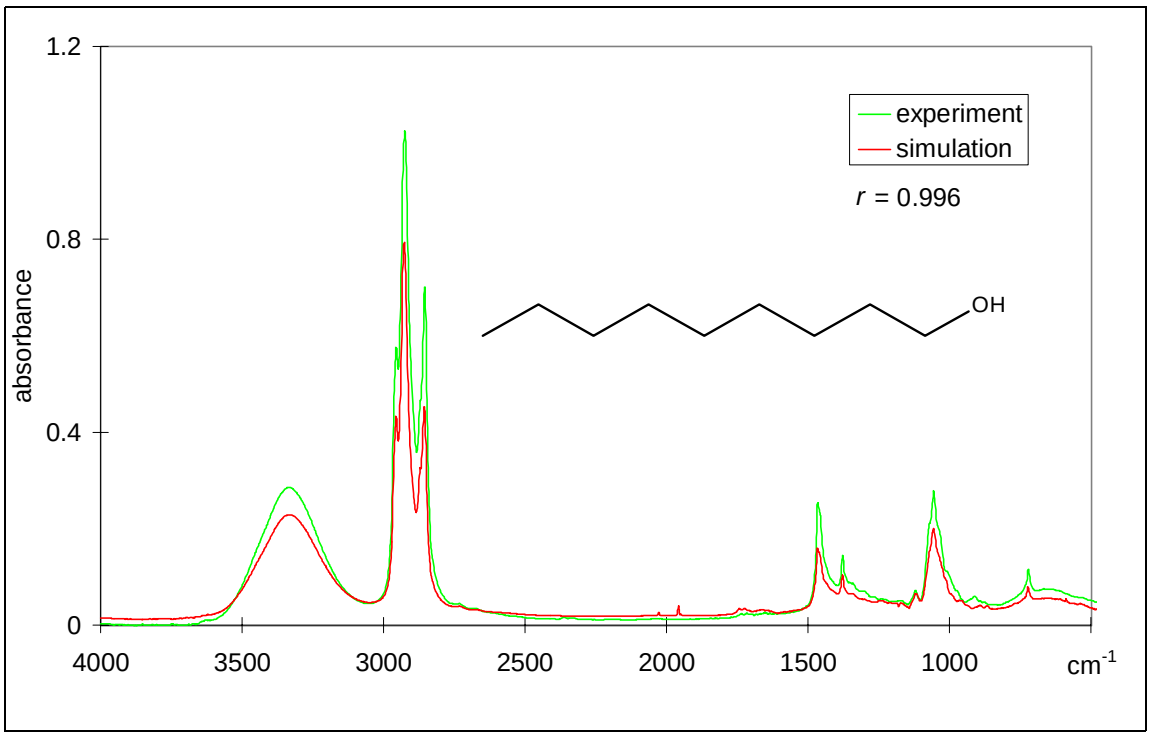


Abb. 2-77: Vergleich von simuliertem und experimentellem Spektrum für 1-Nonanol

Das gute Simulationsergebnis ist leicht nachzuvollziehen, da der Trainingsdatensatz eine Reihe sehr ähnlicher 1-Hydroxyalkane, unter anderem das 1-Undecanol, enthält.

Zusammenfassend kann bemerkt werden, daß auch Simulationen für nichtdatenreduzierte Spektren durchgeführt werden können. Für die Funktionsweise der Methode ist der Datenumfang praktisch unbedeutend. Das einzige auftretende Problem war computertechnischer Art. Die Vorbereitung und Auswertung der Daten wurde mit Tabellenkalkulationsprogrammen durchgeführt. Diese sind jedoch in aller Regel auf die Erfassung von maximal 256 Spalten beschränkt. Somit mußte mit verschiedenen Skriptprogrammen eine Umformatierung des Datenmaterials vorgenommen werden. Prinzipiell muß jedoch die Frage gestellt werden, inwiefern eine Simulation von nichtdatenreduzierte Spektren sinnvoll ist. Um die hohe Bandenvielfalt, wie sie besonders im Fingerprintbereich zu finden ist, qualitativ hochwertig vorherzusagen, muß der Datenraum zum Training des neuronalen Netzes sehr gut abgedeckt sein. Dies kann bei einer eng umrissenen Fragestellung in einem ausführlich spektroskopierten Substanzbereich der Fall sein. Mit dem Anspruch, möglichst für die gesamte organische Chemie Spektrenvorhersagen zu treffen, ist es bei der zur Verfügung stehenden Datenbasis sinnvoller mit datenreduzierten Spektren zu arbeiten. Im Falle einer gelungenen Simulation gibt das vorhergesagte Spektrum die Bandenmuster in Form einer umhüllenden Kurve wieder, wie z.B. bei der Anwendung in nachfolgendem Kapitel (vgl. Abb. 3-5).

3 Praktische Anwendungen der Spektrensimulation

In den nachfolgenden Kapiteln werden drei praktische Anwendungen der Spektrensimulation präsentiert. Auch wenn die thematischen Hintergründe völlig unterschiedlich sind, ist die vorliegende Situation in allen Fällen gleich: Eine oder mehrere Substanzen wurden IR-spektroskopisch analysiert. Eine einfache Identifikation der Verbindungen durch den Vergleich mit einem experimentellen Referenzspektrum ist jedoch nicht möglich, da die entsprechenden Verbindungen in keiner Datenbank und keinemSpektrenkatalog zu finden waren. Durch die Simulation eines IR-Spektrums für den oder die in Frage kommenden Kandidaten soll die vermessene Substanz identifiziert werden.

3.1 Spektrenvorhersage für N,N-Dimethylanilin-N-Oxid

Um die Ausscheidung von lipophilen Fremdstoffen aus einem Organismus zu erleichtern bzw. überhaupt erst zu ermöglichen, ist es notwendig, deren Hydrophylie durch Metabolisierungsreaktionen zu erhöhen.[57] Dazu findet in der Regel als erster Schritt eine Oxidation durch sogenannte Monooxigenasen statt.[58] Böcker et al. beschreibt Arbeiten zur Untersuchung des in vitro-Metabolismus der schwefelhaltigen Pestizide Methiocarb, Ametryn, Prometryn und Terbutryn.[59][60] Dabei galt es weiterhin eine Methode zu entwickeln, mit welcher die Beiträge der beiden Monooxigenasen Cytochrom P-450 und FMO an diesen Metabolismusprozessen bestimmt werden können.[61] In diesem Zusammenhang ist die Diplomarbeit von J. Hardt [62] zu sehen, bei der unter anderem die enzymatisch katalysierte Oxidation von N,N-Dimethylanilin zu N,N-Dimethylanilin-N-Oxid (DMA-NO) durchgeführt wurde. Mit Humanmikrosomen ist die Reaktion unabhängig von Cytochrom P-450 und kann somit als spezifisch für FMO angesehen werden.[63][64] Für diese Experimente wurde von J. Hardt DMA-NO synthetisiert,[65] als Hydrochlorid gefällt und IR-spektroskopisch vermessen.

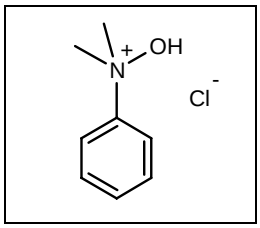


Abb. 3-1: N,N-Dimethylanilin-N-Oxid-Hydrochlorid

Das Spektrum zeigt die erwarteten Banden bei 3400 cm^{-1} (ν OH), 3010 cm^{-1} (ν CH aro-

matisch), 2700 cm^{-1} (ν CH aliphatisch), 1600 cm^{-1} (Ringdeformationsschwingungen) und 1400 cm^{-1} (δ CH_2). Dies ist zwar als positiv zu bewerten, allerdings gäbe es auch eine Vielzahl anderer Verbindungen, die gleiche Strukturmerkmale aufweisen und damit ähnliche Banden im Spektrum zeigen würden. Das zeigt auch, wie wenig aussagekräftig diese Form der Spektrenbeschreibung ist. Die wesentliche Information eines Infrarotspektrums ist nur aus seiner Gesamtform, den Bandenmustern und den relativen Intensitäten, zu gewinnen.

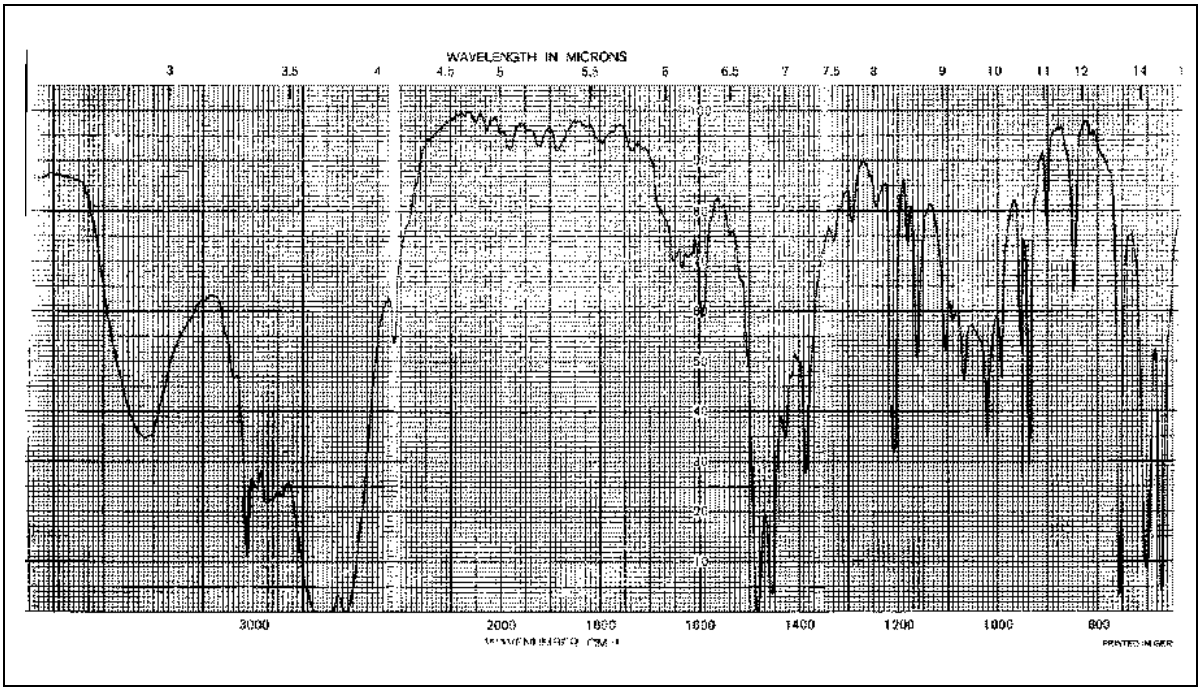


Abb. 3-2: Experimentelles Spektrum von N, N-Dimethylanilin-N-Oxid

Ein Referenzspektrum war weder in verschiedenen Spektrenkatalogen [66][67] noch in der SpecInfo [2] Infrarotdatenbank enthalten. Aus diesem Grund wurde mit den in den vorigen Kapiteln beschriebener Methode ein Infrarotspektrum simuliert, um zu überprüfen, inwieweit eine Korrespondenz bei der Lage der Banden und den Bandenmustern zu finden ist. Dazu wurde die Anfragestruktur zunächst mit den in Tabelle 3-1 aufgeführten Codierungsparametern codiert.

Tab. 3-1: Codierungsparameter

Parameter	Wert
Code	Radial
Anzahl der Codewerte	128
B	100 $Å^{-2}$
R_{max}	12.8 Å
Atomeigenschaft	q_{tot}

Tabelle 3-2 enthält die Parameter des Simulationsexperiments.

Tab. 3-2: Simulationsparameter

Parameter	Einstellung
Trainingsdatensatzauswahl	anfrageorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo, ungeladene H, C, N, O, Hal-Verbindungen (9850 Moleküle)
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Simuliertes und experimentelles Spektrum sind in nachfolgender Abbildung dargestellt. Zum besseren Vergleich wurde das experimentelle Spektrum, welches nur in Papierform vorlag, mittels des Programms UNSCAN-IT [68] digitalisiert. Das simulierte Spektrum wurde skaliert, um es in einen ähnlichen Wertebereich zu bringen wie das experimentelle Spektrum.

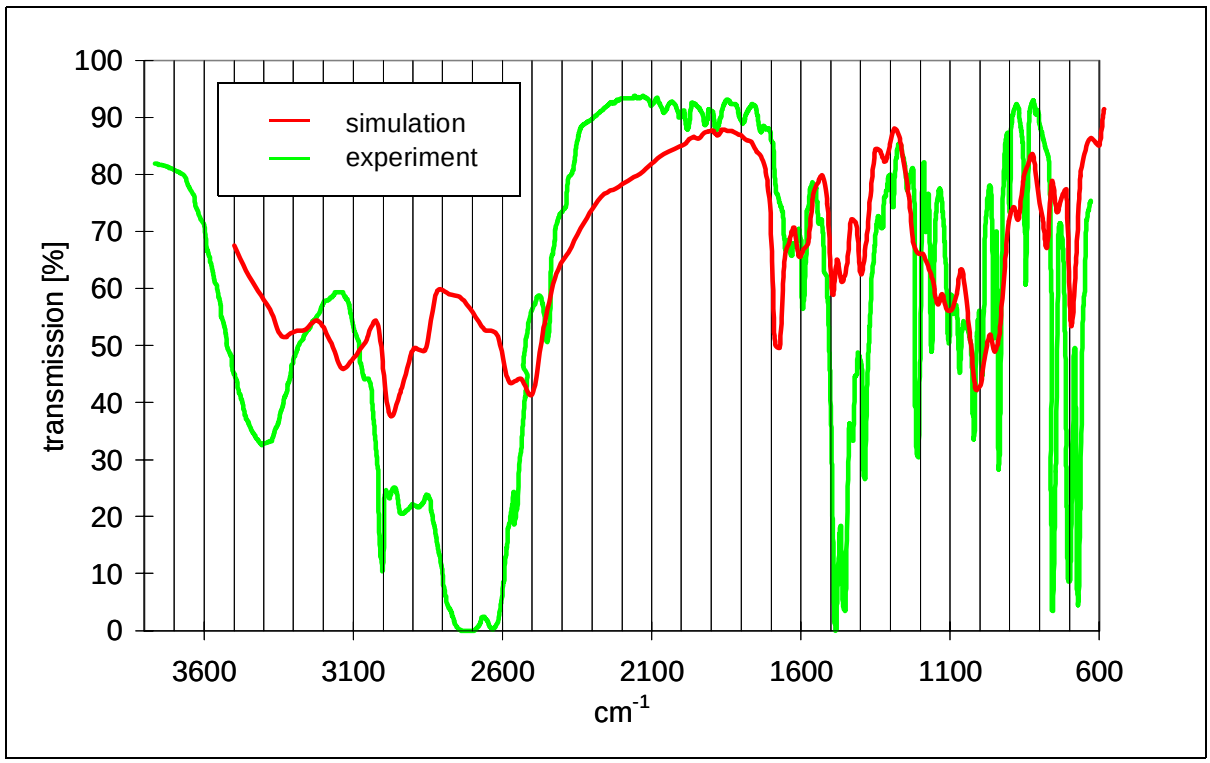


Abb. 3-3: Vergleich von simuliertem und experimentellem Spektrum

Beim Vergleich von simuliertem und experimentellem Spektrum sind sowohl Bereiche mit Übereinstimmungen als auch einige Abweichungen zu erkennen. Das simulierte Spektrum zeigt die entsprechenden Banden für die ν OH-Schwingung bei 3350 cm^{-1}, die aromatischen ν CH Schwingungen bei 3100 cm^{-1} sowie die aliphatischen ν CH Schwingungen bei 2950 - 2700 cm^{-1}. Im Vergleich zum experimentellen Spektrum sind hier jedoch einige Abweichungen zu beobachten. Zunächst unterscheiden sich die Bandenmuster von experimentellem und simuliertem Spektrum deutlich. Weiterhin hat die Bande der aliphatischen ν CH Schwingung des experimentellen Spektrums bei ca. 2700 cm^{-1} kein entsprechendes Gegenstück im simulierten Spektrum. Hier finden sich zwar Banden bei 2900 und 2500 cm^{-1}, wobei letztere eher der NH Schwingung bei protonierten Iminen oder Ammoniumverbindungen zuzuordnen ist. Dies läßt sich bei einer näheren Betrachtung der ausgewählten Trainingsmoleküle leicht nachvollziehen. Nachfolgende Abbildung zeigt die Moleküle, die den Neuronen der ersten Nachbarschaftssphäre des Gewinnerneurons zugeordnet wurden.

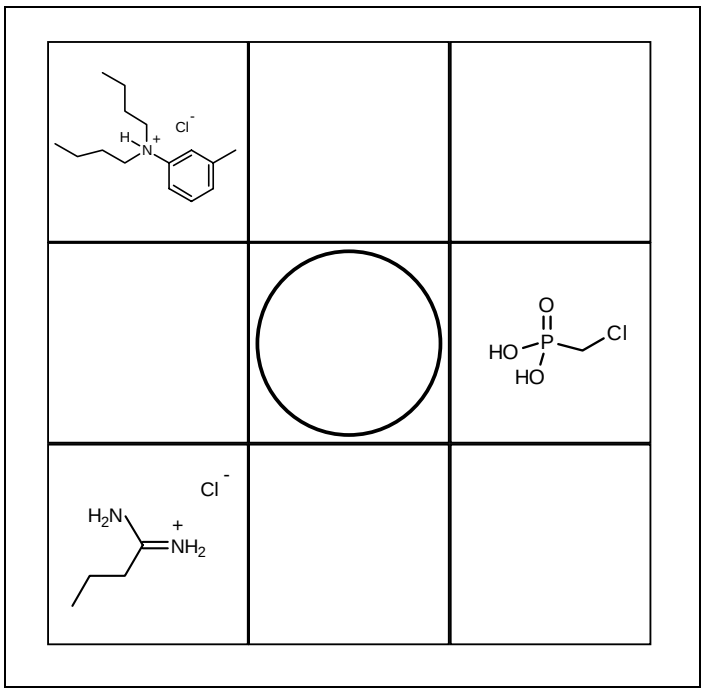


Abb. 3-4: Belegung der Neuronen der ersten Nachbarschaftssphäre des Gewinnerneurons. Das Gewinnerneuron ist mit einem Kreis markiert.

In der ersten Nachbarschaftssphäre befinden sich ein protoniertes Imin und eine Ammoniumverbindung. Das Anfragemolekül weist eine elektronische Situation auf, wie sie in keinem der Trainingsmoleküle zu finden ist: Die positive Ladung, die in Abbildung 3-1 formell dem Stickstoffatom zugeordnet wurde, führt zu einer Schwächung der CH-Bindungen in den beiden Methylgruppen. Entsprechend ist die dazugehörige ν CH-Schwingung bei niedrigeren Wellenzahlen zu beobachten, jedoch nicht so niedrig, wie es bei dem protonierten Imin und der Ammoniumverbindung der Fall ist. Für die weiteren Vergleiche ist der Bereich zwischen 2000 und 600 cm^{-1} vergrößert dargestellt (vgl. Abb 3-5).

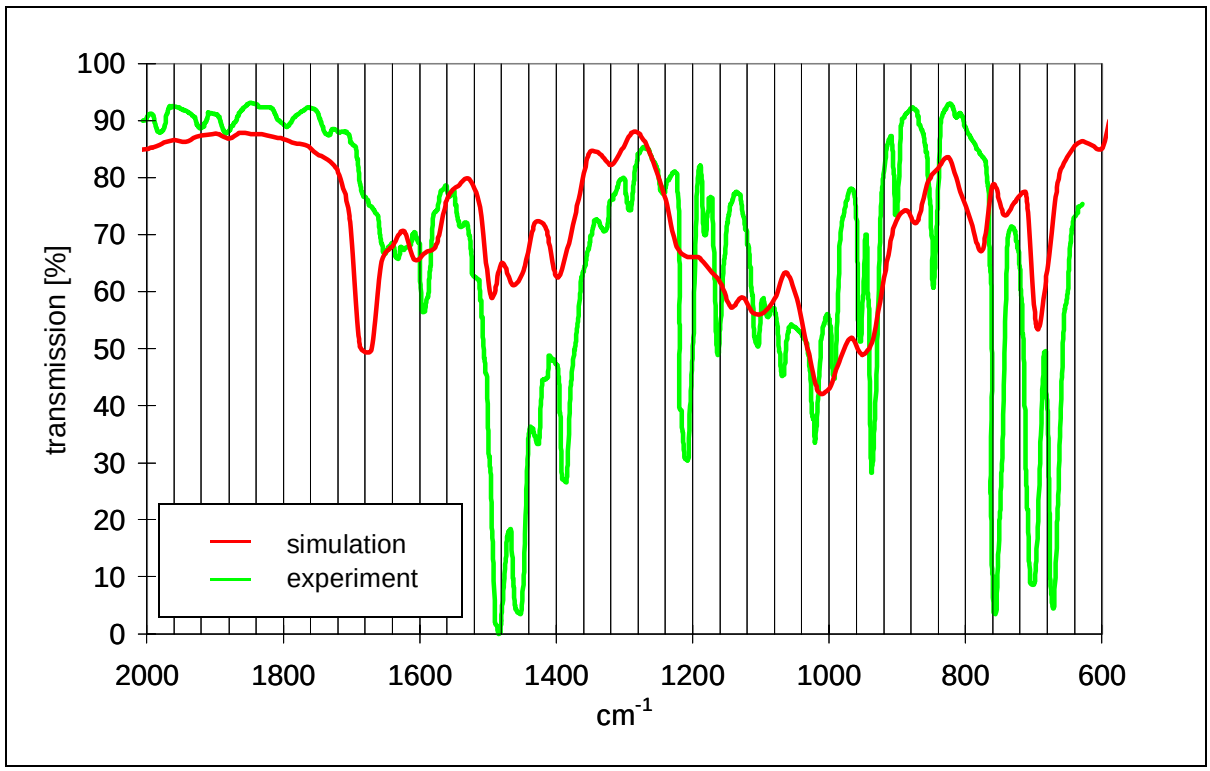


Abb. 3-5: Vergleich von simuliertem und experimentellem Spektrum (Ausschnittsvergrößerung)

In diesem Bereich weisen die Bandenmuster von experimentellem und simuliertem Spektrum einige Ähnlichkeiten auf. Die Bandenform des experimentellen Spektrums im Bereich der ν C-O und ν C-N Schwingungen zwischen 1520 und 1360 cm^{-1} ist im simulierten Spektrum sehr gut wiedergegeben. Zwischen 1300 und 900 cm^{-1} liegt das simulierte Spektrum wie eine umhüllende Kurve über dem experimentellen Spektrum. Die Ähnlichkeit zwischen simuliertem und experimentellem Spektrum in diesem für eine Substanz hochcharakteristischen Wellenzahlenbereich deutet darauf hin, daß es sich bei der vermessenen Substanz um DMA-Hydrochlorid handelt.

Eine bessere Repräsentation der Anfragestruktur durch die Moleküle des Trainingsdatensatzes hätte sicherlich noch zu einer Verbesserung der Simulationsqualität geführt.

3.2 Identifikationsversuche eines Ameisen-Spurpheromons

Mit dem Begriff Kommunikation verbindet man beim Menschen hauptsächlich den Austausch von optischen und akustischen Signalen. Im Tierreich haben sich im Laufe der Evolution auch andere Formen der Verständigung herausgebildet. Insekten orientieren sich beispielsweise in verstärktem Maße olfaktorisch. Bei staatenbildenden Insekten, wie Ameisen, spielt die Kommunikation für das Sozialleben eine herausragende Rolle. Ihre oftmals äußerst umfangreichen Lebensgemeinschaften von bis zu Millionen von Individuen können nur existieren, wenn die Aufgabenteilung innerhalb der Sozietät streng geregelt ist.[69] Pheromone als chemische Botenstoffe ermöglichen hier einen effektiven Informationsaustausch. Für ein besseres Verständnis der chemischen Ökologie ist die Strukturaufklärung und Synthese der Pheromone von großer Bedeutung. Durch die Untersuchung der Spurpheromone mehrerer Ameisenarten einer Gattung sollte untersucht werden, inwieweit die Verwendung bestimmter Signalstoffe bei Ameisen art- bzw. gattungsspezifisch ist. Bei Camponotus-Spezies werden die Spurpheromone in der Rektalblase gespeichert.[70] Die gewählte Gattung, Camponotus, zählt zu den größten Ameisengenera. Zu ihr gehören einige Arten, die als Waldschädlinge auftreten, da sie im Holz von gesunden Bäumen ihre Nestkammern ausnagen.[71]

Allgemein gestaltet sich die Pheromonanalytik sehr kompliziert, da Pheromonmengen im Pico- bis Nanogrammbereich neben einem großen Überschuß an Körperinhaltsstoffen oder Verunreinigungen identifiziert werden müssen. Pro Versuchsansatz werden dabei etwa 10-20 pheromonproduzierende Drüsen präpariert. Durch gaschromatographische und massenspektroskopische Untersuchungen sowie Mikroreaktionen gelang es Bestmann et al., Substanzen zweier Verbindungsklassen zu identifizieren. Hierbei handelte es sich um 3,4-Dihydroisocumarine [72] und δ-Lactone [73], die für die meisten der untersuchten Camponotus-Arten die spuraktiven Komponenten darstellen.

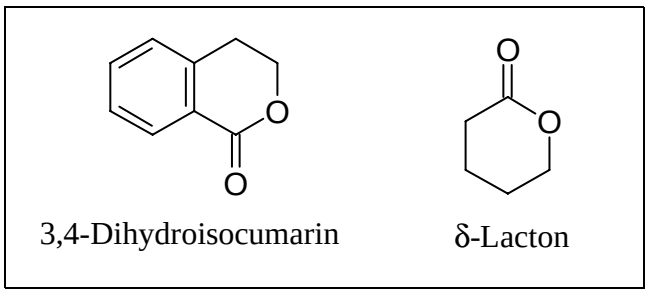


Abb. 3-6: Strukturformeln für 3,4-Dihydroisocumarin und ein δ-Lacton

Bei einer Gruppe weiterer Camponotus-Spezies scheint eine Substanz Spurfolgeverhalten auszulösen, die keiner der genannten Verbindungklassen angehört. Gaschromatographi-

sche und massenspektroskopische Untersuchungen sowie mehrere im Mikromaßstab durchgeführten Reaktionen scheinen auf ein substituiertes γ-Lacton mit einer Hydroxyfunktion in der Seitenkette hinzuweisen, wobei das Vorhandensein einer Carboxylgruppe ausgeschlossen werden kann. Zur Molmassenbestimmung wurde ein CI-Spektrum vermessen. Eine Hochauflösung des Molekülions bei m/e = 186 ergab eine Summenformel $C_{10}H_{18}O_3$. [74]

Trotz der geringen Substanzmengen war es möglich, GC-IR-Spektren in der Gasphase als auch in kondensierter Phase aufzunehmen (vgl. Abb. 3-7).

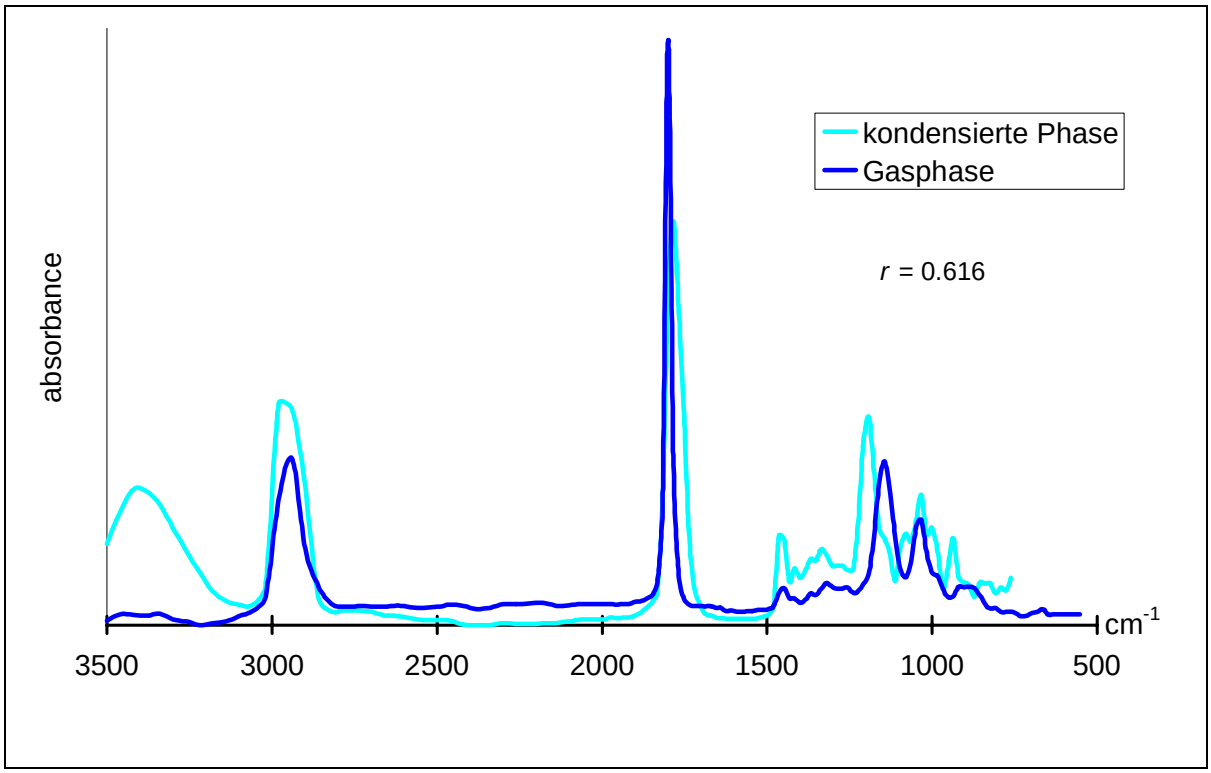


Abb. 3-7: Experimentelle GC-IR-Spektren (kondensierte Phase und Gasphase) des Pheromons

Die sehr schwache Intensität des OH-Signals beim Gasphasenspektrum kann dadurch erklärt werden, daß unter diesen Aufnahmebedingungen die Ausbildung intermolekularer Wasserstoffbrückenbindungen weitgehend unterdrückt wird. Die Aufnahmebedingungen könnten weiterhin der Grund dafür sein, daß die Wellenzahl des Carbonylsignals mit 1813 cm^{-1} ungewöhnlich hoch ist. Allgemein ist das Spektrum aufgrund der geringen Substanzmenge sehr intensitätsschwach. Deutlich zu erkennen sind die ν C-H Schwingungen zwischen 2850 und 3000 cm^{-1}, die ν C=O Schwingung bei 1813 cm^{-1} sowie zwei Signale bei 1050 und 1150 cm^{-1}, die den symmetrischen und antisymmetrischen ν C-O Schwingungen zugeordnet werden können. Das schwache Signal bei 1450 cm^{-1} ließe sich auf die CH_3-Deformationsschwingung zurückführen. Eine Suche in der IR-Spektrendatenbank des Landesuntersuchungsamtes

Erlangen[1)] ergab als ähnlichste Verbindung das Decalacton:

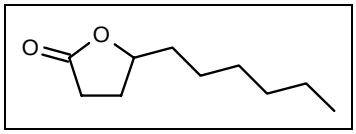


Abb. 3-8: Decalacton

Diese Verbindung läßt sich jedoch nicht mit dem Molekülpeak des Massenspektrums in Einklang bringen. Weiterhin fehlt ihr eine OH-Gruppe, auf deren Vorhandensein obige Experimente sowie das schwache Signal im Infrarotspektrum schließen lassen. Ausgehend von dieser Grundstruktur wurden eine Reihe von Isomeren generiert, die als Kandidaten für die unbekannte Verbindung in Frage kommen könnten. Da bei fehlenden Stereodeskriptoren zufällige 3D-Isomere erzeugt werden, wurden die beiden Stereozentren willkürlich auf R gesetzt, um einheitliche und reproduzierbare Bedingungen zu schaffen.

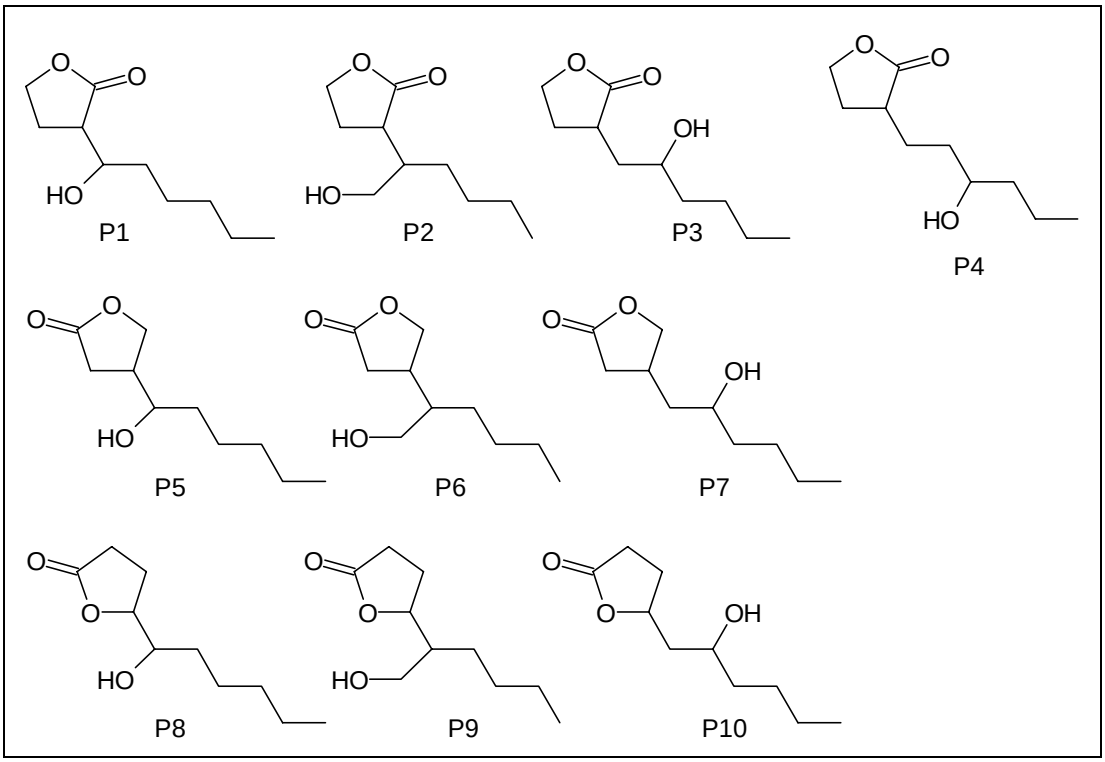


Abb. 3-9: Mögliche Kandidaten zur Identifizierung des unbekannten Pheromons, $C_{10}H_{18}O_3$

Für diese Kandidaten (vgl. Abb. 3-9) wurden die entsprechenden Infarotspektren vorher-

[1)]Landesuntersuchungsamt für das Gesundheitswesen Nordbayern, Eggenreuther Weg 43, 91058 Erlangen

gesagt. Die Simulationsparameter sind in nachfolgender Tabelle aufgeführt:

Tab. 3-3: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfragestrukturorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo gesamt
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Die simulierten Spektren wurden mittels des Korrelationskoeffizienten mit den experimentellen (Gasphase und kondensierte Phase) im Bereich von 3500 - 750 cm^{-1} verglichen. Die Verteilung der Korrelationskoeffizienten sind in Abbildung 3-10 dargestellt.

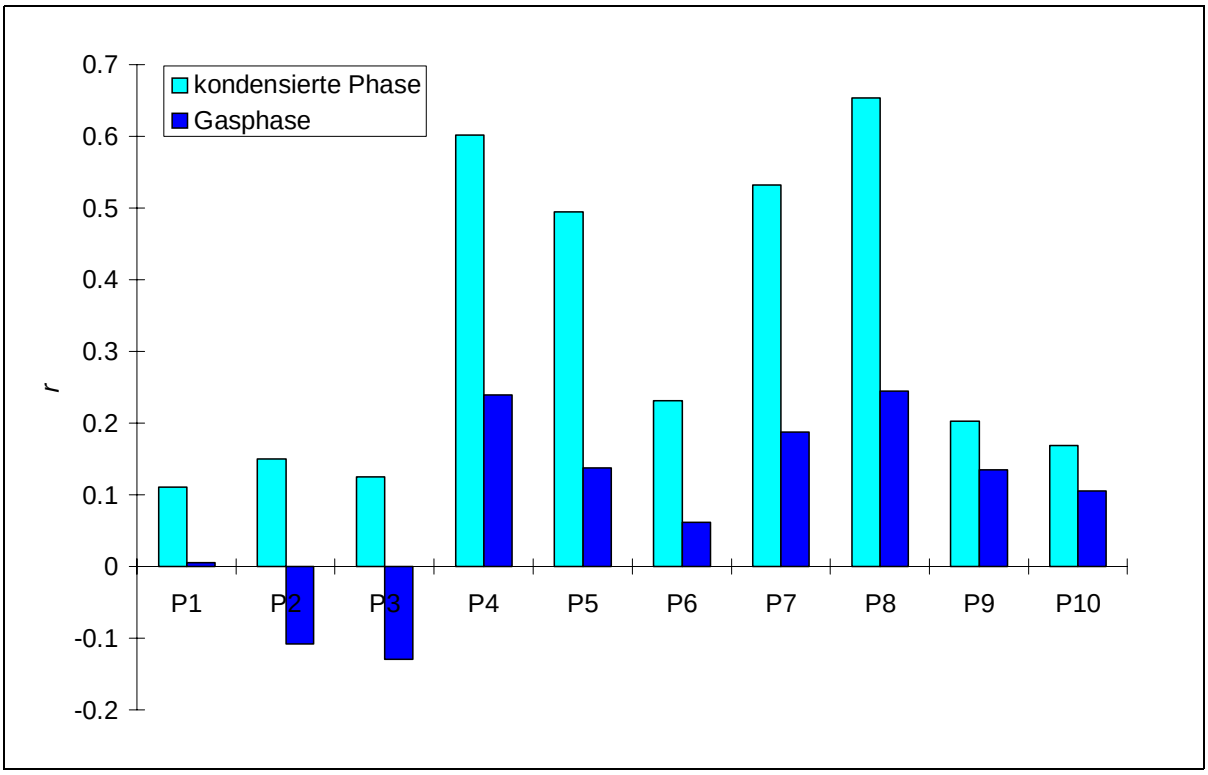


Abb. 3-10: Verteilung der Korrelationskoeffizienten *r* für die Verbindungen P1-P10

Bei allen Verbindungen ist zu erkennen, daß die Ähnlichkeiten zwischen den simulierten

Spektren und dem experimentellen Spektrum der kondensierten Phase höher liegen als bei dem Vergleich mit dem experimentellen Gasphasenspektrum. Dies läßt sich sehr leicht nachvollziehen, da zum Training der neuronalen Netze nur Spektren von festen oder flüssigen Substanzen verwendet wurden. Der größte Korrelationskoeffizient *r* ist bei dem Simulationsexperiment von Verbindung P8 zu beobachten. Prinzipiell stellt sich hier jedoch die Frage, inwieweit der Korrelationskoeffizient in diesem Größenordnungsbereich überhaupt eine sinnvolle Abstufung bezüglich Ähnlichkeiten im spektroskopischen Sinn zulassen. Das soll heißen, daß ein Spektrenpaar mit einem Korrelationskoeffizienten von $r = 0.4$ nicht zwingendermaßen eine größere Ähnlichkeit hat als ein Spektrenpaar mit $r = 0.35$.

Aus diesem Grund werden im folgenden die simulierten Spektren mit einem Korrelationskoeffizien von $r > r_{max}/2$ (= 0.327) visuell mit dem experimentellen Spektrum verglichen. Das experimentelle Spektrum wurde skaliert, indem der niedrigste Absorbanzwert auf 0 gesetzt wurde und alle Absorbanzwerte mit 15 multipliziert wurden.

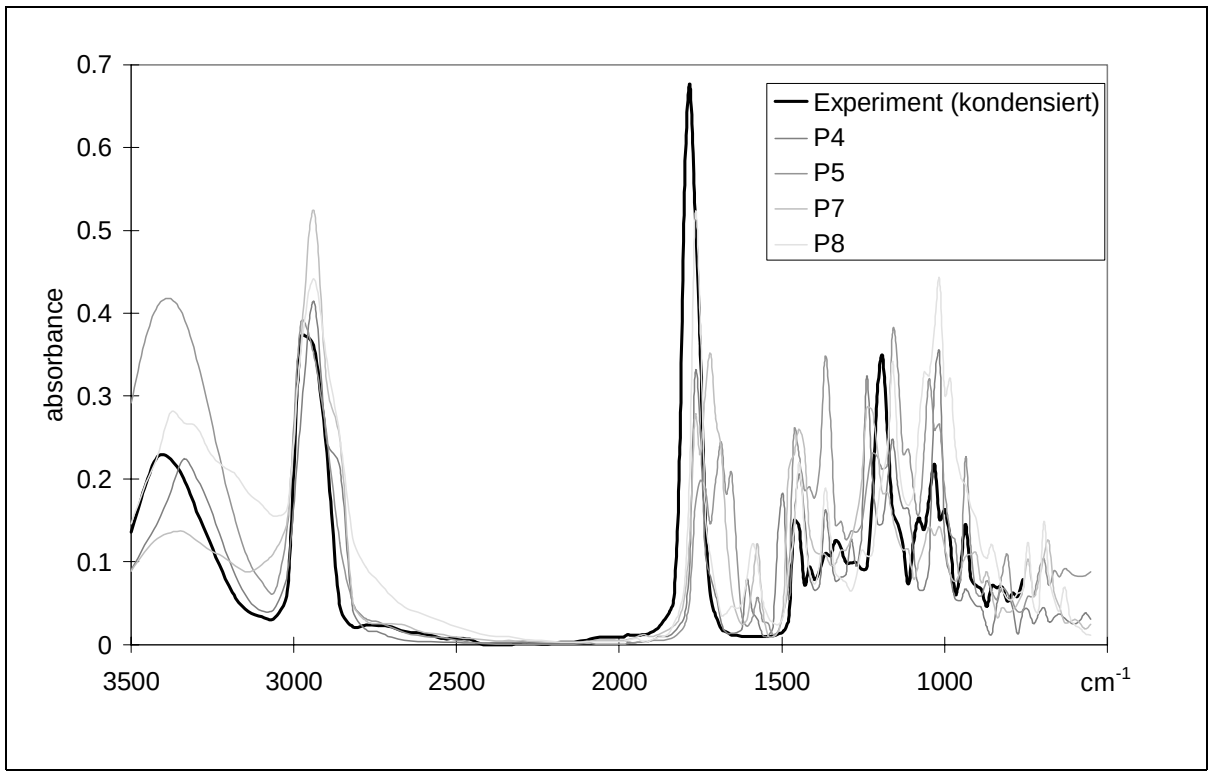
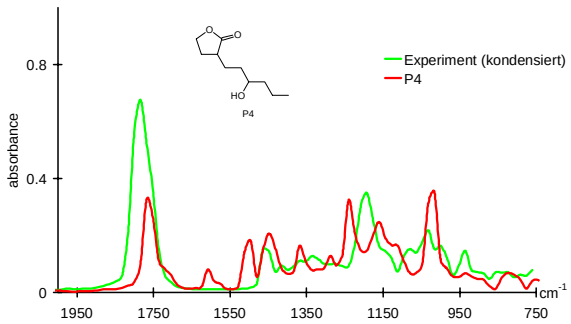
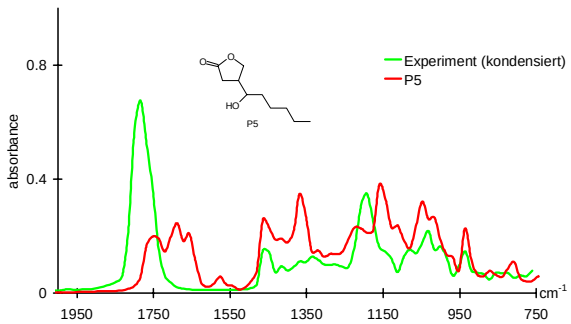
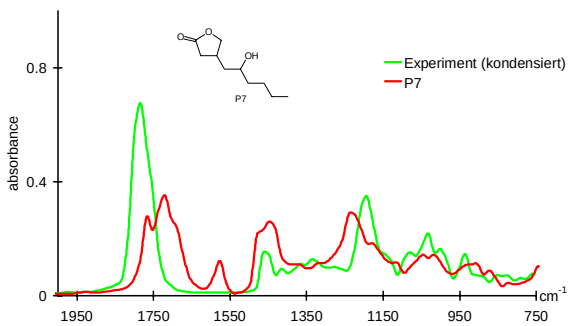


Abb. 3-11: Vergleich der simulierten Spektren mit dem experimentellen Spektrum (kondensierte Phase)

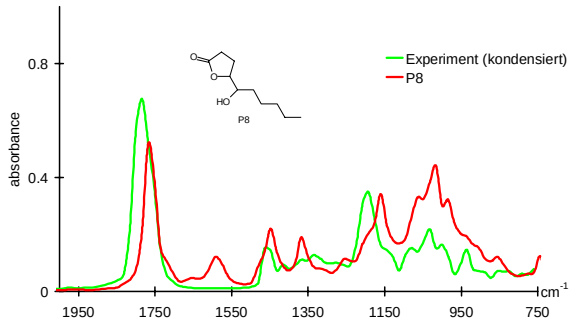
Da der Vergleich oberhalb 3000 cm^{-1} sehr stark von den Aufnahmebedingungen abhängt und die ν C-H Schwingungen zwischen 2850 und 3000 cm^{-1} bei allen Simulationen gut wiedergegeben sind, soll im nachfolgenden nur der Bereich zwischen 2000 und 750 cm^{-1} näher

betrachtet werden.

Tab. 3-4: Visueller Vergleich der simulierten Spektren mit dem experimentellen Gasphasenspektrum

Spektrenvergleich	Diskussion
 Abb. 3-12: Simulation für P4	Die Lage des Carbonylsignals ist bei diesem Simulationsexperiment gut wiedergegeben. Die Bandenmuster von simuliertem und experimentellem Spektrum sind jedoch sehr unähnlich.
 Abb. 3-13: Simulation für P5	In den Bereichen 1350 - 1450 cm^{-1} und bei etwa 900 - 1000 cm^{-1} sind sich die Bandenmuster von simuliertem und experimentellem Spektrum ähnlich. Im übrigen Spektrum sind jedoch deutliche Unterschiede sowohl bei der Form als auch bei der Lage der Signale zu erkennen.
 Abb. 3-14: Simulation für P7	Im Bereich von 750-1550 cm^{-1} sind sich die Spektrenverläufe von Simulation und Experiment sehr ähnlich. Oberhalb 1550 cm^{-1}, insbesonders bei der Lage des Carbonylsignals, unterscheiden sich beide Spektren jedoch deutlich.

Tab. 3-4: Visueller Vergleich der simulierten Spektren mit dem experimentellen Gasphasenspektrum

Spektrenvergleich	Diskussion
 Abb. 3-15: Simulation für P8	Simulation und Experiment sind sich bezüglich der Lage der Banden sehr ähnlich. Im besonderen fällt auf, daß sich die Bandenmuster in Simulation und Experiment sehr gut entsprechen. Einzig bei 1570 cm^{-1} ist im simulierten Spektrum eine Bande zu erkennen, die im experimentellen Spektrum nicht vorhanden ist.

Diskussion der Ergebnisse:

Beim visuellen Vergleich der simulierten Spektren mit dem experimentellen Spektrum sind bei der Simulation für Verbindung P8 (vgl. Abb. 3-15) die meisten Übereinstimmungen mit dem Experiment zu erkennen. Die Verbindung gilt somit als aussichtsreichster Kandidat bei der Identifizierung des Pheromons.

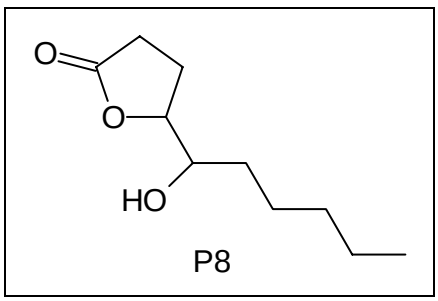


Abb. 3-16: Aussichtsreichster Kandidat bei der Identifizierung des Pheromons mit unbekannter Struktur

Abschließend muß bemerkt werden, daß für die biologische Wirksamkeit einer Verbindung deren Stereochemie und dreidimensionale Struktur von enormer Bedeutung ist. Oft entscheidet die Orientierung eines Stereozentrums über die Wirkungsweise des Moleküls. Bei den dargestellten Simulationsexperimenten wurden beide Stereozentren auf R gesetzt. Dies geschah willkürlich, um einheitliche und reproduzierbare Bedingungen zu schaffen. Bei einigen Experimenten konnte beobachtet werden, daß das Simulationsergebnis deutlich von der eingegebenen Stereochemie abhängt. Dies ist sehr leicht nachzuvollziehen, da verschiedene

Diastereomere zu verschiedenen Strukturcodes und damit zu anderen Molekülen bei der Trainingsdatensatzauswahl führen. Die Datenbasis, aus welcher der Trainingsdatensatz zusammengestellt wird, ist jedoch bezüglich der darin enthaltenen Information sehr heterogen: Einerseits weisen viele der Verbindungen keine Chiralität auf. Andererseits ist die chirale Information bei Verbindung mit Stereozentren nicht immer abgelegt, wobei Enantiomere sowohl infrarotspektroskopisch als auch durch die verwendete Strukturcodierung ohnehin nicht zu unterscheiden sind. Fehlt Stereoinformation in der Datenbank, so wird bei der Generierung der 3D-Struktur ein zufälliges Stereoisomer erzeugt, welches selbstverständlich nicht unbedingt der vermessenen Verbindung entsprechen muß. Dies verläuft bei der Generierung der 3D-Struktur für die eingegebene Verbindung ohne explizite Angabe der Stereochemie ganz analog. Aus diesem Grund muß die Berücksichtigung der Stereochemie bei diesem Experiment prinzipiell kritisch betrachtet werden. Es kann eine Genauigkeit der Ergebnisse vorgetäuscht werden, die durch die zur Verfügung stehenden Daten nicht gewährleistet ist.

Auch hier könnte durch eine qualitativ hochwertige Datenbasis die Simulationsqualität, und damit die Vorhersagesicherheit erhöht werden. Dazu wäre es nötig, daß der Strukturraum, welcher der Fragestellung zugrundeliegt, infrarotspektroskopisch gut abgedeckt ist und die den Spektren entsprechende Stereoinformation ebenfalls in der Datenbank gespeichert ist. Die verwendete Strukturcodierung eignet sich sehr gut zur Transformation dieser wichtigen Information, da sie sehr empfindlich auf kleine Änderungen der 3D-Struktur reagiert und sich somit Diastereomere anhand des Strukturcodes gut unterscheiden lassen.

3.3 Identifikation von Herbizid-Abbauprodukten

Jährlich werden in der Bundesrepublik (alte Bundesländer) etwa 30000 Tonnen Pestizide eingesetzt,[75] wobei Schätzungen zufolge nur ca. 0.1% der ausgebrachten Pestizidmenge die Zielorganismen erreichen.[76] Während in Ländern der Dritten Welt jährlich etwa 9000 Menschen bei der Herstellung und dem Einsatz von Pestiziden tödliche Vergiftungen erleiden,[77] ist in den Industrieländern die Kontamination von Lebensmitteln und Trinkwasser die Hauptgefahr.

Im nachfolgenden soll die Gruppe der Triazin-Herbizide (Unkrautvernichtungsmittel), wegen der symmetrischen Anordnung der Stickstoffatome im aromatischen Sechsring auch als s-Triazine bezeichnet, näher betrachtet werden. Die Wirkungsweise der Triazin-Pestizide beruht auf der Hemmung des Photosynthesemechanismus.[78] Diese Hemmung tritt jedoch nicht bei Mais, Zuckerrohr oder Ananaspflanzen auf, weshalb der Anbau dieser Pflanzen sowie die Behandlung von Nichtkulturland das Haupteinsatzgebiet für diese Pestizide ist. Im Wechselanbau mit Getreide kann jedoch durch die relativ hohe Persistenz, die in Abhängigkeit von verschiedenen äußeren Bedingungen bis zu 100 d beträgt, auch eine Störung des Nutzpflanzenwachstums verursacht werden. Wegen der hohen Persistenz im Boden ist der Einsatz der wohl bekanntesten Verbindung dieser Gruppe, dem Atrazin (2-Chloro-4-ethylamino-6-isopropylamino-1,3,5-triazin), seit 1991 in Deutschland verboten. Da eine EU-einheitliche Regelung angestrebt wird, ist jedoch mit einer Wiederzulassung des Produkts zu rechnen. Es kommt zuem vor, daß Atrazin in Deutschland illegalerweise eingesetzt wird, wobei die Landwirte auf alte Vorräte zurückgreifen bzw. in Nachbarländern einkaufen.

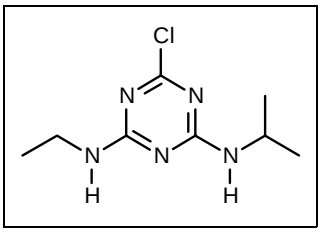


Abb. 3-17: Strukturzeichnung des Atrazins

Aus diesen Gründen ist eine zuverlässige Analytik dieser Stoffe von hoher toxikologischer, ökologischer und ökonomischer Relevanz. Oftmals gestaltet sich die Analytik in Ökosystemen jedoch schwierig, da neben den verschiedenen Wirk- bzw. Schadstoffen auch deren Abbauprodukte analysiert werden müssen. Eine einfache Identifikation durch den Vergleich der experimentellen Spektren mit Referenzspektren ist meist nicht möglich, da eine Vielzahl dieser oftmals isomeren Verbindungen in keiner Spektrendatenbank zu finden sind. In der

Regel ist jedoch in etwa bekannt, mit welchen Komponenten bei der Analyse eines Systems zu rechnen ist, da man den Ausgangsstoff der Kontamination kennt. Ein Lösungsansatz wäre also beispielsweise, für eine bekannte Ausgangsverbindung mittels eines Strukturgenerators alle möglichen Verbindungen zu generieren, um so potentielle Kandidaten für die Identifikation zu erhalten. Bei einem Strukturgenerator, der nur auf eine korrekte Konnektivität bei der Erzeugung möglicher Isomeren achtet, erreicht man schnell eine unüberschaubare Menge an Verbindungen und damit die Kapazitätsgrenzen des Identifikationssystems. Effektiver wäre hier ein intelligentes System einzusetzen, welches bei der Generierung möglicher Verbindungen berücksichtigt, ob diese überhaupt chemisch sinnvoll sind, z.B. durch das Verbot zu hoher Ringspannungen. Bei der Berücksichtigung möglicher Abbauprodukte würde jedoch auch so ein Strukturgenerator noch zu viele zu überprüfende Kandidaten anbieten. Als Konsequenz erscheint es sinnvoll, ein Reaktionsvorhersagesystem als Strukturgenerator einzusetzen, das für die bekannte Ausgangsverbindung mögliche Abbauprodukte generiert.

3.3.1 Beschreibung des Experiments

Bei dem Computerexperiment wurde angenommen, daß ein Boden mit Triazinderivaten behandelt worden war.[79] Bei der Analyse des Bodens, wie sie beispielsweise mittels eines GC-IR-, oder LC-IR-Geräts hätte durchgeführt werden können, wurden mehrere Komponenten isoliert, die es galt zu identifizieren. Dazu wurde für die Ausgangskomponente, die ursprünglich auf das Feld ausgebracht worden war, ein Baum möglicher Abbauprodukte mittels EROS7 generiert.[80][81] EROS (***E***laboration of ***R***eactions for ***O***rganic ***S***ynthesis) ist ein regelbasiertes Reaktionsvorhersagesystem. Die Regelbasis wird aus Gründen der Flexibilität extern zum Kernprogramm gehalten.

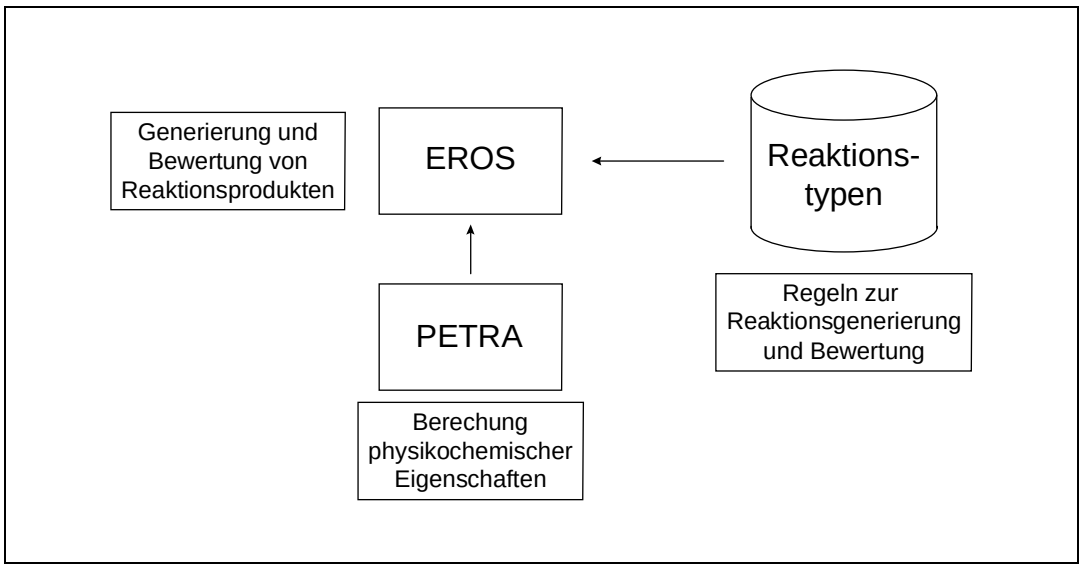


Abb. 3-18: Schematischer Aufbau des Reaktionsvorhersagesystems EROS

EROS untersucht inwiefern sich die Reaktionsregeln unter Berücksichtigung physikochemischer Eigenschaften auf ein Molekül anwenden lassen. Beispielsweise, ob das Molekül eine C-Cl Bindung enthält, die eine entsprechenden Polarität aufweist, so daß sie hydrolysiert werden kann. Bei dem nachfolgenden Versuch wurden die reduktive Dealkylierung sowie die Hydrolyse als Abbaureaktionen zugelassen (vgl. Abb. 3-19).

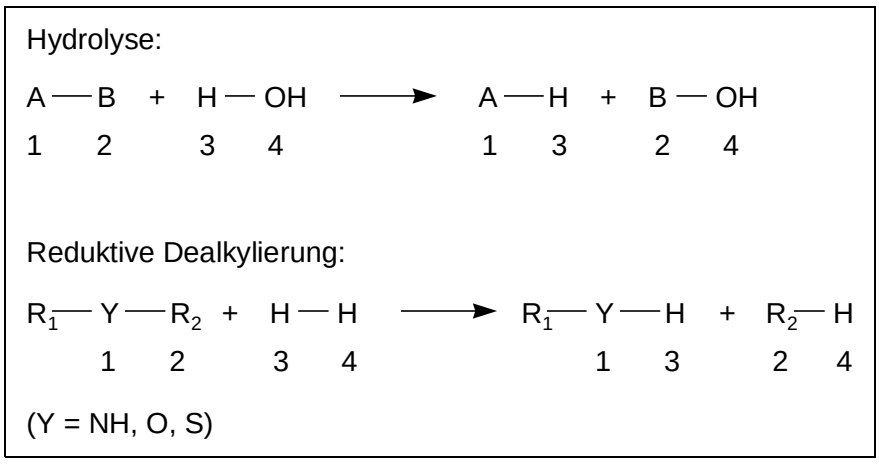


Abb. 3-19: Reaktionsschemata für die Reaktionstypen Hydrolyse und reduktive Dealkylierung

Für diesen Baum an möglichen Abbauprodukten wurden die entsprechenden IR-Spektren simuliert und mit den experimentellen Spektren verglichen. Für jedes experimentelle Spektrum wurde das ähnlichste, simulierte Spektrum ermittelt. Die der Simulation zugrundeliegende Verbindung wurde als Lösung für die jeweilige Identifikation präsentiert. Das

Gesamtschema des Experiments ist in Abbildung 3-20 abgebildet.

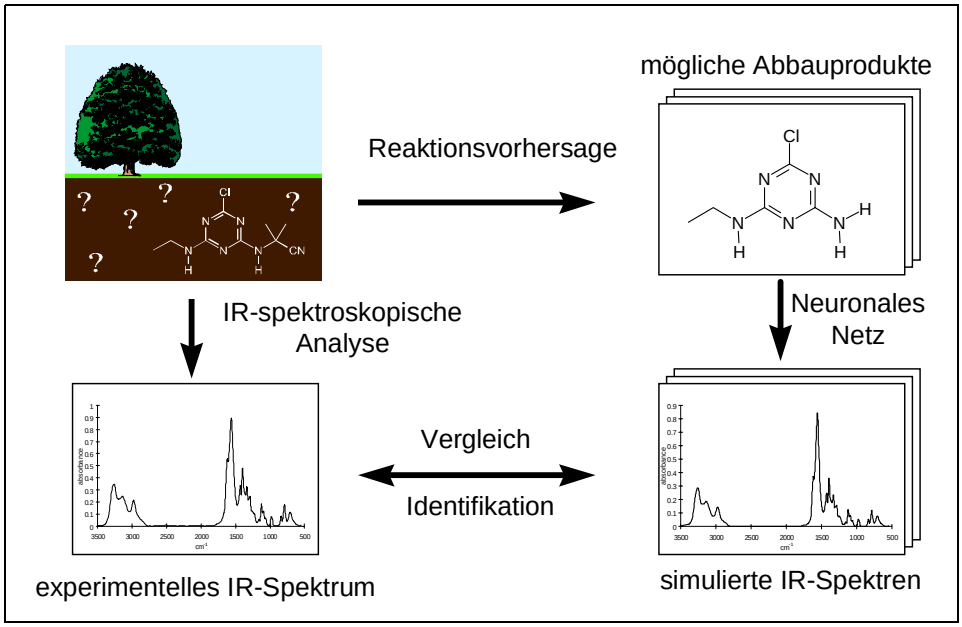


Abb. 3-20: Schematischer Ablauf des Computerexperiments

3.3.2 Cyanazin

Bei dem ersten Experiment wurde angenommen, daß der untersuchte Boden mit Cyanazin behandelt worden war.

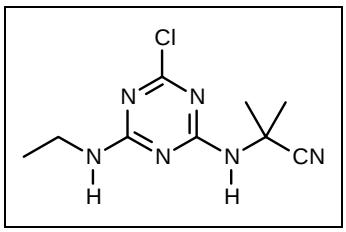


Abb. 3-21: Strukturformel von Cyanazin

Die folgenden drei Spektren sollten nun identifiziert werden:

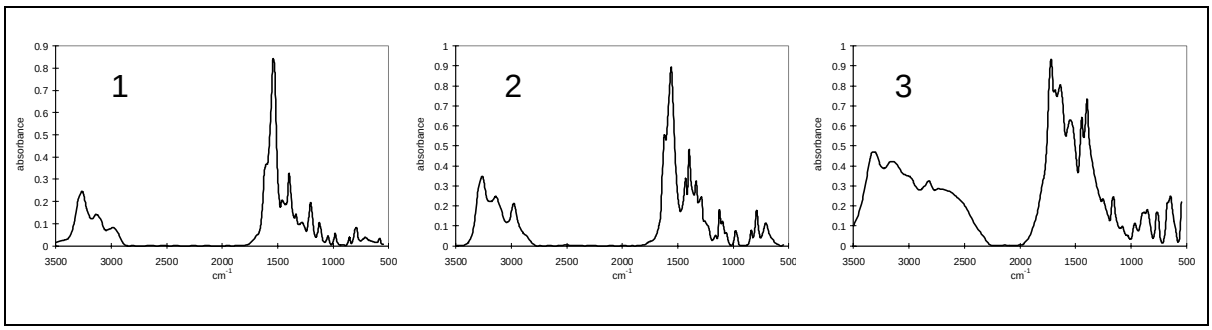


Abb. 3-22: Zu identifizierende experimentelle Spektren

Bei diesem Experiment war selbstverständlich bekannt, von welchen Verbindungen die experimentellen Spektren stammten, damit die Richtigkeit der Identifikation überprüft werden konnte. Diese Information wurde jedoch beim eigentlichen Identifikationsvorgang nicht mit-einbezogen.

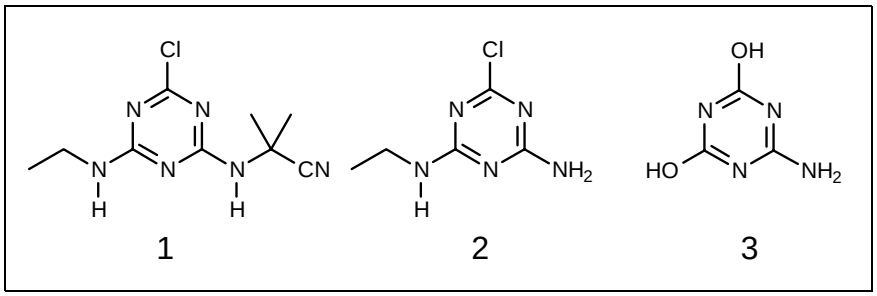


Abb. 3-23: Zu identifizierende Verbindungen

Ausgehend von Cyanazin wurde folgender Reaktionsbaum aufgestellt:

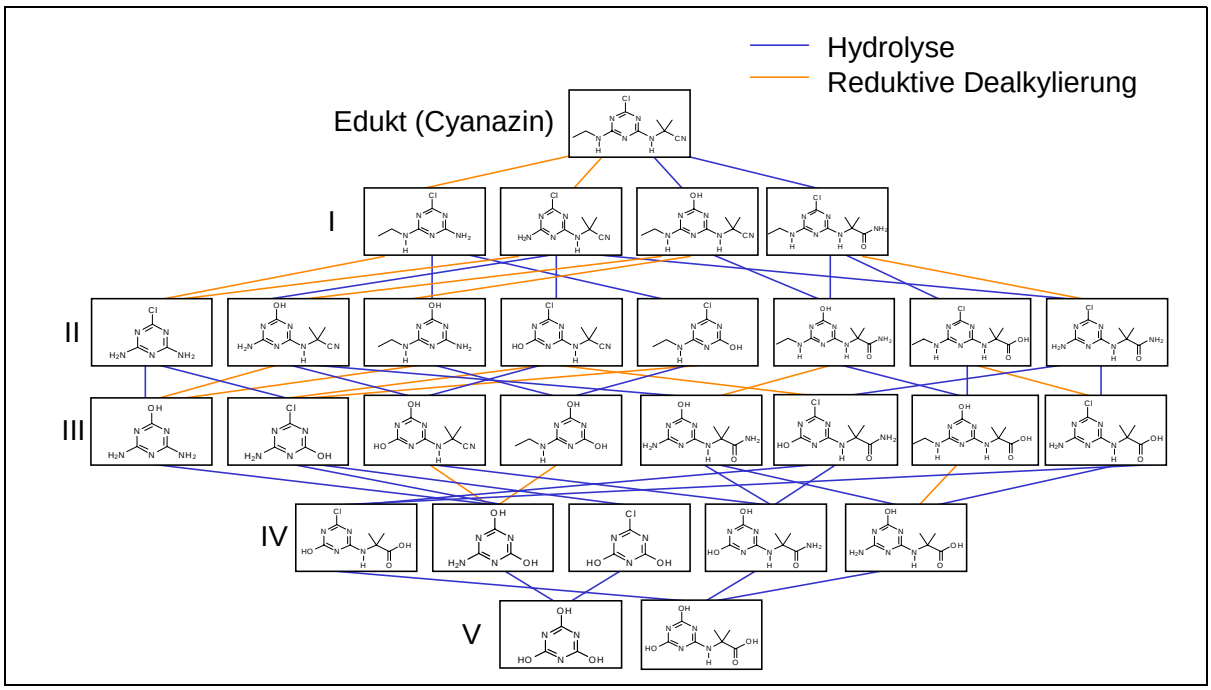


Abb. 3-24: Baumschema möglicher Abbauprodukte des Cyanazins (vgl. Anhang A.6)

Bei der Generierung des Abbaubaumes (vgl. Abb. 3-24) wurden die reduktive Dealkylierung und die Hydrolyse als Reaktionen berücksichtigt. Diese beiden Reaktionstypen stellen laut Kearney et al. [82] die Hauptabbauwege für diese Verbindungsspezies dar. Für alle Verbindungen des Abbaubaumes wurden die entsprechenden Infrarotspektren vorhergesagt. Die Parameter für das Simulationsexperiment finden sich in nachfolgender Tabelle:

Tab. 3-5: Simulationsparameter

Strukturcodierung	64 3D-MoRSE-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfrageorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo, ungeladene H, C, N, O, Hal -Verbindungen (9850) Zusätzlich wurden die drei zu identifizierenden Moleküle aus dem Datensatz entfernt, so daß sie in keinem der Trainingsdatensätze enthalten waren.
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht

Somit ergibt sich für den Baum an Abbauprodukten ein entsprechender Baum an Infrarotspektren.

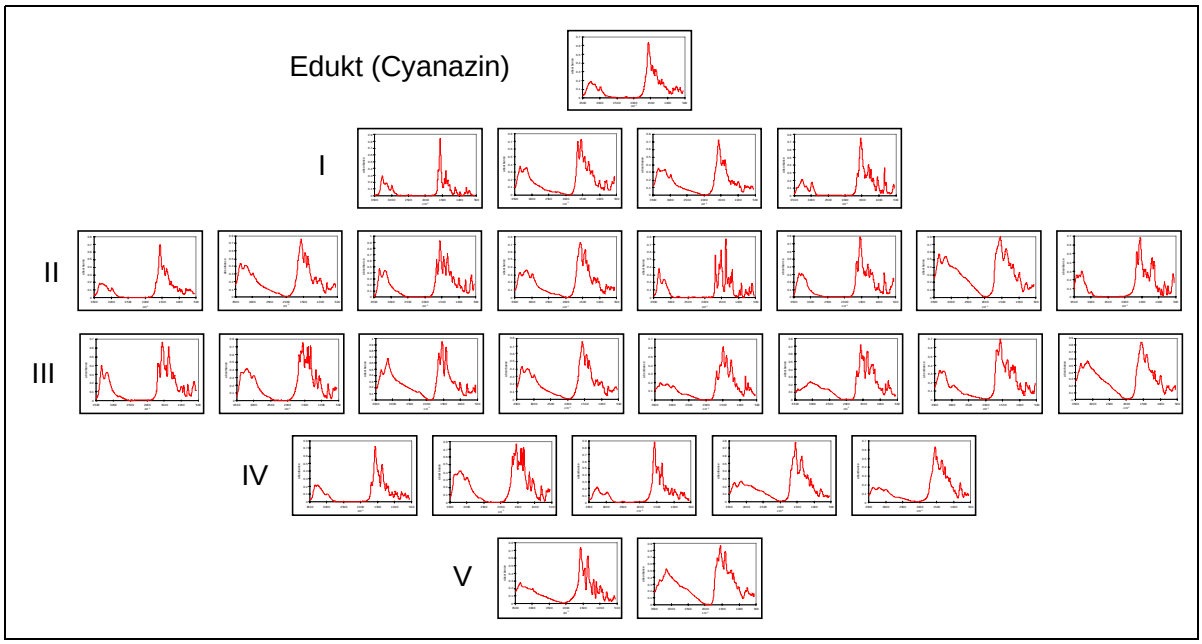


Abb. 3-25: Vorhergesagte Infrarotspektren

Die simulierten Spektren wurden mit den drei experimentellen Spektren verglichen und entsprechend den ermittelten Ähnlichkeiten geordnet. Als Vergleichsmaß wurde der bereichsgewichtete Korrelationskoeffizient r_b berechnet, da dieser bei der Erkennung von Spektrenidentität die größte Trefferquote aufweisen konnte (vgl. Kap. 2.3.1). Bei der Bereichsgewichtung wurden die Gewichte für die Wellenzahlen zwischen 2740 und 1800 cm^{-1} auf Null gesetzt, ansonsten auf Eins.

Bei den beiden ersten Spektren sind relativ hohe Ähnlichkeiten (r_b = 0.885 und r_b = 0.987) mit zwei simulierten Spektren zu beobachten. Bei dem dritten Spektrum ist das ähnlichste simulierte Spektrum mit r_b = 0.719 relativ unähnlich.

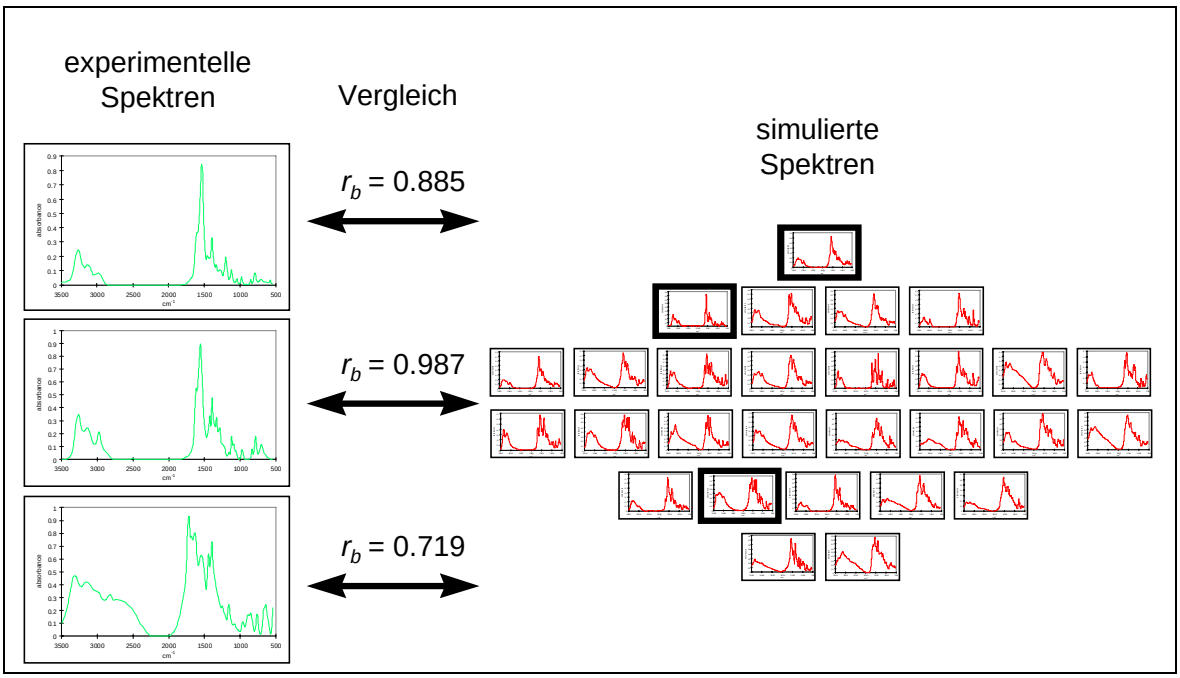


Abb. 3-26: Vergleich von simulierten und experimentellen Spektren

Folgende Strukturen werden als Lösungen präsentiert:

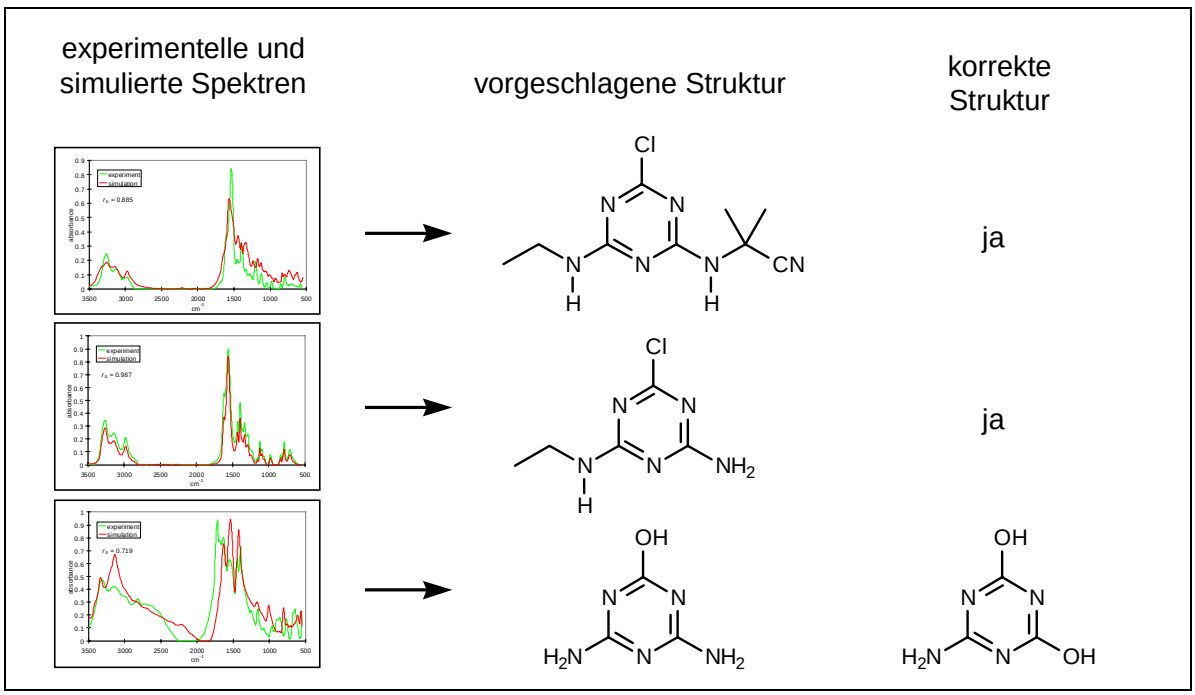


Abb. 3-27: Überprüfung der vorgeschlagenen Strukturen

Bei den ersten beiden Spektren werden die korrekten Strukturen als Lösungen präsen-

tiert. Beim dritten Spektrum ist die vorgeschlagene Struktur nicht ganz korrekt. Während die vom System ausgegebene Struktur zwei Aminofunktionen und eine Hydroxyfunktion enthält, hat die korrekte Struktur zwei Hydroxy- und eine Aminofunktion. Eine Simulation für das entsprechende Tautomere (vgl. Abb. 3-28) bringt keine Verbesserung. Der bereichsgewichtete Korrelationskoeffizient r_b zwischen dem Tautomerenspektrum und dem experimentellen Spektrum beträgt 0.648.

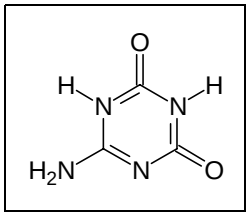


Abb. 3-28: Tautomeres von Verbindung 3

3.3.3 Trietazin

Das Experiment wurde analog für Trietazin durchgeführt:

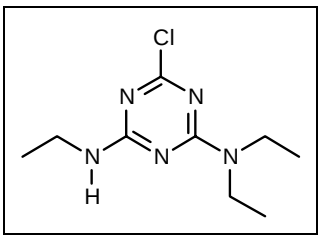


Abb. 3-29: Strukturformel von Trietazin

Mit den folgenden drei Verbindungen und den dazugehörigen Spektren wurde das Identifikationsexperiment durchgeführt:

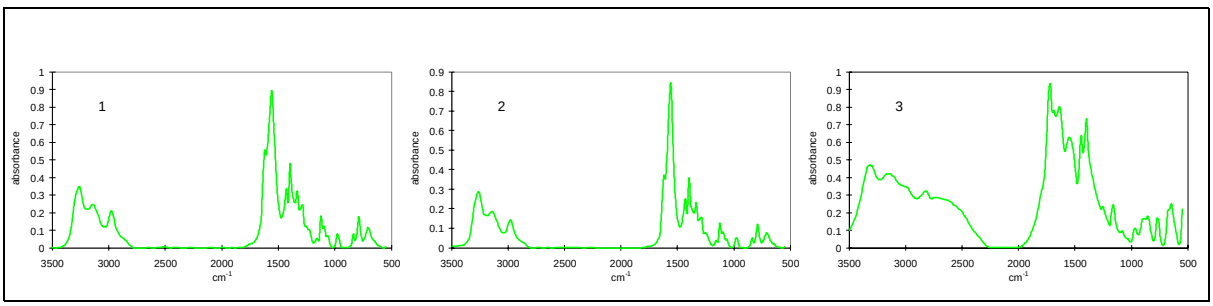


Abb. 3-30: Experimentelle IR-Spektren der drei Testverbindungen

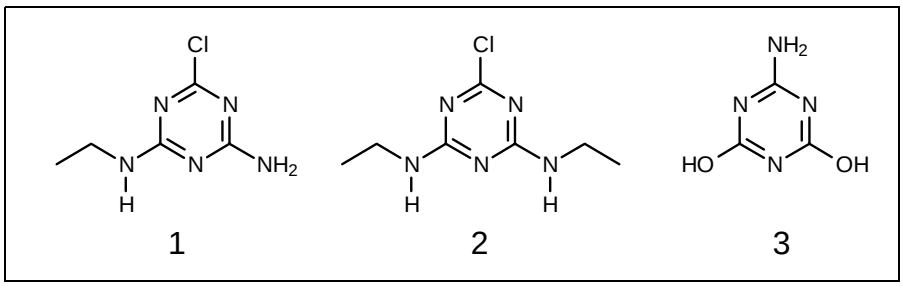


Abb. 3-31: Strukturzeichnungen der drei Testmoleküle

Mit dem Reaktionsvorhersagesystem EROS wurde ausgehend vom Trietazin ein Reaktionsbaum möglicher Abbauprodukte erstellt:

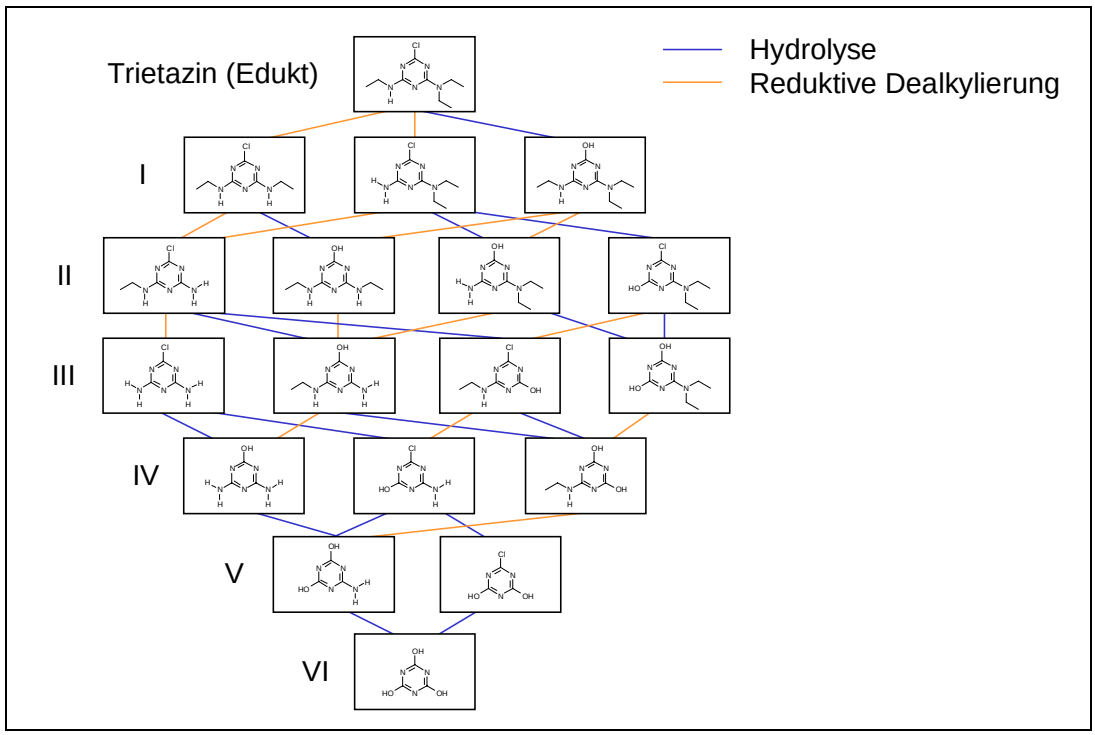


Abb. 3-32: Baumschema möglicher Abbauprodukte des Trietazins

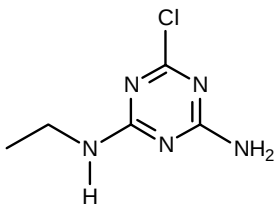
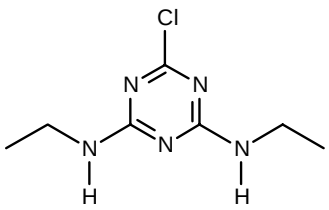
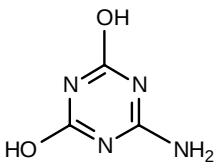
Es wurden nicht nur die Abbauprodukte, sondern auch noch mögliche Tautomere berücksichtigt, welche jedoch in obiger Graphik aus Gründen der Übersichtlichkeit nicht dargestellt sind. Eine Darstellung aller generierten Verbindungen befindet sich in Anhang A.6. Für diese Verbindungen wurden die entsprechenden Infrarotspektren simuliert. Die Simulationsparameter sind in nachfolgender Graphik aufgeführt:

Tab. 3-6: Simulationsparameter

Strukturcodierung	128 Radialcode-Werte mit $A_i = q_{tot}$
Trainingsdatensatzauswahl	anfrageorientiert
Anzahl der Trainingsmoleküle	50
Datenbasis	SpecInfo, ungeladene H, C, N, O, Hal -Verbindungen (9850 Moleküle) Zusätzlich wurden die drei zu identifizierenden Moleküle aus dem Datensatz entfernt, so daß sie in keinem der Trainingsdatensätze enthalten waren.
Neuronen	10 x 10
Netzwerkform	toroidal
Training	unüberwacht und überwacht

Die simulierten Spektren, einschließlich der Tautomerenspektren, wurden mit den drei experimentellen Spektren verglichen. Als Vergleichsmaß wurde, wie im vorherigen Cyanazin-Experiment, der bereichsgewichtete Korrelationskoeffizient r_b berechnet. Die Rangliste des Experiments mit unüberwachtem Training ist in nachfolgender Tabelle dargestellt. Die Tabellenfelder mit den den Testmolekülen entsprechenden Verbindungen des Abbaubaumes sind grau schraffiert. Entsprechende Tautomere von Verbindung 3 (≡ V1), die ebenfalls als Treffer gewertet werden können, sind ebenfalls mit einem fetten Rahmen markiert.

Tab. 3-7: Rangliste beim Vergleich der simulierten und experimentellen Spektren (unüberwachtes Training)

	1 ≡ II1 	2 ≡ I1 	3 ≡ V1 
1	V2 Tautomer	V2 Tautomer	VI1 Tautomer
2	V2 Tautomer	V2 Tautomer	V2 Tautomer
3	II1	II1	III4 Tautomer
4	III3	III3	V1 Tautomer

Tab. 3-7: Rangliste beim Vergleich der simulierten und experimentellen Spektren (unüberwachtes Training)

5	I3 Tautomer	I3 Tautomer	IV3 Tautomer
6	I3 Tautomer	I§ Tautomer	VI1
7	IV1	IV1	V1 Tautomer
8	III 2 Tautomer	III2 Tautomer	IV1 Tautomer
9	II4 Tautomer	II4 Tautomer	III3 Tautomer
10	IV2 Tautomer	I1	V1 Tautomer
11	I1	IV2	IV3 Tautomer
12	III3 Tautomer	IV2 Tautomer	VI1 Tautomer
13	IV2	III3 Tautomer	V1 Tautomer
14	III2 Tautomer	III2 Tautomer	V1

Für einen besseren Vergleich ist der Abbaubaum nochmals dargestellt:

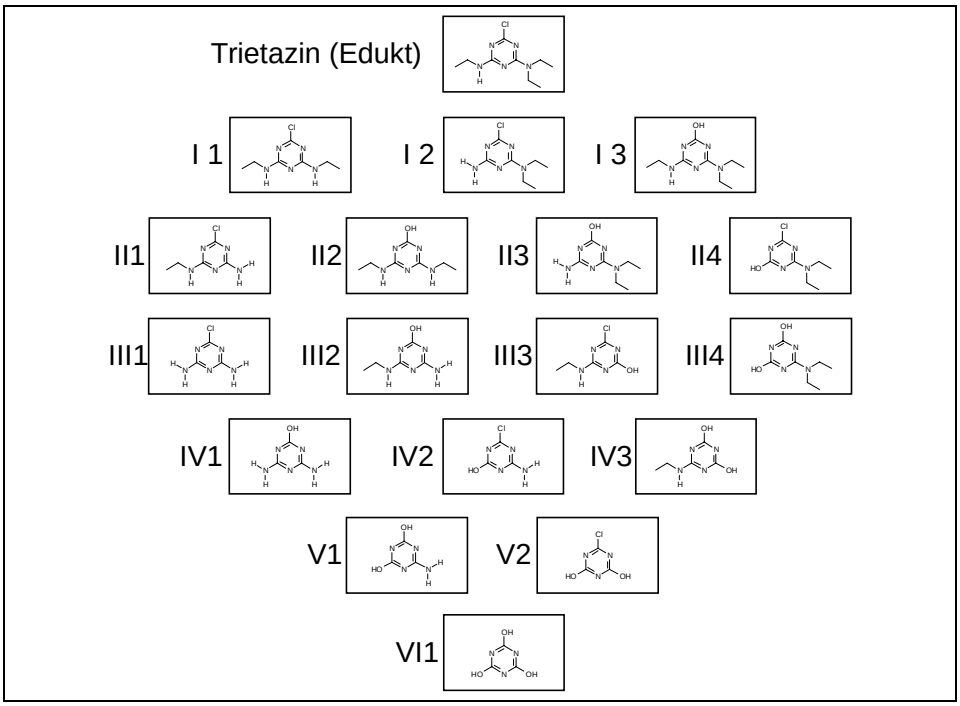


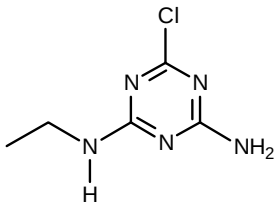
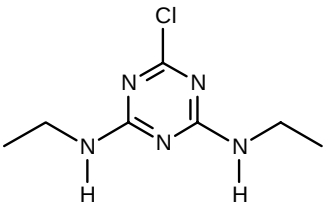
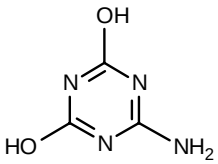
Abb. 3-33: Abbaubaum für Trietazin

Auf den ersten Blick ist das Ergebnis des Simulationsexperiments sehr schlecht. Der korrekte Lösungsvorschlag für Verbindung 1 kommt erst auf Platz 3, für Verbindung 2 auf Platz 10 und für Verbindung 3 auf Platz 14 bzw. ein entsprechendes Tautomer auf Platz 4. Werden

die Tautomeren bei der Erstellung der Rangliste nicht berücksichtigt, so sind die Ergebnisse mit denen des Cyanazinexperiments vergleichbar. Verbindung 1 liegt dann auf Platz 1, Verbindung 2 auf Platz 3 und Verbindung 3 auf Platz 2.

Das gleiche Experiment wurde wiederholt, wobei das Training des neuronalen Netzes diesmal unüberwacht durchgeführt wurde. Die Ergebnisse sind in nachfolgender Tabelle aufgeführt, wobei die entsprechenden Felder, wie oben beschrieben, markiert sind.

Tab. 3-8: Rangliste beim Vergleich der simulierten und experimentellen Spektren (überwachtes Training)

	1 ≡ II1 	2 ≡ I1 	3 ≡ V1 
1	I1	III3	VI1 Tautomer
2	III3	I1	V1 Tautomer
3	II1	II4	V1
4	II4	II1	IV3
5	Edukt	Edukt	IV2

Die Ergebnisse der Experimente mit Berücksichtigung der Tautomeren und mit überwachtem Training fallen deutlich besser aus als beim unüberwachten Training. Läßt man auch hier die Tautomere weg, so entsprechen die Plazierungen durchschnittlich denen beim Experiment mit unüberwachtem Netztraining: Verbindung 1 liegt auf Platz 3, Verbindung 2 auf Platz 2 und Verbindung 3 auf Platz 1.

3.3.4 Diskussion der Ergebnisse

Die Ergebnisse für das Cyanazin-Beispiel fallen etwas besser aus als bei dem Trietazin-Beispiel. Dies läßt sich insofern leicht nachvollziehen, da die zu indentifizierenden Moleküle beim Cyanazin-Beispiel größere strukturelle Unterschiede aufweisen als beim Trietazin-Beispiel. Besonders beim Trietazin-Experiment mit überwachtem Training wurde bei den strukturell sehr ähnlichen Molekülen Verbindung 1 und Verbindung 2 die jeweils andere Verbindung

höher bewertet als die korrekte (vgl. Tab. 3-8). Werden die Tautomere bei der Betrachtung miteinbezogen, so ist kaum mehr eine Systematik bei der Festlegung der Reihenfolge zu beobachten. Dies ist insofern verständlich, da die Fragestellung durch die große Ähnlichkeit der Abbauprodukte sehr komplex ist. Die experimentellen Spektren der Abbauprodukte unterscheiden sich in manchen Fällen (z.B. IV1 und V1) nur marginal. Es ist sogar zu vermuten, daß in solchen speziellen Fällen die spektralen Unterschiede eher auf unterschiedliche experimentelle Bedingungen zurückzuführen sind als auf strukturelle Ursachen. Inwieweit sich solche möglicherweise zufälligen Spektrenunterschiede mit Unterschieden im Strukturcode, welche ja einer klaren Systematik folgen, überhaupt korrelieren lassen, wäre eine zusätzliche interessante Fragestellung. Ein weiterer Aspekt, bei dem zufällige Einflüsse eine Rolle spielen, ist die Erstellung der 3D-Struktur. Je nachdem von welcher Seite des Moleküls der 3D-Strukturgenerator beginnt das Modell zu erstellen, was wiederum von der Numerierung der Atome abhängt, können sich die resultierenden, internen Koordinaten unterscheiden. Dies wirkt sich auf den berechneten Strukturcode aus. Nachfolgende Abbildung zeigt zwei Radialcodes für Verbindung 1, wobei die Strukturen unterschiedlichen Datensätzen entnommen wurden und leicht unterschiedliche 3D-Strukturen aufweisen. Die Strukturcodes (Radialcodierung, $A_i = q_{tot}$) unterscheiden sich mit einem *rms*-Wert von 0.028.

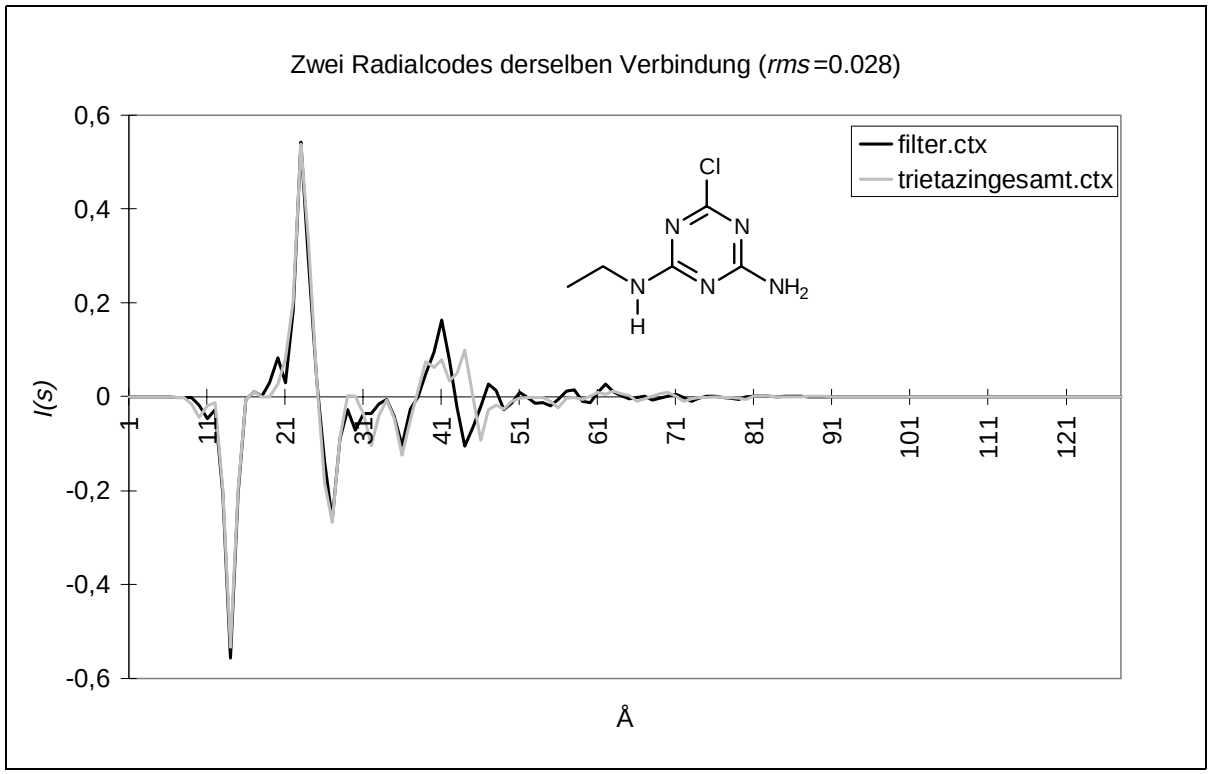


Abb. 3-34: Zwei Strukturcodes von Verbindung 1

Die Auswirkung dieses Effekts soll durch Abbildung 3-35 verdeutlicht werden. Hier wurde der Strukturcode für Verbindung 1 mit den Gewichten des neuronalen Netzes verglichen, so wie es auch im Schritt der Netzwerkabfrage bei einem Simulationsexperiment durchgeführt wird. Die Trainingsdatensätze waren identisch (anfrageorientierte Trainingsdatensatzauswahl), wobei das Training in einem Fall unüberwacht und im anderen Fall überwacht durchgeführt wurde. Die Abbildung zeigt die Neuronen, wobei die Farben der Neuronen deren Ähnlichkeit mit dem Strukturcode der Anfragestruktur ausdrücken. Die ähnlichsten Neuronen sind mit einem Kreis markiert. Diese Neuronen sind die Gewinnerneuronen aus denen das simulierte Spektrum stammt. Es fällt auf, daß der Wertebereich der *rms*-Werte von 0.009 bis 0.035 geht. Der *rms*-Wert zwischen zwei Strukturcodes in Abbildung 3-34 liegt mit 0.028 somit eher im oberen Grenzbereich.

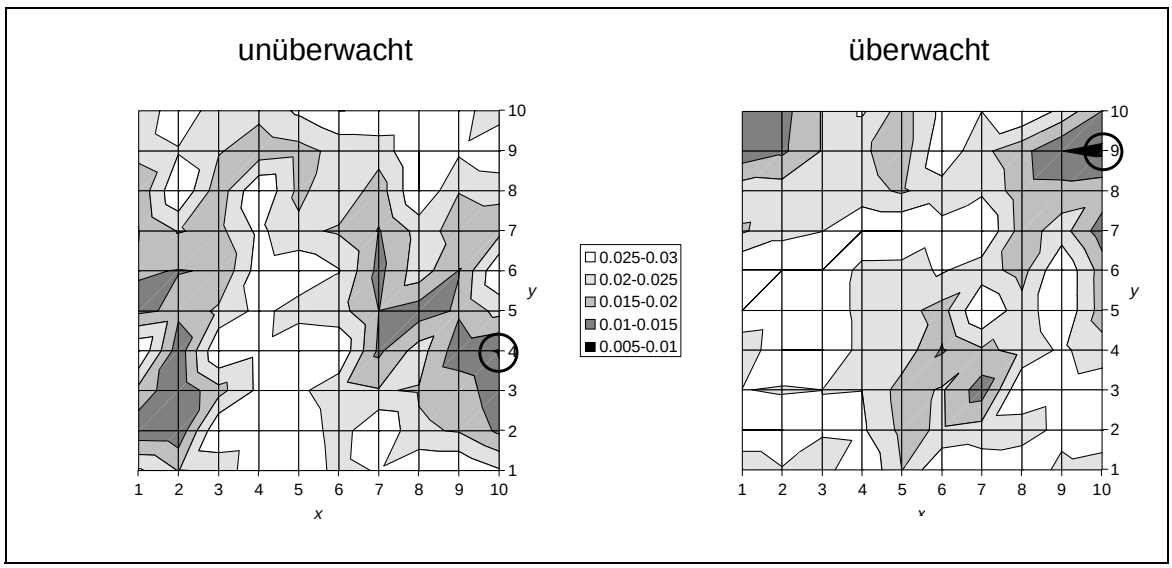


Abb. 3-35: *rms*-Werte zwischen dem Anfragestrukturcode und den Neuronengewichten

Daß sich in so einem spektroskopisch und strukturell enorm sensiblen Bereich dennoch derart gute Vorhersagen treffen lassen, kann als Hinweis auf die große Leistungsfähigkeit dieser Spektrenvorhersagemethode gewertet werden. Auch hier wäre natürlich die Qualität der Vorhersage durch eine bessere Abdeckung des zugrundeliegenden Datenbereichs zu erreichen. Da sich an der Datenbasis kurzfristig nichts ändern läßt, muß auf der Ebene der Erkennung und Abbildung von Strukturcodeähnlichkeiten nach Verbesserungsmöglichkeiten gesucht werden (vgl. Kap. 5.1).

3.4 Zusammenfassung der Anwendungsbeispiele

Die drei Beispiele in den Kapiteln 3.1 - 3.3 zeigen typische Anwendungsmöglichkeiten für die Spektrensimulation. Eine oder mehrere Substanzen sind IR-spektroskopisch analysiert worden. Eine einfache Substanzidentifikation durch den Vergleich mit experimentellen Referenzspektren ist nicht möglich, da die Datenbank keine identischen Spektren enthält. Für eine Reihe möglicher Kandidaten werden die entsprechenden Spektren simuliert und mit den experimentellen Spektren verglichen. Für die Erstellung dieser Kandidatenliste gibt es mehrere Möglichkeiten:

- Der Benutzer ist sich bezüglich der analysierten Substanz sehr sicher und benötigt nur eine Bestätigung (vgl. Kap. 3.1)
- Der Benutzer kennt die Edukte und die Reaktionsbedingungen, so daß er das Auftreten von einem oder mehreren Nebenprodukten für möglich hält
- Das bei einer Datenbanksuche als ähnlichste Verbindung gefundene Molekül wird als Startmolekül genommen und es werden Simulationen für Moleküle durchgeführt, die zu dem Startmolekül strukturelle Ähnlichkeit aufweisen (vgl. Kap. 3.2)
- Es liegt keine detaillierte Information vor, so daß eine größere Anzahl von Kandidaten in Frage kommen. Aus der Anfangsinformation, z.B. der Kenntnis des Edukts eines Abbauprozesses, wird mit einem weiteren System, z.B. einem Reaktionsvorhersagesystem, eine Liste potentieller Kandidaten erzeugt (vgl. Kap. 3.3)

Bei dem anschließenden Vergleich zur Identifikation ist es nicht unbedingt notwendig, daß der Korrelationskoeffizient r zwischen simuliertem und experimentellem Spektrum sehr hoch ist. Bei manchen Beispielen, gerade wenn die Anfragestruktur nicht besonders gut durch den Trainingsdatensatz repräsentiert wurde, konnte die Substanz trotzdem identifiziert werden, indem das am wenigsten unähnliche Spektrum als Lösung präsentiert wurde.

Allgemein gehört Vorhersageverfahren sicherlich die Zukunft, da sie die möglicherweise sehr zeitaufwendigen und kostenintensiven Synthesen von Referenzverbindungen überflüssig machen können. Zum aktuellen Zeitpunkt kann dieses Identifikationssystem nicht vollautomatisch und unbeaufsichtigt arbeiten. Es kann jedoch eine wertvolle Entscheidungshilfe bei Interpretations- und Identifikationsfragen bieten.

4 Infrarotspektrenvorhersage über das Internet

Im Rahmen eines durch den DFN-Verein [83] geförderten Projektes „Telekooperation in der Spektroskopie - TeleSpek" wurde die in den vorhergehenden Kapiteln beschriebene Methode über Internet zur Verfügung gestellt.[84][85][86] Ziel war es, neben der Einrichtung eines Spektren-Diskussionsforums mit Hochschul- und Industriepartnern, einen schnellen Zugang zu Referenzspektren zu bieten.

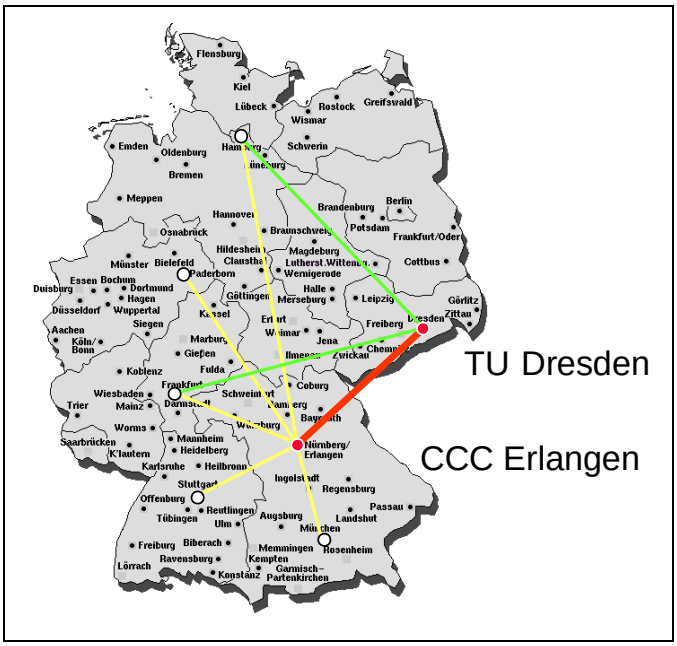


Abb. 4-1: Aufbau eines internetbasierten Spektrenvorhersagesystems

In enger Zusammenarbeit mit dem Arbeitskreis von Prof. Salzer am Institut für analytische Chemie der TU Dresden wurden die Methoden online getestet und die Benutzerfreundlichkeit verbessert. Die TeleSpek-Internetseiten bieten Information über die entwickelten Methoden der Strukturcodierung und der Spektrenvorhersage. Die einzelnen Schritte der Methode, beispielsweise das Training des neuronalen Netzes, werden detailliert erläutert. Weiterhin hat der Benutzer die Möglichkeit, sich in eine Email-Liste einzutragen, um über Neuerungen auf diesen Seiten informiert zu werden. Die TeleSpek-Einstiegsseite unter http://www2.ccc.uni-erlangen.de/IR/ ist in nachfolgender Abbildung dargestellt:

Abb. 4-2: TeleSpek Startseite (http://www2.ccc.uni-erlangen.de/IR/)

Auf der Startseite befinden sich eine Reihe weiterer Verweise. Durch einen Mausklick auf die Verweise gelangt man zu den Seiten, auf welchen die Methoden näher erläutert werden oder zu den interaktiven Seiten. Bei den interaktiven Seiten besteht die Möglichkeit, selbständig ein Simulationsexperiment zu starten. Dazu ist es notwendig, eine Molekülstruktur einzugeben. Die Struktur wird in ein Eingabeformular auf der Internetseite in Form eines SMILES-Strings [87][88] (**S**implified **M**olecular **I**nput **L**ine **S**pecification) eingegeben. Bei SMILES handelt es sich um eine ASCII-Codierung der Konnektivität und der Stereochemie des Moleküls. Durch die Beschränkung auf ASCII-Zeichen ist diese Form der Strukturcodierung unter anderem sehr gut zur Eingabe in Internetseiten geeignet. Atome werden in Form ihrer chemischen Elementsymbole beschrieben, wobei die Einbindung der Atome in ein π-System durch Kleinschreibung der Elementsymbole ausgedrückt wird. Verzweigungen im Molekülbau werden durch das Setzen von Klammern, Ringsysteme durch die Verwendung von Zahlenindizes beschrieben. Implizite Wasserstoffatome werden nicht notiert, wobei betont werden muß, daß

es für dasselbe Molekül mehrere syntaktisch korrekte SMILES-Codes geben kann. In nachfolgender Abbildung sind einige Beispiel für Molekülstrukturen und deren SMILES-Codes dargestellt:

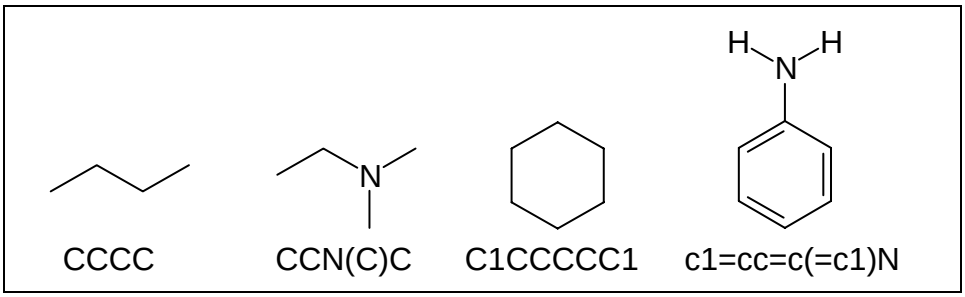


Abb. 4-3: SMILES-Strings für Butan, N,N-Dimethyl-N-Ethylamin, Cyclohexan und Anilin

Für einfach gebaute Moleküle ist es sehr gut möglich, sich den Code zu überlegen. Für komplexere Strukturen wird der SMILES-Code leicht unübersichtlich und es empfiehlt sich die Benutzung eines graphischen Struktureingabeprogrammes. Zur Eingabe des SMILES-Codes in die TeleSpek-Seiten kann der Benutzer entweder ein lokal installiertes Programm verwenden oder durch einen Druck auf den Knopf „Create Molecule" den systemeigenen Editor[1)] starten:

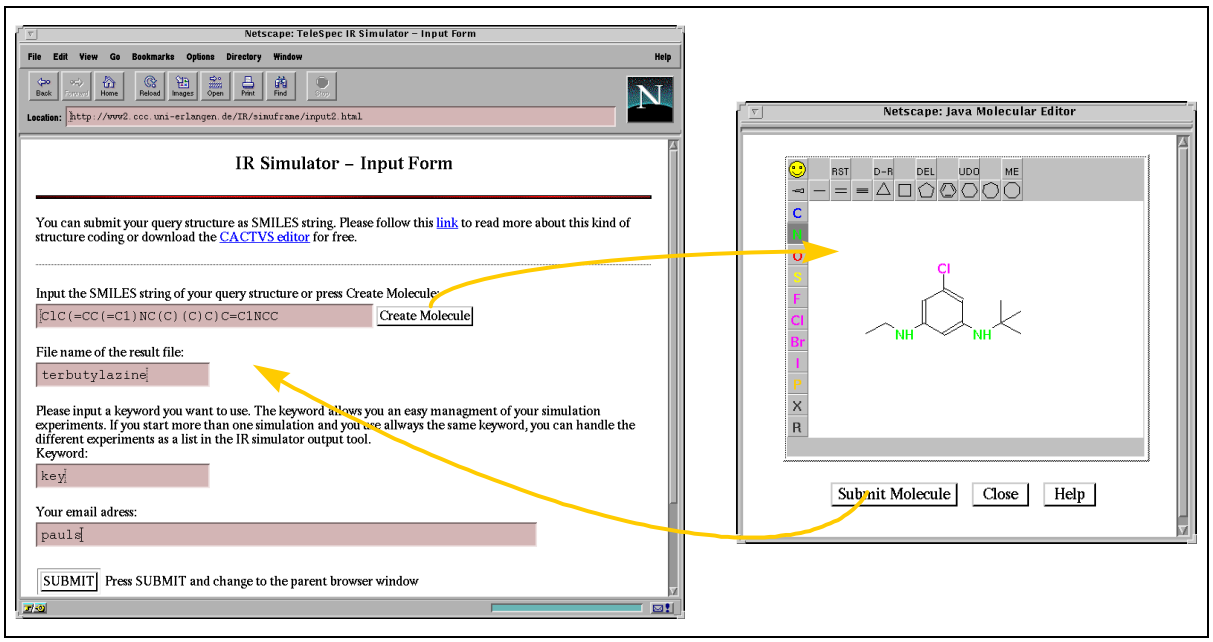


Abb. 4-4: TeleSpek-Formular zur Eingabe des SMILES-Codes (links). Mit dem Strukturedtor (rechts) kann eine Strukturformel gezeichnet werden. Nach einem Klick auf den Knopf „Submit Molecule" im Editor-Fenster erscheint der entsprechende SMILES-Code im Eingabeformular.

[1)] Der Editor wurde freundlicherweise von Dr. Peter Ertl von der Novartis Crop Protection AG in Basel zur Verfügung gestellt

Im Editor kann der Benutzer per Maus eine Strukturformel zeichnen. Nach einem Klick auf den „SUBMIT“-Knopf im Moleküleditor erscheint der entsprechende SMILES-String in der Zeile des Eingabeformulars. Hier wird der Benutzer weiterhin um die Eingabe eines Dateinamens gebeten unter welchem das Experiment auf dem TeleSpek-Zentralrechner für 14 Tage gespeichert wird. Zudem kann der Benutzer ein selbstgewähltes Keyword eingeben. Hier ist es sinnvoll, für alle Experimente dasselbe Keyword zu verwenden, da so bei einem späteren Besuch der Seiten alle gestarteten Experimente als Liste abgerufen werden können. In das letzte Eingabefeld soll die Email-Adresse eingegeben werden, um bei fehlgeschlagenen Experimenten, z.B. durch die Eingabe eines falschen SMILES-Codes oder Simulationen, die zu unsinnigen Ergebnissen führen, z.B. wenn das Anfragemolekül deutlich außerhalb des Datenbereichs der Trainingsdaten liegt, Rücksprache mit dem Benutzer halten zu können. Nach einem Maus-Klick auf „Submit“ muß der Benutzer dann wieder zum Simulator-Hauptfenster zurückkehren (Abb. 4-5):

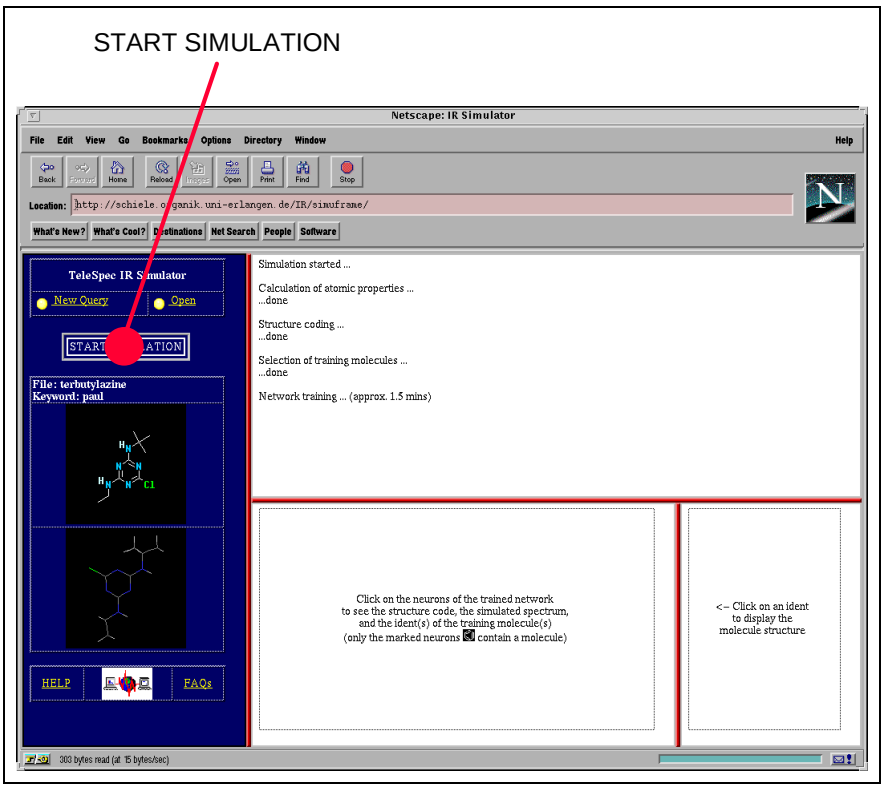


Abb. 4-5: Interaktiver IR-Simulator in TeleSpek

Auf der linken Seite ist im oberen Teil des Rahmens ein Molekülbild der eingegebenen Struktur dargestellt. Darunter befindet sich ein rotierbares Stäbchenmodell des Moleküls[1)] .

[1)]ChemSymphony JAVA Applet der Firma Cherwell Scientific

Nach einem Maus-Klick auf „START” wird das Simulationsexperiment gestartet. Im rechten oberen Fenster erscheinen die Meldungen, welche Schritte der Simulation gerade abgearbeitet werden: Berechnung physikochemischer Atomeigenschaften, Strukturcodierung, Auswahl des Trainingsdatensatzes und das Netztraining. Stellt das System bei der Auswahl des Trainingsdatensatzes fest, daß die Anfragestruktur in der SpecInfo-Datenbank enthalten ist, so ist keine Simulation notwendig und das Simulationsexperiment wird abgebrochen. Der Benutzer wird über das Vorhandensein seiner Anfragestruktur in der Datenbank informiert. Läuft das Simulationsexperiment weiter, so wird nach etwa 90 s das Ergebnis dargestellt (Abb. 4-6):

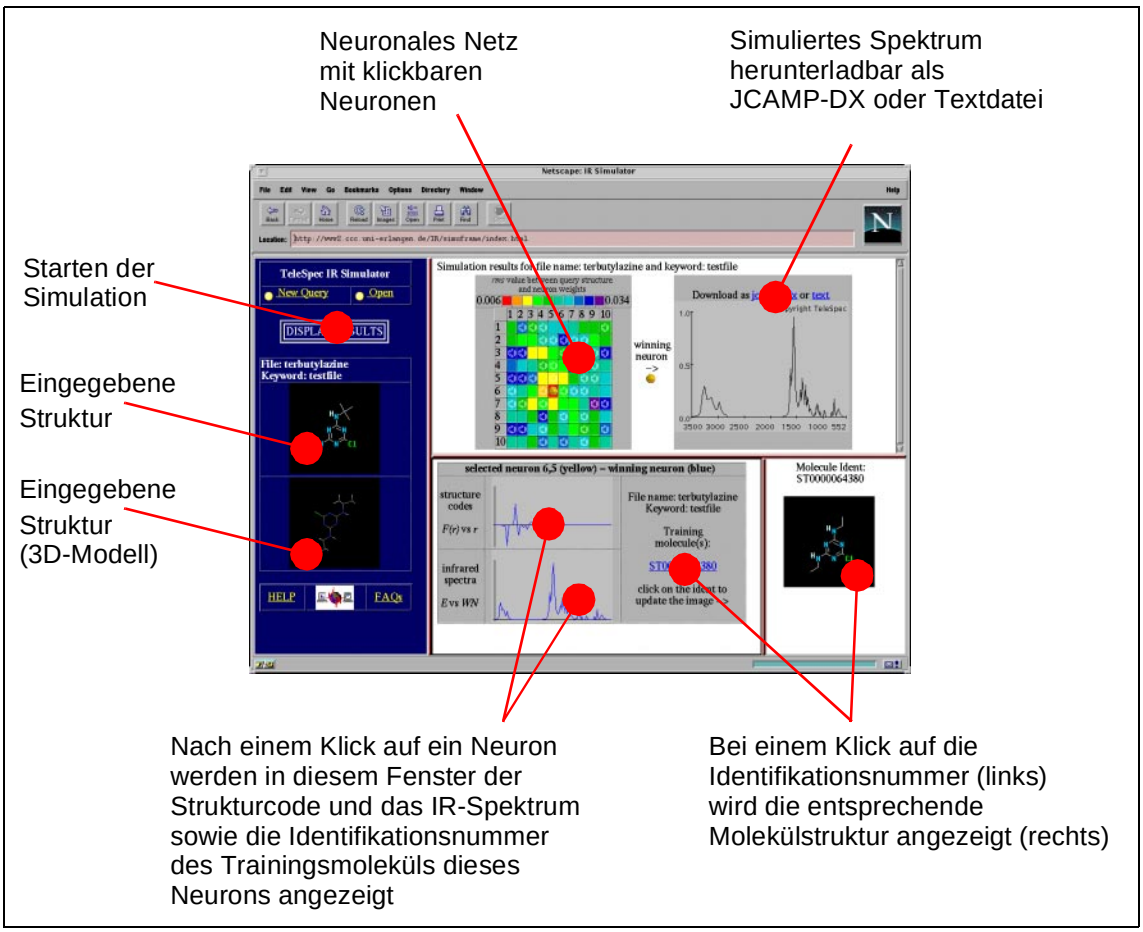


Abb. 4-6: Möglichkeiten zur Analyse des Simulationsexperiments

Im rechten oberen Fenster erscheint das simulierte IR-Spektrum, welches als JCAMP-DX [89] oder Textdatei heruntergeladen werden kann. Daneben ist das trainierte neuronale Netz abgebildet, wobei die verschiedenen Neuronen entsprechend ihrer Ähnlichkeit zum Strukturcode der Anfragestruktur eingefärbt sind. Weiterhin sind diejenigen Neuronen, die im Training mit einem Trainingsmolekül belegt wurden, mit einem Benzolring markiert. Bei einem Klick auf ein Neuron werden im darunterliegenden Fenster der Strukturcode und das

IR-Spektrum dieses Neurons sowie des Gewinnerneurons dargestellt. Wurde das angeklickte Neuron im Training mit einem Molekül belegt, so wird zusätzlich die Identifikationsnummer dieses Moleküls angezeigt. Bei einem Klick auf die Identifikationsnummer wird die dazugehörige Struktur dargestellt.

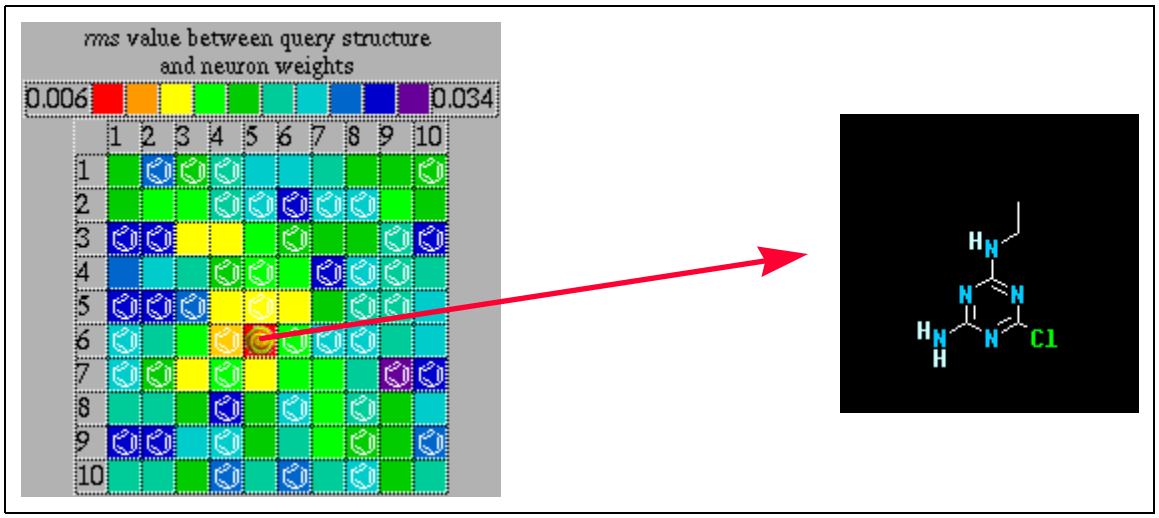


Abb. 4-7: Interaktive Analyse des neuronalen Netzes. Bei einem Maus-Klick auf ein Neuron wird die entsprechende Trainingsstruktur angezeigt.

So hat der Benutzer die Möglichkeit zu analysieren, welche Moleküle zum Training verwendet wurden und ob diese Moleküle eine Ähnlichkeit mit der Anfragestruktur aufweisen. Die Ähnlichkeit zwischen Anfragestruktur und Trainingsmolekülen kann einen Hinweis darauf geben, inwieweit eine qualitativ hochwertige Spektrensimulation zu erwarten ist oder nicht. Zusammenfassend muß die Förderung dieses Projekts als sehr positives Signal gesehen werden. Zum einen, da nun eine breite wissenschaftliche Öffentlichkeit Zugang zur Methode hat und zum anderen, weil die Methode durch die Nutzung und die Anforderungen und Rückmeldungen der Benutzer weiter verbessert werden kann. Ein Teil der in dieser Arbeit vorgestellten Simulationsbeispiele können über die TeleSpek-Internetseiten abgerufen werden. Die Keywords für die verschiedenen Experimente sind in nachfolgender Tabelle aufgeführt:

Tab. 4-1: TeleSpek Keywords

Kurzbezeichnung des Experiments	In Kapitel	TeleSpek-Keyword
Identifikation von N,N-Dimethylanilin-N-Oxid	3.1	nndma
Identifikation eines Ameisen-Spurpheromens	3.2	pheromon
Moleküle eines repräsentativen Datensatzes	2.6	dresden

5 Ansätze zur Weiterentwicklung der Methode

In den einzelnen Kapiteln wurden bereits Möglichkeiten zur Verbesserung der Methode erwähnt. Diese sollen nun zusammengefaßt und näher erläutert werden.

5.1 Verbesserung der Strukturcodierung

Im Kapitel zur Strukturcodierung wurde des öfteren darauf hingewiesen, daß es ein Hauptziel war, eine Form der Strukturbeschreibung zu finden, bei der möglichst viel infrarotrelevante Strukturinformation erhalten bleibt und durch den Strukturcode widergespiegelt wird. Ist diese Voraussetzung erfüllt, so besteht zwischen dem Strukturcode eines Moleküls und dem Infrarotspektrum des Moleküls, das im Prinzip nichts anderes als eine spezielle Form der Strukturbeschreibung ist, ein enger Zusammenhang. Erst dann besteht die Grundlage, den Zusammenhang zwischen Infrarotspektrum und Strukturcode mittels eines neuronalen Netzes zu modellieren. Letztendlich läßt sich das Problem auf eine ausreichend genaue Abbildung von Ähnlichkeiten, nichts anderes ist die Hauptfunktion eines neuronalen Netzes, reduzieren. Durch Korrelationsuntersuchungen zwischen Ähnlichkeiten im Strukturcode und spektralen Ähnlichkeiten kann überprüft werden, inwieweit der Strukturcode molekulare Eigenschaften, in diesem Fall das Absorptionsverhalten des Moleküls im infraroten Spektralbereich, beschreibt.[90] Im Idealfall sollten also Moleküle mit ähnlichen Strukturcodes auch ähnliche Infrarotspektren und Moleküle mit unterschiedlichen Strukturcodes auch unterschiedliche Infrarotspektren haben. Läge ein derartiger linearer Zusammenhang zwischen den Ähnlichkeiten von Strukturcodes und den Ähnlichkeiten von Infrarotspektren vor, so würde dies zu einer Situation führen, wie sie in Abbildung 5-1 dargestellt ist.

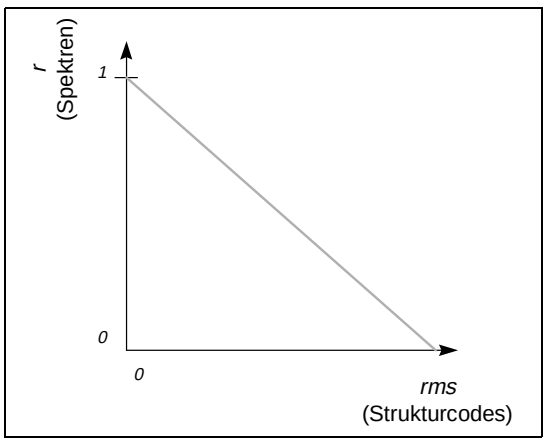


Abb. 5-1: Zusammenhang zwischen Vergleichen von Strukturcodes (*rms*) und Vergleichen von Infrarotspektren (*r*). Beim Vergleich von Molekülen entsprechen sich die jeweiligen Ähnlichkeiten von Strukturcodes und Spektren. (Ideales Verhalten)

Die Punkte für die Vergleiche der Strukturcode- und Spektrenpaare würden auf der diagonal verlaufenden Linie liegen. Ein derartiges Verhalten von Strukturcode- und Spektrenähnlichkeiten wäre zwar eine denkbar günstige Voraussetzung für Strukturcode-Spektren-Korrelationsuntersuchungen, ist jedoch für den Fall der Spektrenvorhersage nicht zwingend notwendig. Bei der Spektrenvorhersage wird der Strukturcode des Anfragemoleküls mit dem Eingabe-(Struktur)-Teil des neuronalen Netzes verglichen. Aus der Ausgabeschicht des dem Strukturcode der Anfragestruktur ähnlichsten Neurons wird das simulierte Spektrum ausgegeben. Grundvoraussetzung für das Funktionieren dieser Methode zur Spektrenvorhersage ist, daß Moleküle mit ähnlichen Strukturcodes auch ähnliche Infrarotspektren besitzen. Die Umkehrung dieser Voraussetzung, nämlich, daß Moleküle mit ähnlichen Infrarotspektren auch ähnliche Strukturcodes besitzen, würde die Vorhersagbarkeit der Simulationsqualität zwar erleichtern, ist für eine erfolgreiche Simulation aber nicht zwingend notwendig. Abbildung 5-2 soll diesen Zusammenhang noch einmal erläutern.

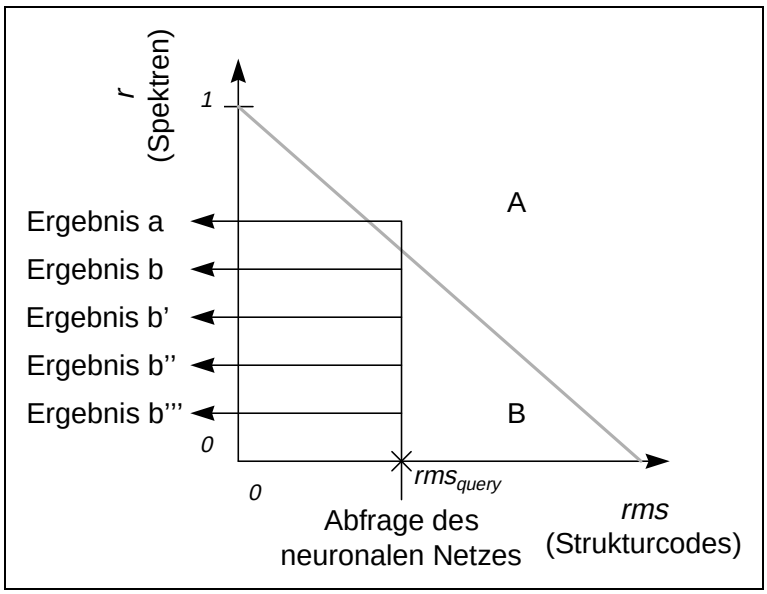


Abb. 5-2: Zusammenhang zwischen Ähnlichkeiten von Strukturcodes und Ähnlichkeiten von Infrarotspektren - Einfluß der Population verschiedener Bereiche A und B auf die Vorhersagequalität. Der *rms*-Wert zwischen Anfragestrukturcode und Neuronengewichten ist rms_{query}. Wenn Bereich B belegt ist, so liegt die Qualität der Vorhersage zwischen b und b'''. Wenn Bereich A belegt ist, so liegt die Qualität der Vorhersage bei a oder höher.

Für einen Anfragestrukturcode wird ein ähnlichstes Neuron bestimmt. Der entsprechende *rms*-Wert zwischen Anfragestrukturcode und Neuronengewichten ist rms_{query}. Ist das Verhalten des Strukturcodes derart, daß nur Bereich A belegt ist, so wird das vorhergesagte Infrarotspektrum einen Korrelationskoeffizienten mit dem entsprechenden experimentellen Spektrum aufweisen, der in etwa Ergebnis A entspricht; möglicherweise sogar besser ($r \geq$

Ergebnis a). Ist das Verhalten des Strukturcodes so, daß Bereich B belegt ist, so ist das Simulationsergebnis nicht unbedingt schlechter. Eine sinnvolle Vorhersage für eine Mindestqualität der Vorhersage ist jedoch nicht mehr zu treffen. Das Qualität des Ergebnisses ist vielmehr zufällig und liegt zwischen Ergebnis b und b’’’. Für eine gut geeignete Strukturcodierung wäre es somit wünschenswert, wenn die Vergleichspaare im Bereich A zu liegen kommen und Bereich B möglichst unbesetzt bleibt. Um in möglichst vielen Fällen eine gute Simulation zu erreichen, wäre es zudem günstig, wenn eine Häufung der Vergleichspaare im Bereich C zu beobachten wäre (vgl. Abb. 5-3).

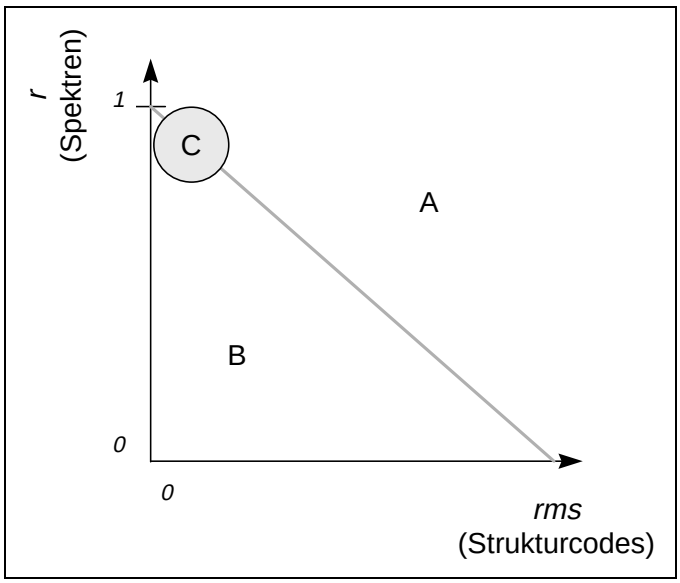


Abb. 5-3: Zusammenhang zwischen Strukturcode und Infrarotspektrum - Bedeutung der Population verschiedener Bereiche

Um zu untersuchen, wie das Verhalten der Radialcodierung (128 Werte) mit $A_i = q_{tot}$ bei derartigen Korrelationsexperimenten ist, wurde jedes einzelne ungeladene H, C, N, O, Hal-Molekül (9850 Moleküle) der SpecInfo IR-Datenbank mit allen anderen Molekülen der Datenbank verglichen (48506325 Vergleiche). Dabei wurde jeweils der *rms*-Wert zwischen den Strukturcodes und der Korrelationskoeffizient *r* zwischen den Infrarotspektren bestimmt (vgl. Abb. 5-4).

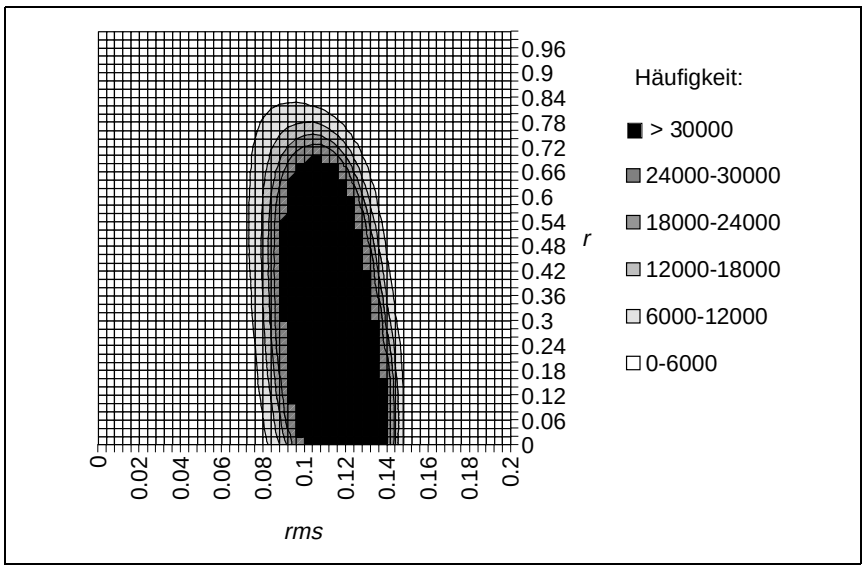


Abb. 5-4: Korrelationsdiagramm aller paarweisen Vergleiche aller 9850 ungeladenen H, C, N, O, Hal-Moleküle der SpecInfo IR-Datenbank. Es wurde jeweils der *rms*-Wert zwischen den Strukturcodes und der Korrelationskoeffizient *r* zwischen den Infrarotspektren bestimmt. Die Beschreibung der Molekülstrukturen erfolgte in 128 Radialcodewerten mit $A_i = q_{tot}$.

Hier ist zu erkennen, daß es Molekülpaare mit ähnlichen Spektren ($r > 0.8$) und ähnlichen Strukturcodes gibt ($rms < 0.1$). Allgemein ist die Abbildung, aufgrund der hohen Anzahl von Einzelpunkten und der daraus resultierenden niedrigen Auflösung von Bereichen mit geringer Belegung, jedoch wenig aussagekräftig. Aus diesem Grund wurde in einem sehr ähnlich gelagerten Experiment für jede ungeladene H, C, N, O, Hal - Verbindung der SpecInfo IR-Datenbank eine anfragestrukturorientierte Spektrenvorhersage durchgeführt. Dieses Experiment wurde bereits in Kapitel 2.5 näher beschrieben. Auch hier wurden die Molekülstrukturen in 128 Radialcodewerte mit $A_i = q_{tot}$ transformiert. Für jede Simulation wurde der *rms*-Wert zwischen dem Anfragestrukturcode und dem Gewinnerneuron (a), dem ähnlichsten Molekül des Trainingsdatensatzes (b) sowie der gemittelte *rms*-Wert zwischen dem Anfragestrukturcode und allen Molekülen des Trainingsdatensatzes (c) gespeichert. Diese drei Werte wurden dann jeweils gegen den Korrelationskoeffizienten *r* zwischen simuliertem und experimentellem Spektrum aufgetragen (vgl. Abb. 5-5).

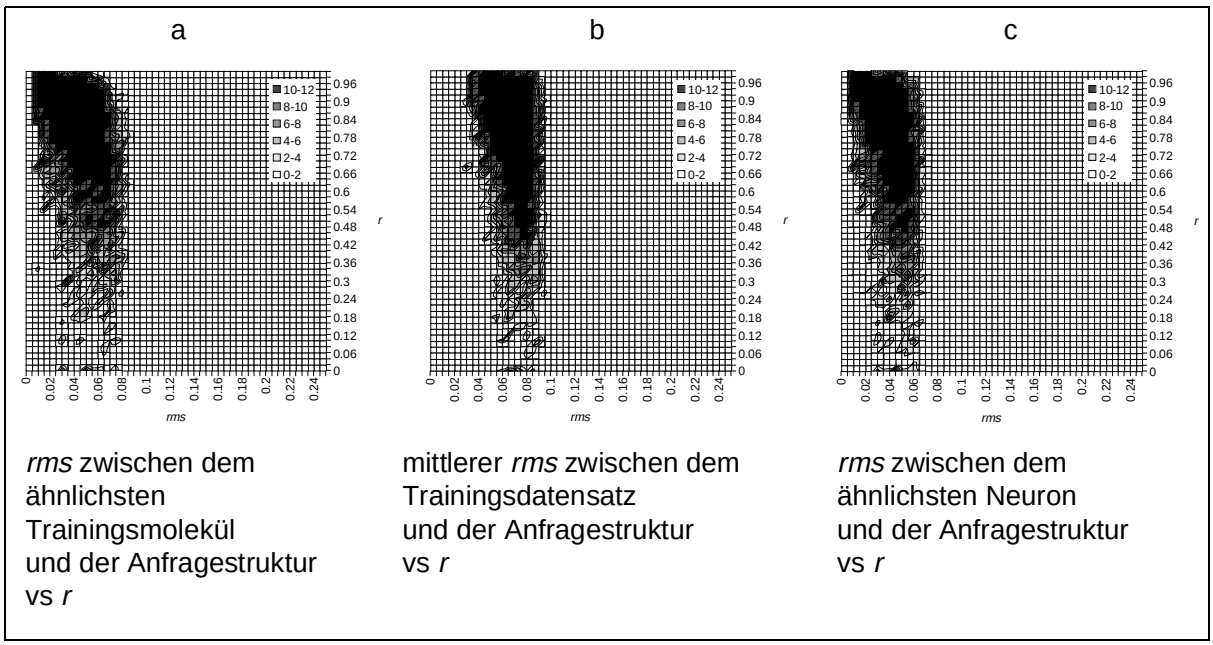


Abb. 5-5: *rms*-Werte vs Korrelationskoeffizienten *r* für anfrageorientierte Simulationen aller 9850 H, C, N, O, Hal -Verbindungen der SpecInfo IR-Datenbank. Die Beschreibung der Molekülstrukturen erfolgte in 128 Radialcodewerten mit $A_i = q_{tot}$.

Diese drei Experimente (vgl. Abb. 5-5) zeigen ein ähnliches Verhalten bei der Auftragung der *rms*-Werte der Strukturcodes vs den Korrelationskoeffizienten *r* der Spektren wie beim vorherigen Experiment (vgl. Abb. 5-4). Als sehr positiv kann gewertet werden, daß im Bereich sehr ähnlicher Strukturcodes und sehr ähnlicher Spektren (Bereich C in Abbildung 5-3) eine Häufung der Datenpunkte zu beobachten ist. Daraus läßt sich folgern, daß eine Vielzahl der Moleküle der SpecInfo-Infrarotdatenbank in einem datenmäßig gut abgedeckten Bereich liegen und somit für diese Moleküle gute Vorhersageergebnisse zu erwarten sind. Als negativ ist zu bewerten, daß der Bereich mit relativ ähnlichen Strukturcodes aber variierender Spektrenähnlichkeit (Bereich B in Abbildung 5-3) ebenfalls mit Datenpunkten belegt ist. Liegt für ein Anfragemolekül der *rms*-Wert zwischen dem Anfragestrukturcode und dem ähnlichsten Neuron in diesem Bereich, so fällt das Simulationsergebnis möglicherweise schlechter aus als es aufgrund der Datenbasis nötig wäre. Anders ausgedrückt: Für diese bestimmte Anfragestruktur gibt es Neuronen, die im Training mit Molekülen belegt wurden, welche ähnlichere Spektren haben. Zufälligerweise wird jedoch ein Neuron mit einem Molekül ausgewählt, dessen Infrarotspektrum eine geringere Ähnlichkeit mit dem experimentellen Spektrum der Anfragestruktur aufweist und somit zu einem schlechteren Simulationsergebnis führt. Eine Verbesserung dieses Verhaltens könnte durch eine weitere Erhöhung des infrarotrelevanten Informationsgehalts des Strukturcodes erzielt werden. Eine wesentliche Verbesserung der Codierung durch eine weitere Variation der Atomeigenschaften ist kaum mehr zu erreichen.

Bei den Experimenten in Kapitel 2.4.1 hatte sich jedoch gezeigt, daß die Skalierungsmethode einen sehr großen Einfluß auf die Simulationsergebnisse hat. Ebenso wäre zu überprüfen, ob beispielsweise durch eine explizite Berücksichtigung von Kraftkonstanten [91][92][93] der vorhandenen Bindungen eine Verbesserung zu erzielen ist. Bei der aktuellen Form der Codierung werden diese Kraftkonstanten nur implizit durch die Beschreibung der Atomabstände und der Atomladungen ausgedrückt. Bei der Codierung wird die Codefunktion für jedes mögliche Atompaar des Moleküls berechnet, unabhängig davon, ob zwischen den Atomen eine Bindung besteht oder nicht. Möglicherweise wäre auch schon durch eine Beschränkung auf gebundene Atompaare oder Atompaare bis zu einer Nachbarschaftsbeziehung von vier Bindungsschritten ein positiver Effekt zu beobachten. Die Berücksichtigung von Nachbarschaftsbeziehungen bis zu vier Bindungsschritten erscheint sinnvoll, da so die vorkommenden Diederwinkel in die Codierung miteinfließen würden.

Um Simulationsfehlern, wie bei der β-Alanin-Simulation in Kapitel 2.6 zu beobachten war, entgegenzuwirken, wäre auch eine Kombination der Codierungsmethode mit einer fragmentbasierten Codierung denkbar. Vorausgesetzt, die entsprechenden Strukturen sind in der Datenbank korrekt abgelegt, könnten nun solche Sonderfälle, wie diese Aminosäure, erfaßt und die Simulation entsprechend optimiert werden. Dies birgt natürlich auch Probleme. Die erfolgreiche Simulation für Cyanazin (vgl. Abb. 2-55 in Kap. 2.4.2.3) zeigt wegen des captodativen Effekts keine Nitrilbande. Wäre die Nitrilfunktion des Anfragemoleküls durch eine fragmentbasierte Strukturcodierung erfaßt worden, wäre eine Nitrilbande simuliert worden. Das Simulationsergebnis hätte sich damit verschlechtert.

5.2 Vorhersage der Simulationsqualität

Bei den Beispielen, die in dieser Arbeit vorgestellt wurden, konnte die Qualität der Vorhersage durch den Vergleich von simuliertem und experimentellem Spektrum bestimmt werden. Wie jedoch bereits erwähnt wurde, kommt diese Methode genau dann zum Einsatz, wenn eben kein experimentelles Referenzspektrum vorhanden ist. Da die Simulationsmethode datenbasiert ist, ist die Qualität der Vorhersage davon abhängig, ob die zugrundeliegende Datenbasis Moleküle enthält, die der Anfragestruktur ähnlich sind. Dies kann von Anfragestruktur zu Anfragestruktur sehr unterschiedlich sein. Der Benutzer, der sich dieser Vorhersagemethode zur Simulation eines Referenzspektrums bedient, benötigt somit ein Maß, um die Simulationsgüte abzuschätzen. Die Information aus der sich dieses Maß ableiten kann, ist der Vergleich von Anfragestruktur und Trainingsstrukturen. Basierend auf der Ähnlichkeit zwischen den Strukturcodes der Anfragestruktur und dem ähnlichsten Molekül des Trainingsdatensatzes wurde in Kapitel 2.5 ein Verfahren entwickelt, das in begrenztem Umfang eine

Vorhersage der Simulationsqualität erlaubt. Die so ermittelten Werte für den Korrelationskoeffizienten *r* zwischen simuliertem und experimentellem Spektrum stellen jedoch untere Grenzwerte dar. Bei den durchgeführten Simulationsexperimenten waren die erzielten Ergebnisse deutlich besser als es nach dieser Vorhersage zu erwarten gewesen wären. Unabhängig davon besteht für den Benutzer die Möglichkeit, das trainierte neuronale Netz zu untersuchen (vgl. z.B. Abb. 2-68). Dabei ist von Interesse, welche Moleküle auf das Gewinnerneuron oder auf benachbarte Neuronen zugeordnet wurden und somit starken Einfluß auf das simulierte Spektrum haben. Dies setzt allerdings spektroskopisch-chemischen Sachverstand beim Benutzer voraus. Hier wäre zu untersuchen, inwieweit sich diese Methode durch einen intelligenten Substruktur-Suchalgorithmus automatisieren ließe, um in Kombination mit obigem Vergleich von Strukturcodes, eine zuverlässige und realistische Vorhersage der Simulationsqualität zu erreichen.

5.3 Vorhersage von Gemischspektren

Prinzipiell sollte es mit der vorgestellten Methode möglich sein, Gemischspektren vorherzusagen. Der entsprechende Strukturcode könnte dabei eine Kombination der beteiligten Komponenten sein, die entsprechend ihres prozentualen Anteils unterschiedlich gewichtet werden. Eine Umkehrung dieser Methode zur Vorhersage der Komponentenanteile eines Gemisches anhand des Gemischspektrums wäre eine weitere interessante Fragestellung.

5.4 Erweiterung auf andere spektroskopischen Methoden

Da die Strukturcodierung sehr gut die dreidimensionale Struktur eines Moleküls erfaßt, wäre es durchaus denkbar, die Methode der Spektrenvorhersage auf spektroskopische Methoden anzuwenden, die sehr stark von der dreidimensionalen Molekülstruktur abhängen, wie z.B. die ^{1}H-NMR-Spektroskopie. Durch die Variation der Atomeigenschaften könnte der Informationsgehalt des Strukturcodes den Anforderungen der jeweiligen spektroskopischen Methode angepaßt werden.

6 Zusammenfassung

Durch den raschen Fortschritt der Technologien im Bereich der Spektrometrie und der Datenverarbeitung ist es möglich geworden, große Datenmengen aufzunehmen und zu speichern. Der begrenzende Faktor ist in den meisten Fällen nun das Zeitfenster der spektroskopischen Methode selbst und nicht mehr die langsam fließenden Datenkanäle. Das eigentliche Problem hat sich damit vom Bereich der Datenakquisition in den Bereich der Datenaufarbeitung verschoben. Es ist ein Bedarf an Methoden entstanden, die bei der Auswertung und Interpretation der Meßergebnisse Unterstützung bieten. Gerade in der Spektroskopie existieren bereits eine Vielzahl experimenteller Daten und täglich werden neue erzeugt, aus denen sich wertvolle Information extrahieren läßt. Eine häufige Fragestellung in der infrarotspektroskopischen Analytik ist die Identifikation einer unbekannten Probe. Dazu wird in der Regel das gemessene Spektrum mit einem Referenzspektrum verglichen. Hier stehen jedoch einer Zahl von 16 Millionen bekannten chemischen Verbindungen nur etwa 100000 archivierte und zugängliche Infrarotspektren gegenüber. Durch dieses Mißverhältnis wird es oftmals vorkommen, daß ein Referenzspektrum für eine bestimmte Substanz nicht zur Verfügung steht. In diesem Zusammenhang wurde eine datenbasierte Vorhersagemethode für Infrarotspektren entwickelt, die einen schnellen Zugang zu benötigten Referenzspektren bietet. Grundlage dieser Arbeiten war die Entwicklung einer Transformation der dreidimensionalen Struktur eines Moleküls in einen Code konstanter Länge, wodurch es möglich wurde, den Zusammenhang zwischen 3D-Struktur und Infrarotspektrum mittels eines neuronalen Netzes zu modellieren. Ausgehend vom Strukturcode kann dann das entsprechende Infrarotspektrum vorhergesagt werden.

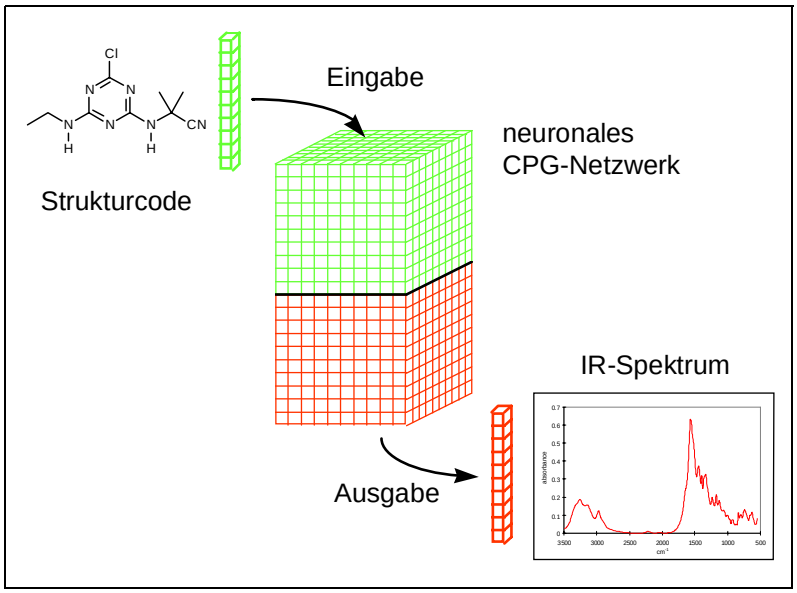


Abb. 6-1: Vorhersage des Infrarotspektrums aus der Molekülstruktur

Die Entwicklung dieser Methode und deren Anwendung auf experimentelle Fragestellungen war Inhalt dieser Arbeit. In diesem Kapitel sollen die Vorzüge und Einschränkungen zusammengefaßt werden.

6.1 Möglichkeiten der Methode

Die Verwendung neuronaler Netze ist eine elegante Methode zur Modellierung nichtlinearer Zusammenhänge, wie sie beispielsweise zwischen Molekülstruktur und Infrarotspektrum bestehen. Die Vorteile dieses Ansatzes sollen nun kurz erläutert werden:

- **Vorhersage von Spektren über den gesamten Bereich des mittleren Infrarot**
  Durch die Korrelation der codierten 3D-Struktur mit dem gesamten Infrarotspektrum ist es möglich, für eine Anfragestruktur das entsprechende Vollspektrum, einschließlich der charakteristischen Bandenmuster im Fingerprintbereich, vorherzusagen. Die Methode ist somit nicht auf die Vorhersage bestimmter Signale und Signallagen ausgehend von Strukturmerkmalen der Anfragestruktur beschränkt.

- **Induktives Lernverfahren**
  Die eingesetzten künstlichen neuronalen Netze bedienen sich eines induktiven Lernverfahrens, d.h., sie sind in der Lage, sich ihr Wissen selbständig aus einer Reihe von Beispielen abzuleiten. Dies hat den großen Vorteil, daß bei einem komplexen Sachverhalt, wie der Infrarotspektroskopie, zunächst kein *a priori* Wissen in das System gegeben werden muß.

- **Hohe Vorhersagegeschwindigkeit**
  Die Methode arbeitet sehr schnell. Innerhalb von 1.5 Minuten kann auf einer SGI ORIGIN 200 eine anfrageorientierte Standardsimulation einschließlich der Trainingsdatensatzauswahl und dem Netztraining durchgeführt werden. Die Vorhersagegeschwindigkeit ist dabei unabhängig von der Molekülgröße.

- **Keine Beschränkung für die Molekülgröße**
  Ebenso wie die Simulationsdauer ist die Vorhersagequalität weitgehend unabhängig von der Größe und Komplexität des Moleküls sowie der gebundenen Atome. Eine prinzipielle Beschränkung auf bestimmte Substanzen oder Verbindungsklassen besteht nicht.

6.2 Grenzen der Methode

Aus der Methode an sich und der verwendeten Datenbasis ergeben sich jedoch auch einige Einschränkungen und Nachteile:

- **Schlechte Repräsentation eines Anfragemoleküls durch die Moleküle des Trainingsdatensatzes**
  Die Vorhersagequalität hängt in hohem Maße davon ab, wie gut die Anfragestruktur durch die Moleküle des Trainingsdatensatzes repräsentiert wird. Sind im Trainingsdatensatz Moleküle enthalten, die der Anfragestruktur sehr ähnlich sind, so wird die Vorhersage in der Regel sehr gut ausfallen. Umgekehrt kann die Simulation auch für ein relativ einfach gebautes Molekül schlecht ausfallen, wenn keine ähnlichen Verbindungen in der Datenbank enthalten sind

- **Qualitativ minderwertige Trainingsspektren**
  Die Qualität des vorhergesagten Spektrums kann nur so gut sein, wie die Qualität der zugrundeliegenden Daten. Sind die Datenbankspektren beispielsweise verunreinigt oder wurden sie mit feuchtem KBr aufgenommen, so werden die simulierten Spektren ebenfalls die Signale der Verunreinigungen bzw. die typischen „Wasserbäuche" bei 3400 cm^{-1} enthalten, obwohl die Anfragestruktur keine OH-Gruppe hat.

- **Geringe Extrapolationsfähigkeiten neuronaler Netze**
  Die eingesetzten Counterpropagation-(CPG)-Netze haben sehr gute Interpolationseigenschaften. Bei Simulationsexperimenten ist oft zu beobachten, daß das im Simulationsschritt ausgewählte Gewinnerneuron im Training mit keinem Molekül belegt wurde. Allerdings wurden Nachbarneuronen belegt, die einzelne oft auch unterschiedliche Strukturmerkmale der Anfragestruktur enthalten. Das vorhergesagte Spektrum ist in einem solchen Fall eine Interpolation zwischen verschiedenen Spektren des Trainingsdatensatzes. Im Gegensatz dazu sind die Extrapolationsfähigkeiten eines CPG-Netzes eher als gering einzustufen. Aus dieser Tatsache heraus ergeben sich verschiedene Grenzen des Systems: Simulationen für Substanzen, die in gewisser Hinsicht ein Ende einer systematischen Reihe darstellen und damit nicht durch Interpolation aus ähnlichen Verbindungen abgeleitet werden können, werden von geringer Qualität sein. Eine solche Substanz ist beispielsweise Methan.

Ein Schwerpunkt gegen Ende dieser Arbeit war das DFN-Projekt „Telekooperation in der Spektroskopie - TeleSpek“. Im Rahmen dieses Projekts wurde die Methode zur Spektrenvorhersage über Internet zur Verfügung gestellt. Benutzer haben hier nun die Möglichkeit, interaktive Spektrensimulationsexperimente durchzuführen (vgl. Kap. 4). Seit Beginn des Projekts im November 1996 wurden bis Juni 1998 ca. 400 Spektrensimulationsexperimente von externen Benutzern durchgeführt.

Zusammenfassend kann bemerkt werden, daß sich die vorgestellte Methode zur Vorhersage von Infrarotspektren als sehr leistungsstark erwiesen hat. Die beiden eingesetzten Verfahren zur Codierung der molekularen 3D-Struktur, die 3D-MoRSE-Codierung und die Radialcodierung, bieten ein großes Potential für Untersuchungen, bei denen es gilt, Strukturinformation mit molekularen Eigenschaften zu korrelieren. Für die Methode zur Spektrenvorhersage und für die Verfahren zur Strukturcodierung ergeben sich, wie auch bereits in Kapitel 5 dargelegt wurde, eine Reihe von interessanten Ansätzen für weiterführende Forschungsarbeiten.

Literaturverzeichnis

[1] Sadtler Division Infrarotdatenbank, Bio-Rad Laboratories Ltd. Maylands Avenue, Hemel Hempstead, Hertfordshire HP2 7TD England

[2] SpecInfo Infrarotdatenbank, Chemical Concepts, Boschstr 12, D-69469 Weinheim

[3] Wedler, G.: *Lehrbuch der Physikalischen Chemie*; 3. Auflage, Verlag Chemie, Weinheim 1987, S. 544.

[4] Günzler, H.; Heise, H. M.: *IR-Spektroskopie - Eine Einführung*; 3. Auflage, Verlag Chemie, Weinheim 1996.

[5] Williams, D. H.; Fleming, I.: *Spektroskopische Methoden in der organischen Chemie*; Georg Thieme Verlag, Stuttgart 1968, S. 40.

[6] Otto, M.: *Analytische Chemie*, Verlag Chemie, Weinheim 1995.

[7] Naumer, H.; Heller, W.: *Untersuchungsmethoden in der Chemie*; Georg Thieme Verlag, Stuttgart 1985.

[8] Razinger, M.; Novic, M.: *Reduction of the Information Space for Data Collections, PCs for Chemists*; Zupan, J. Ed.; Elsevier: Amsterdam 1990; pp 89-103;

[9] Zupan, J.: *Algorithms for Chemists*, Wiley, New York 1989.

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

[11] Munk, M. E.; Madison, M. S.; Robb, E. W. Neural Network Models for Infrared Spectrum Interpretation *Microchim. Acta (Wien)* **1991**, *2*, 505-514.

[12] Huixiao, H.; Xinquan, X. Essessa: An Expert System for Elucidation of Structures. 1. Knowledge Base of Infrared Spectra and Analysis and Interpretation Programs *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 302-210.

[13] Weigel, U.-M.; Herges, R. Automatic Interpretation of Infrared Spectra: Recognition of Aromatic Substitution Patterns Using Neural Networks *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 723,731.

[14] Weigel, U.-M.; Herges, R. Simulation of infrared spectra using artificial neural networks based on semiempirical and empirical data *Anal. Chim. Acta*. **1996**, *331*, 63-74.

[15] Affolter, C.; Baumann, K.; Clerc, J. T.; Schriber, H.; Pretsch, E. Automatic Interpretation of Infrared Spectra *Microchim. Acta* **1997**, *14*, 143-147.

[16] Dubois, J. E.; Mathieu, G.; Peguet, P.; Panaye, A.; Doucet, J. P. Simulation of Infrared Spectra: An Infrared Spectral Simulation Program (SIRS) Which Uses DARC Topological Substructures *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 290-302.

[17] Clerc, J. T.; Terkovics, A. L. Versatile topological descriptor for quantitative structure/property studies *Anal. Chim. Acta* **1990**,*235*, 93-102.

[18] Debye, P. Zerstreuung von Röntgenstrahlen an amorphen Körpern, *Phys. Zeitschr.* **1927**, *31*, 135-141.

[19] Wierl, R. Elektronenbeugung und Molekülbau. *Ann. Phys. (Leipzig)* **1931**, *8*, 521-564.

[20] Soltzberg, L. J.; Wilkins, C. L., Molecular Transforms: A Potential Tool for Structure-Activity Studies. *J. Am. Chem. Soc.* **1977**, *99*, 439-443.

[21] 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.

[22] Selzer, P.; Schuur, J.; Gasteiger, J. Simulation of IR Spectra with Neural Networks Using the 3D-MoRSE Code. In *Software Development in Chemistry 10*; Gasteiger, J. Ed.; Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany, 1996; p. 293-302.

[23] Schuur, J.; Gasteiger, J. 3D-MoRSE Code - A New Method for Coding the 3D Structure of Molecules. In *Software Development in Chemistry 10*; Gasteiger, J. Ed.; Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany, 1996; p. 67-78.

[24] 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.

[25] 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.

[26] Náray-Szabó, G.; Harmat, V. The Molecular Transform as a Tool for Quantification of Molecular Similarity *communications in mathematical and computer chemistry, match, matcdy*, **1997**, *35*, 29-40.

[27] Csorvássy, I.; Tözsér, L. The molecular transform as a similarity measure, *J. Math. Chem.* **1993**, *13*, 343-357.

[28] Steinhauer, L.; Steinhauer, V.; Gasteiger, J. Obtaining the 3D Structure from Infrared Spectra of Organic Compounds Using Neural Networks. In *Software Development in Chemistry 10*; Gasteiger, J. Ed.; Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany, 1996; p. 315-322.

[29] 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.

[30] Hemmer, M. C. *Strukturvorhersage aus Infrarotspektren mit neuronalen Netzen*, Diplomarbeit, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1998.

[31] Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules *Tetrahedron Comput. Method.* **1992**, *3*, 537-547.

[32] Gasteiger, J.; Sadowski, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders *Chem. Rev.* **1993**, *93*, 2576-2581.

[33] Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model-Builders Using 639 X-Ray Structures *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.

[34] Gasteiger, J.; Marsili, M. Iterative Partial Equalisation of Orbital Electronegativity - A rapid Access to Atomic Charges *J. Chem. Soc. Perkin* **1984**, *2*, 559-564.

[35] Guillen, M. D.; Gasteiger, J. Extension of the Method of Iterative Partial Equalization of Orbital Electronegativity to Small Ring Systems *Tetrahedron* **1983**, *39*, 1331-1335.

[36] Gasteiger, J.; Saller, H. Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomeriekonzeptes *Angew. Chem.* **1985**, *97*, 699-701.
Angew. Chem. Int. Ed. Engl. **1985**, *24*, 687-689.

[37] Selzer, P. Infrared Data Correlations with Chemical Structure, in Ecyclopedia of Computational Chemistry, P. v. R. Schleyer Ed.; Wiley, Chichester, 1997, in Druck.

[38] Ricard, D.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. Neural Network Approach to Structural Feature Recognition from Infrared Spectra *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202-210.

[39] Zell, A.: *Simulation Neuronaler Netze*; Addison-Wesley Publishing Company, Bonn 1994.

[40] Fausett, L.: *Fundamentals of Neural Networks-Architectures, Algorithms, and Applications*; Prentice Hall International Editions, London 1994.

[41] Zupan; J. Gasteiger, J.: *Neural Networks for Chemists - An Introduction*; VCH, Weinheim 1993.

[42] Gasteiger, J.; Zupan, J. Neuronale Netze in der Chemie *Angew. Chem.* **1993**, *105*, 510-536.
Angew. Chem. Int. Ed. Engl. **1993**, *32*, 503-527.

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

[44] Hopfield, J. J. Neural Networks and Physical Systems With Emergent Collective Computational Abilities *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554-2558.

[45] Kohonen, T.: Self-organized Formation of Topologically Correct Feature Maps, *Biol. Cybern.*, **1982**, *43*, 59-69

[46] Kohonen, T. *Self-Organisation and Associative Memory*; third Edition, Springer, Berlin 1989

[47] Kohonen, T. *Self-Organizing Maps*, Huang, T. S.; Kohonen, T.; Schröder, M. R. Eds.; Springer, Berlin 1995.

[48] 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.

[49] Schuur, J. H. *Die Codierung der 3D-Struktur von Molekülen und ihre Anwendung zur Simulation von IR-Spektren und für QSAR-Untersuchungen*, Dissertation, Computer-Chemie-Centrum, Institut für Organische Chemie der Universität Erlangen-Nürnberg, 1998.

[50] Hecht-Nielsen, R. Counterpropagation Networks, *Applied Optics* **1987**, *26*, 4979-4984.

[51] Gottwald, S.; Küstner, H.; Hellwich, M.; Kästner, H. *Handbuch der Mathematik*, Buch und Zeit Verlagsgesellschaft mbH, Köln 1986, S. 624.

[52] Box, G. E. P.; Hunter, W. G.; Hunter, J. S.: *Statistics for Experimenters*, J. Wiley & Sons, New York 1978.

[53] J. Hartung; B. Elpelt: *Multivariate Statistik, Lehr- und Handbuch der angewandten Statistik*; R. Oldenbourg Verlag, München 1984, S. 512.

[54] J. Bortz: *Statistik für Sozialwissenschaftler*; 3. Auflage. Springer Verlag, Berlin 1989, S. 213.

[55] Steinhauer, V. Interner Bericht, Friedrich-Alexander Universität Erlangen-Nürnberg

[56] Merck-Index (10.) Nr. 3714

[57] Birgerson, B.; Sterner, O.; Zimerson, E.: *Chemie und Gesundheit - Eine verständliche Einführung in die Toxikologie*, Verlag Chemie, Weinheim 1988, S. 111.

[58] Eisenbrand, G.; Metzler, M.: *Toxikologie für Chemiker - Stoffe, Mechanismen, Prüfverfahren*, Georg Thieme Verlag, Stuttgart 1994, S. 39.

[59] Lang, D.; Criegee, D.; Grothusen, A.; Saalfrank, R. W.; Böcker, R. H. In Vitro Metabolism of Atrazine, Terbutylazine, Ametryne, and Terbutryne in Rats, Pigs, and Humans. *Drug Metab. Dispos.* **1996**, *24*, 859-865.

[60] Lang, D. H.; Rettie, A. E.; Böcker, R. H. Identification of Enzymes Involved in the Metabolism of Atrazine, Terbutylazine, Ametryne, and Terbutryne in Human Liver Microsomes. *Chem. Res. Toxicol.* **1997**, *10*, 1037-1044.

[61] Grothusen, A.; Hardt, J.; Bräutigam, L.; Lang, D.; Böcker, R. A conveniant method to discriminate between cytochrome P450 enzymes and flavin-containing monooxygenases in human liver microsomes, *Arch Toxicol.* **1996**, *71*, 64-71.

[62] Hardt, J. *Untersuchungen in vitro zum oxidativen Metabolismus schwefelhaltiger Pestizide durch Enzymsysteme des Menschen*, Diplomarbeit, Institut für Toxikologie und Arbeitsmedizin der Universität Erlangen-Nürnberg, 1994.

[63] Pandey, R. N.; Armstrong, A. P.; Hollenberg, P. F. Oxidative N-demethylation of N,N-Dimethylaniline by purified isozymes of cytochrome P-450, *Biochem. Pharmacol.* **1989**, *38*, 2181-2185.

[64] Sadeque, A. J. M.; Thummel, K. E.; Rettie, A. E. Purification of macaque liver flavin-containing monooxygenase: A form of the enzyme related immunochemically to an isozyme expressed selectively in adult human liver, *Biochim. Biophys. Acta* **1993**, *1162*, 127-134.

[65] Cymerman Craig, J.; Purushothaman, K. K. An Improved Preparation of Tertiary Amine N-Oxides *J. Org. Chem.* **1970**, *35*, 1721-1722.

[66] Pouchert, C. J. *The Aldrich Library of Infrared Spectra*, 3.rd Edition, Aldrich Chemical Company, Milwaukee 1981.

[67] Schrader, B. *Raman/Infrared Atlas of Organic Compounds*, Verlag Chemie, 2.nd Edition, Weinheim 1989.

[68] Un-Scan-It, Automated Digitizing System, Version 4.0 for Windows, 1996, Silk Scientific Corporation.

[69] Hölldobler, B.; Wilson, E. O.: *The Ants* Springer Verlag, Berlin 1990, S. 160.

[70] Haak, U. *Strukturaufklärung und Synthese von Spurpheromonen bei Ameisen der Gattung Camponotus*, Dissertation, Institut für Organische Chemie der Universität Erlangen-Nürnberg, 1995.

[71] Gößwald, K.: *Organisation und Leben der Ameisen*, Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart 1985.

[72] Bestmann, H. J.; Kern, F.; Schäfer, D.; Witschel, M. C. 3,4-Dihydroisocumarine, eine neue Klasse von Spurpheromomnen bei Ameisen, *Angew. Chem.* **1992**, *104*, 757-758.

[73] Bestmann, H. J.; Haak, U.; Kern, F. 2,4-Dimethyl-5-hexanolide, a Trail Pheromone Component of the Carpenter Ant Camponotus herculeanus, *Naturwissenschaften*, **1995**, *82*, 142-144.

[74] Kohl, E. Interne Mitteilung, Arbeitskreis Prof. Bestmann, Institut für Organische Chemie, Friedrich-Alexander Universität Erlangen-Nürnberg, 1998.

[75] Heintz, A.; Reinhardt, G.: *Chemie und Umwelt*; 2. Auflage, Vieweg Verlag, Braunschweig 1991, S. 201.

[76] Pimentel, D.; Levitan, L. Pesticides: Amounts applied and amounts reaching pests, *Biosci. Rep.* **1986**, *2*, 86-91.

[77] Bödecker, W.; Dümmler, C.: *Pestizide und Gesundheit: Vorkommen, Bedeutung und Prävention von Pestizidvergiftungen*, Verlag C. F. Müller, Karlsruhe 1990, S. 12.

[78] Esser, H. O.; Dupuis, G.; Ebert, E.; Vogel, C.; Marco, G. J. s-Triazines. In *Herbicides, chemistry, degradation, and mode of action,* Vol. 1;

[79] Kostka, T.; Selzer, P.; Gasteiger, J. Computer-Assisted Prediction of the Degradation Products and Infrared Spectra of s-Triazine-Herbizides, In *Software Development in Chemistry 11*; Fels, G.; Schubert, V. Eds.; Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany, 1997; p. 226.

[80] Höllering, R.; Schulz, K. P.; Gasteiger, J.; Steinhauer, L.; Kostka, T. The Simulation of Organic Reactions: From the Degradation of Chemicals, *J. Am. Chem. Soc.*, **1998**, submitted.

[81] Höllering, R. *Simulation von Massenspektren und Entwicklung eines Systems zur Reaktionsvorhersage*, Dissertation, Computer-Chemie-Centrum, Institut für Organische Chemie der Universität Erlangen-Nürnberg, 1998.
Zugriff über Internet: http://www2.ccc.uni-erlangen.de/dissertationen/data/dissertation/Robert_Hoellering/html/

[82] Kearney, P. C.; Kaufmann, D. D.; Sheets, T. J. *J. Agr. Food Chem.* **1965**, *13*, 369

[83] Verein zur Förderung eines Deutschen Forschungsnetzes e.V. (http://www.dfn.de)

[84] Selzer, P.; Hemmer, M.; Schuur, J.; Steinhauer, V.; Gasteiger, J. TeleSpek-Telekooperation in der Spektroskopie, *Nachr. Chem. Tech. Lab.*, **1998**, *46*, A78-A82.

[85] Ihlenfeldt, W.-D. Chemie-Multimedial *Chemie in unserer Zeit*, **1997**, *3*, 150-151.

[86] Schuur, J. H.; Selzer, P.; Steinhauer, V.; Gasteiger, J. Kooperative, rechnergestützte IR-Spektreninterpretation - neue Wege für die Infrarotspektroskopie *GIT Labor-Fachzeitschrift*, **1997**, *3*, 283-286.

[87] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.

[88] SMILES Erläuterungen im Internet:
http://www.daylight.com/dayhtml/smiles/smiles-intro.html
http://www2.ccc.uni-erlangen.de/services/smiles.html

[89] McDonald, R. S.; Wilks, P. A. *Appl. Spectrosc.* JCAMP-DX: A Standard Form for the Exchange of Infrared Spectra in Computer Readable Form, **1988**, *42*, 151-162.

[90] Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors *J. Med. Chem.* **1996**, *39*, 3049-3059.

[91] Engelke, F.: *Aufbau der Moleküle*; B. G. Teubner Verlag, Stuttgart 1992, S. 58.

[92] Böhlig, H.; Geiseler, G.: *Molekülschwingungen und Kraftkonstanten*; NOVA ACTA LEOPOLDINA, Halle (Saale) 1988.

[93] Fadini, A.; Schnepel, F.-M.; *Schwingungsspektroskopie - Methoden/Anwendungen*; Georg Thieme Verlag, Stuttgart 1985, S 44.

A. Anhang

Der Anhang enthält die folgenden Datensätze und Spektrensammlungen:

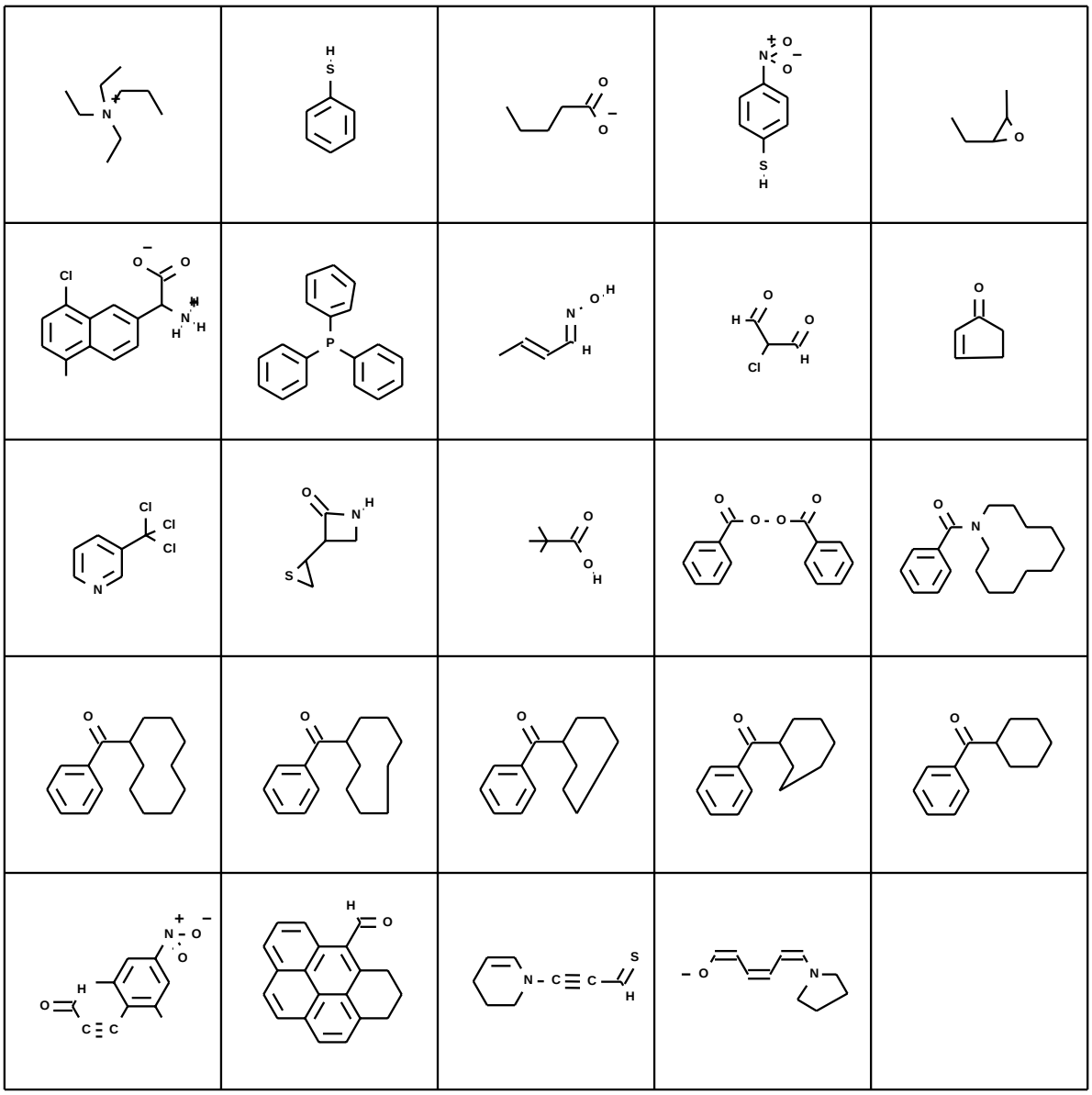
- Moleküldatensatz zur Bestimmung der Skalierungsfaktoren für den 3D-MoRSE Code
- Alle Moleküle der SpecInfo IR-Datenbank mit einem s-Triazin-Gerüst (vgl. Kap. 2.4.2.4)
- Darstellung der Simulationsergebnisse von Kapitel 2.6
- Datensatz mit 81 Molekülen mit nicht-datenreduzierten Spektren (vgl. Kap. 2.7)
- Darstellung der Simulationsergebnisse für 81 Moleküle mit nicht-datenreduzierten Spektren (vgl. Kap. 2.7)
- Datensatz der generierten Trietazin-Abbauprodukte sowie der entsprechenden Tautomere (vgl. Kap. 3.3.3)

Weiterhin befindet sich im Anhang:

- Publikationsliste
- Lebenslauf

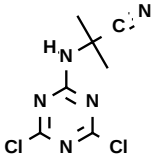
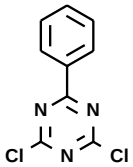
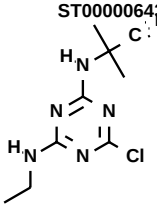
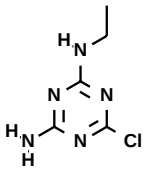
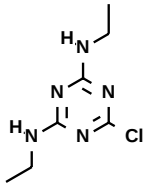
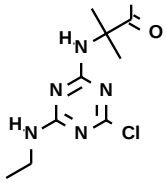
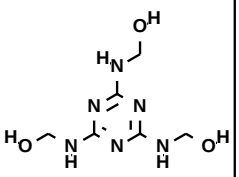
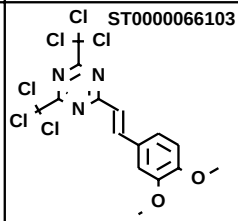
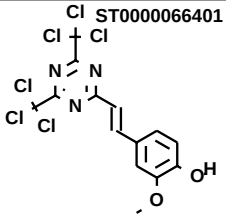
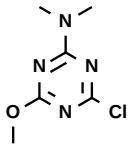
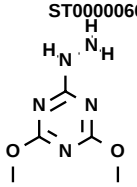
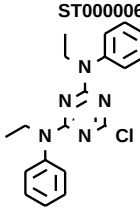
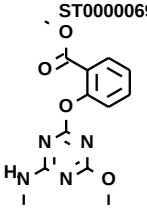
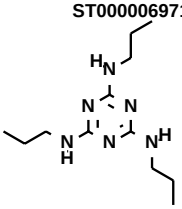
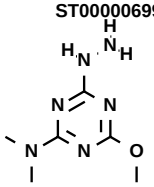
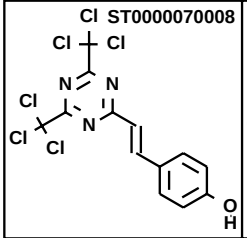
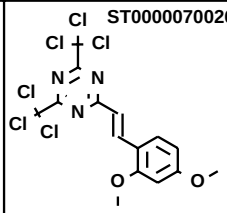
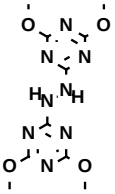
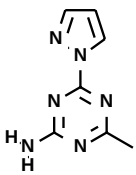
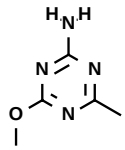
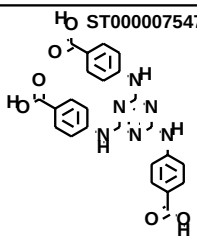
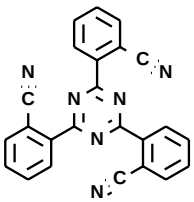
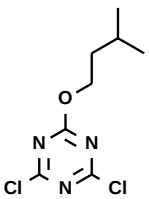
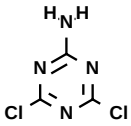
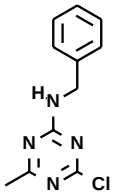
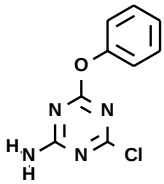
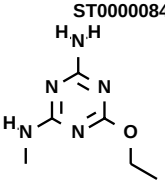
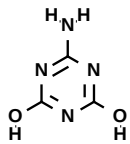
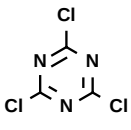
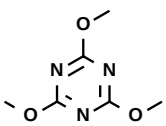
A.1 Skalierungsdatensatz für den 3D-MoRSE Code

#1 /home/jan/prog/skale.ctx [1 – 24]

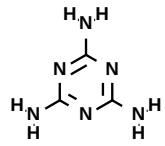
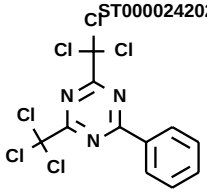
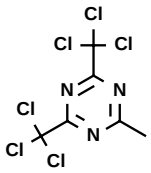
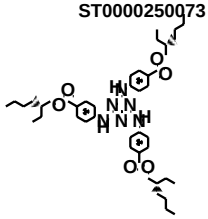
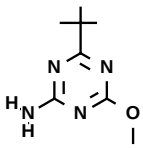
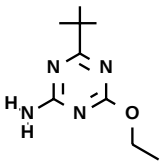
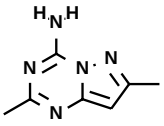
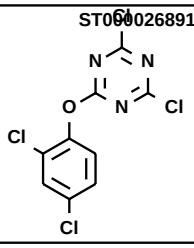
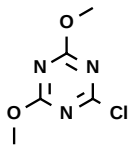
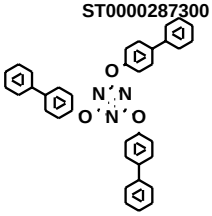
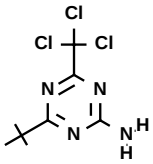
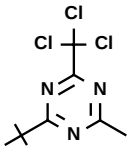
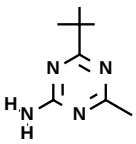
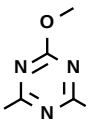
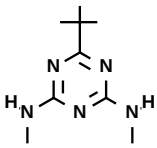
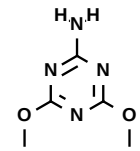


A.2 Triazin-Datensatz

#1 /home/pauls/molgraph/triazine/triazin.ctx [1 – 30]

ST0000056520 	ST0000063772 	ST0000064327 	ST0000064355 	ST0000064380 
ST0000064515 	ST0000064910 	ST0000066103 	ST0000066401 	ST0000066428 
ST0000066457 	ST0000066800 	ST0000069525 	ST0000069712 	ST0000069995 
ST0000070008 	ST0000070026 	ST0000070185 	ST0000071293 	ST0000072279 
ST0000075473 	ST0000075902 	ST0000077509 	ST0000079861 	ST0000081574 
ST0000084141 	ST0000084646 	ST0000168701 	ST0000222903 	ST0000225837 

#2 /home/pauls/molgraph/triazine/triazin.ctx [31 – 46]

ST0000227889 	ST0000242025 	ST0000246533 	ST0000250073 	ST0000259963 
ST0000262969 	ST0000268811 	ST0000268919 	ST0000269695 	ST0000287300 
ST0000330940 	ST0000332179 	ST0000335529 	ST0000337144 	ST0000337902 
ST0000361133 				

A.3 Darstellung der Simulationsergebnisse von Kapitel 2.6

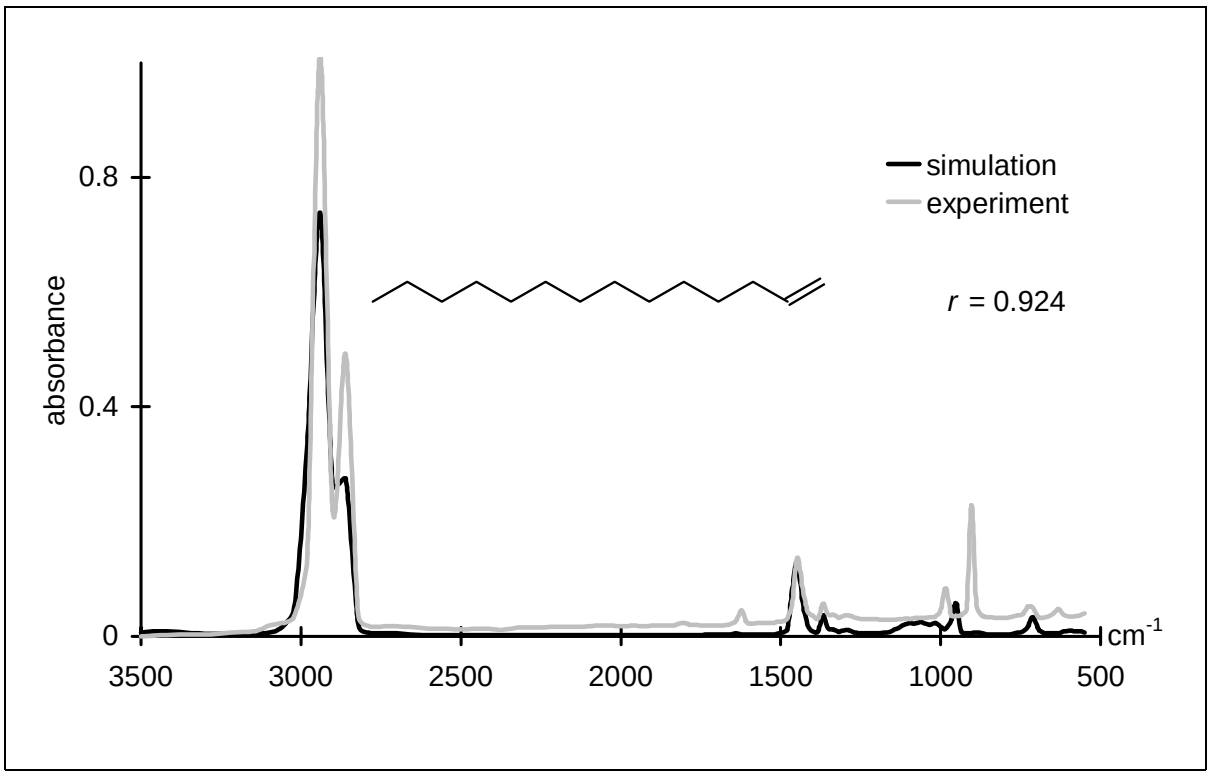


Abb. 7-1: Simulationsergebnis für Verbindung dresden_1

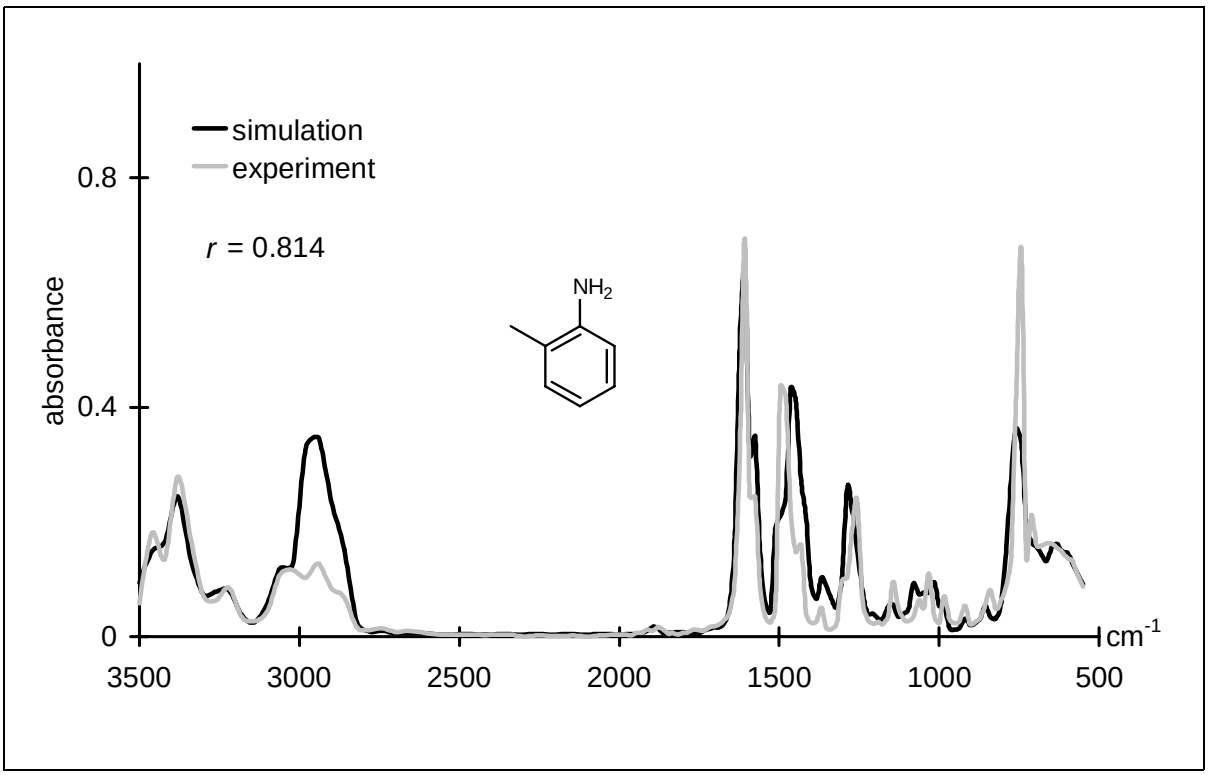


Abb. 7-2: Simulationsergebnis für Verbindung dresden_2

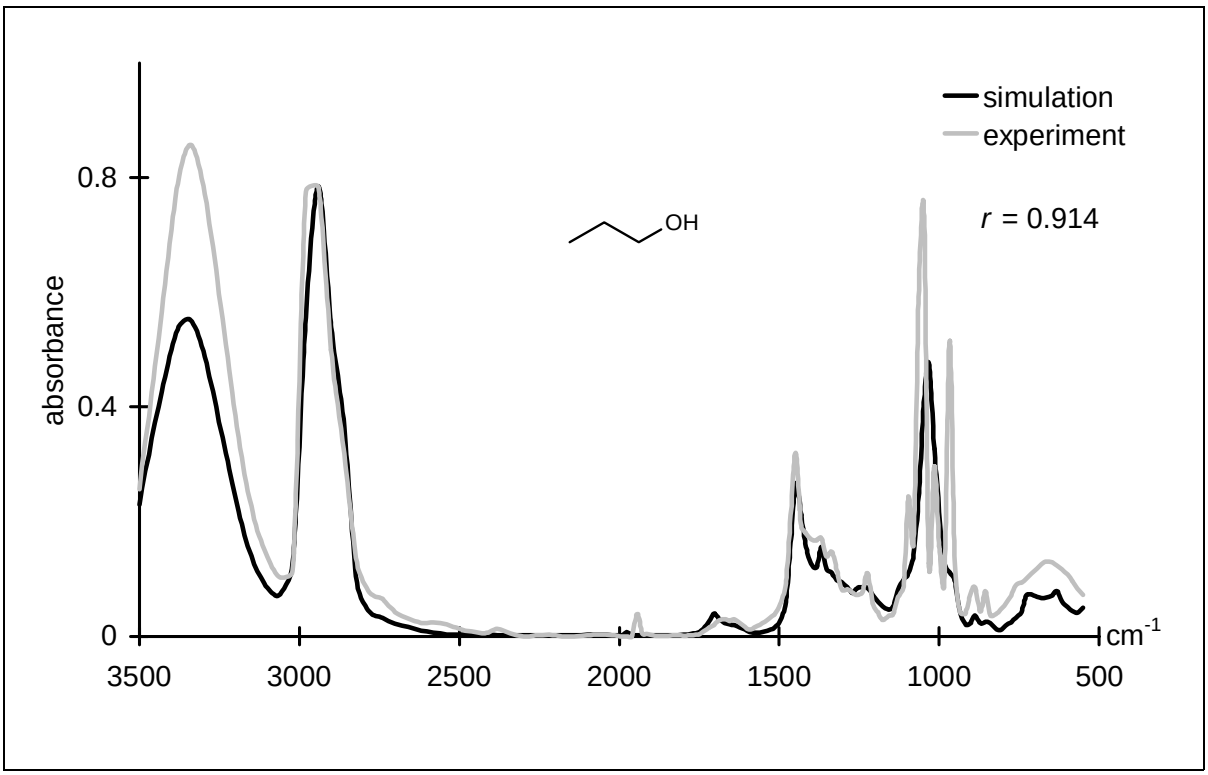


Abb. 7-3: Simulationsergebnis für Verbindung dresden_3

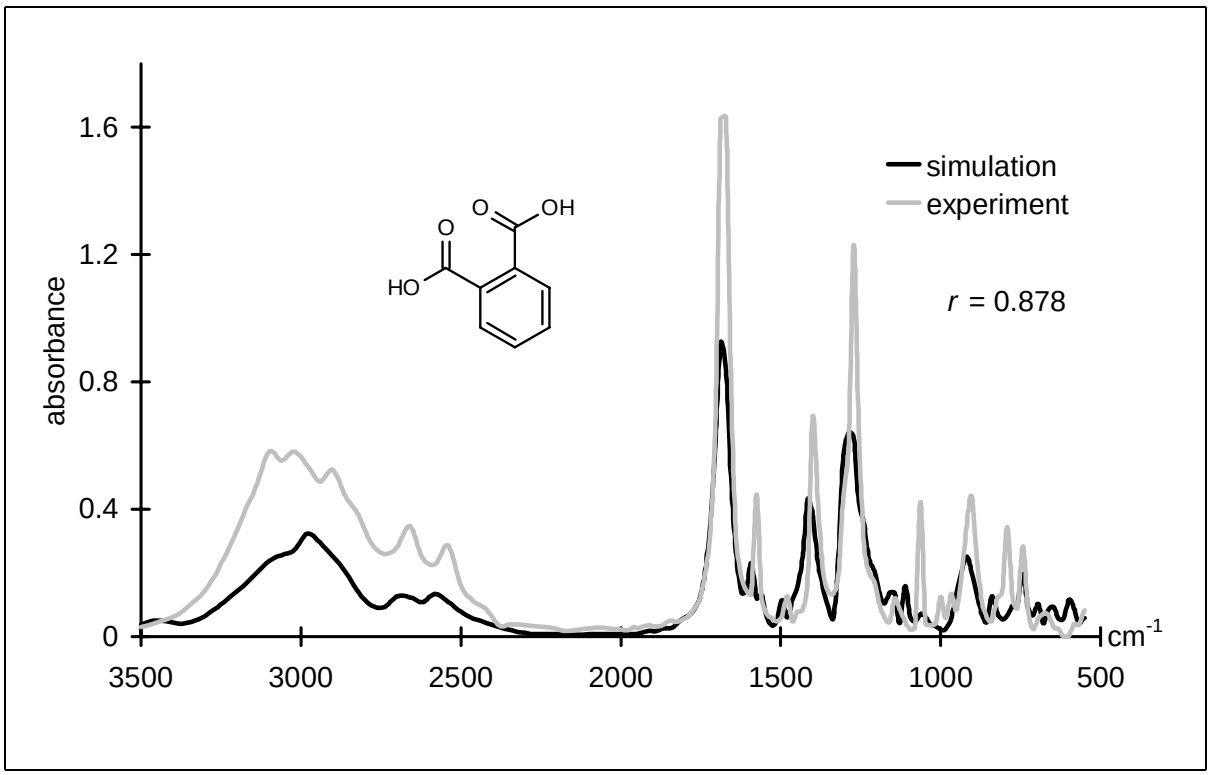


Abb. 7-4: Simulationsergebnis für Verbindung dresden_4

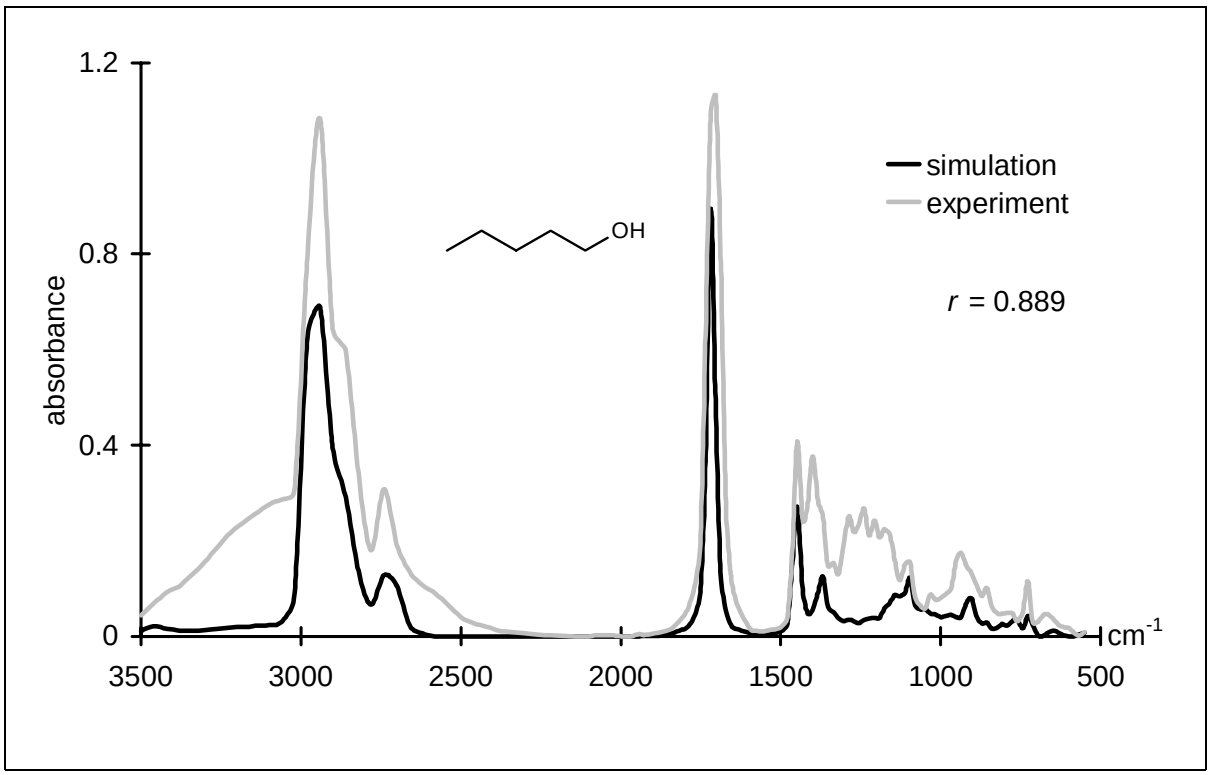


Abb. 7-5: Simulationsergebnis für Verbindung dresden_5

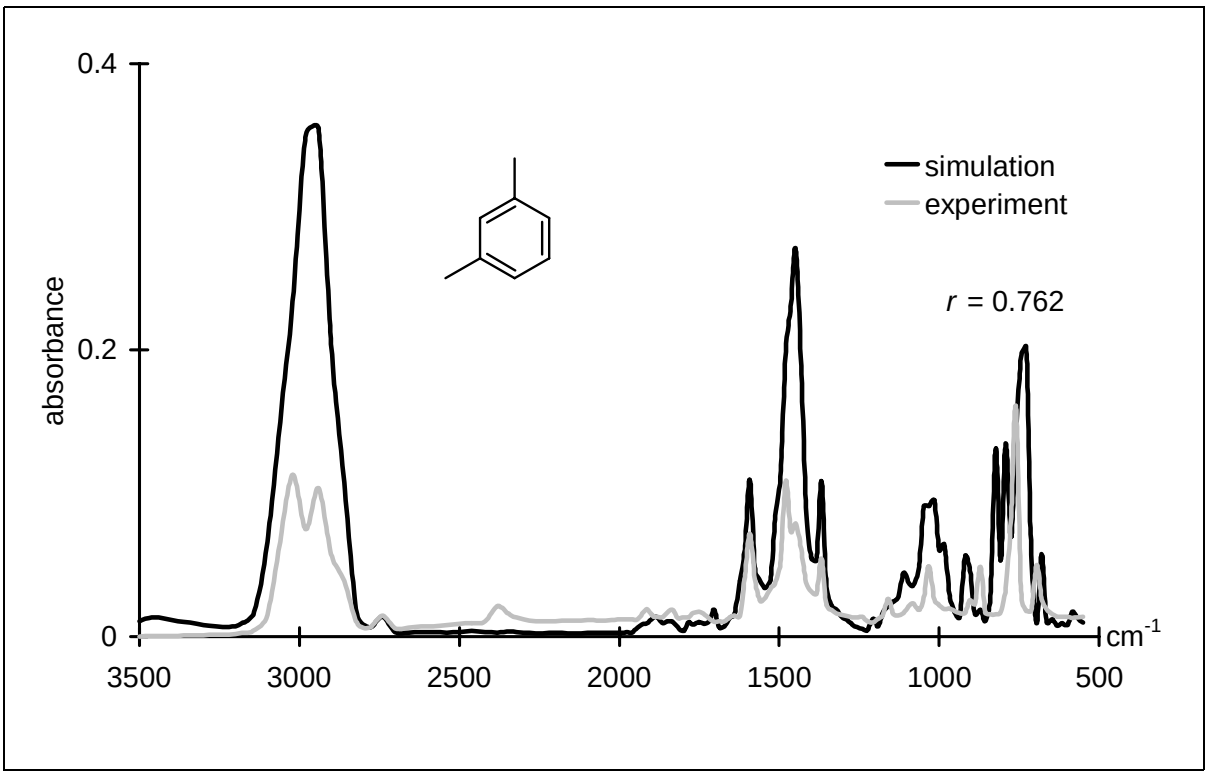


Abb. 7-6: Simulationsergebnis für Verbindung dresden_6

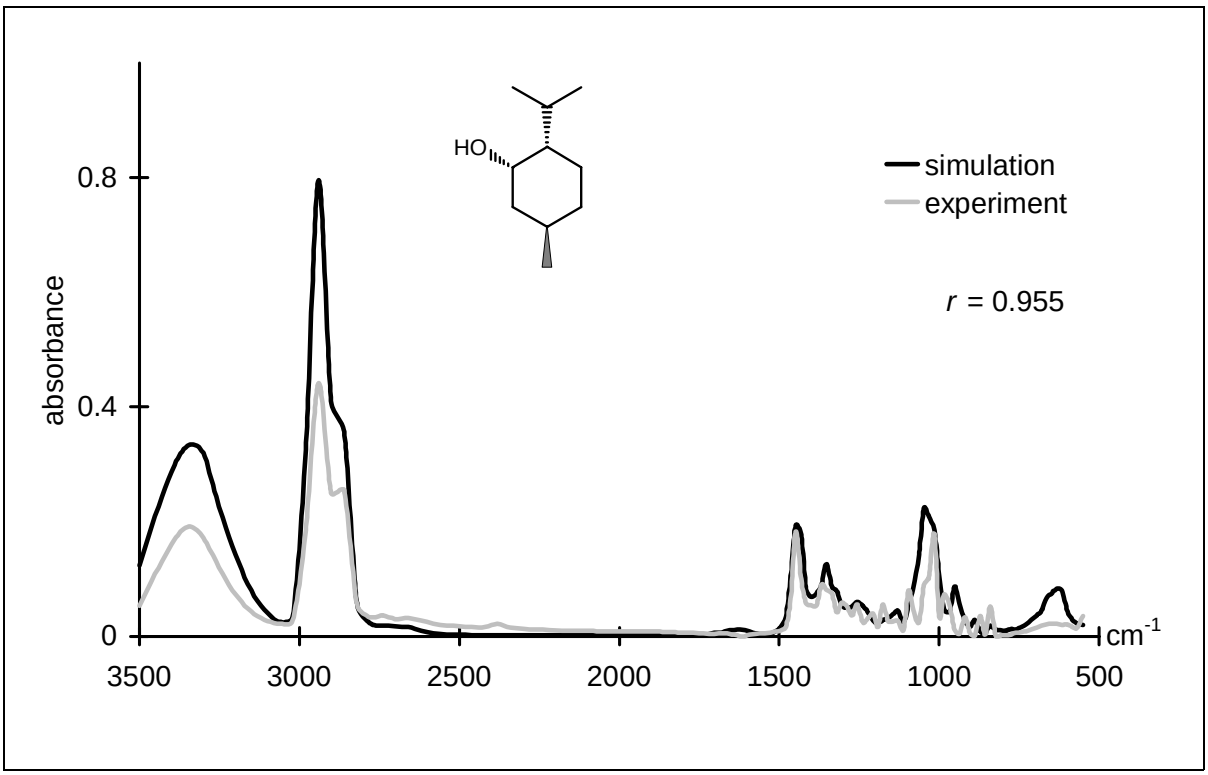


Abb. 7-7: Simulationsergebnis für Verbindung dresden_7

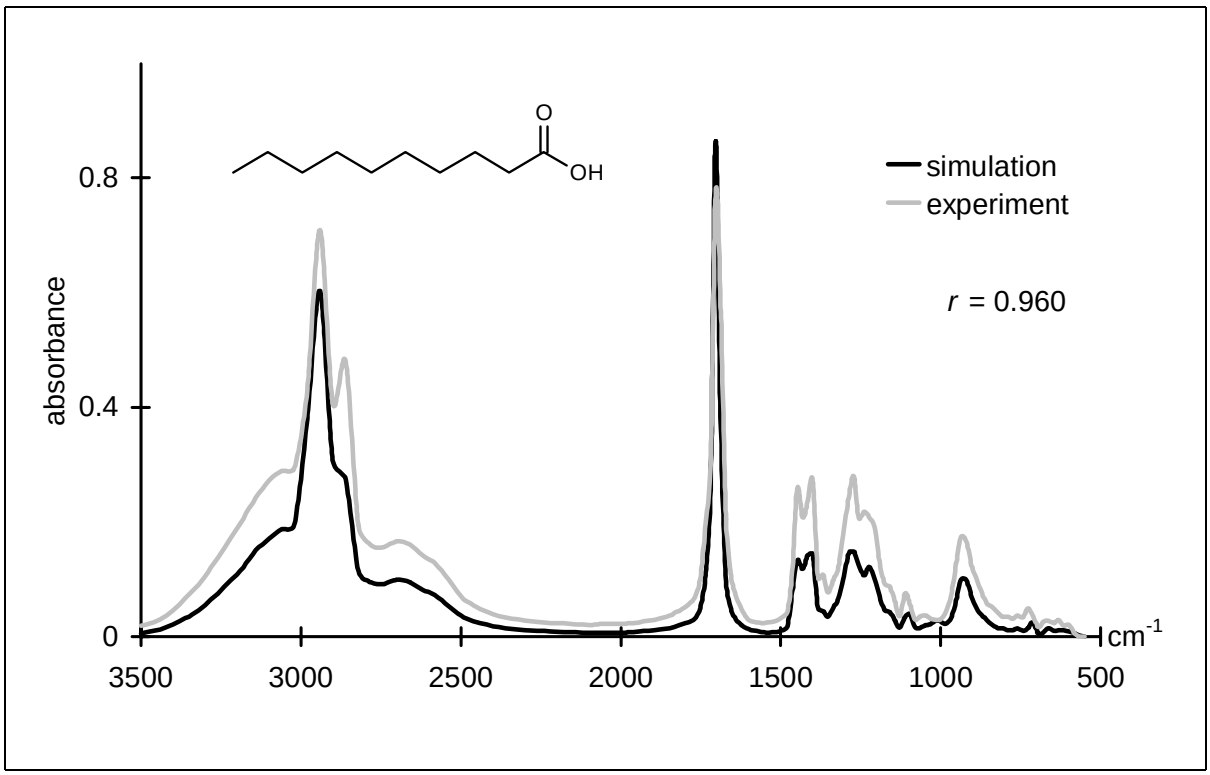


Abb. 7-8: Simulationsergebnis für Verbindung dresden_8

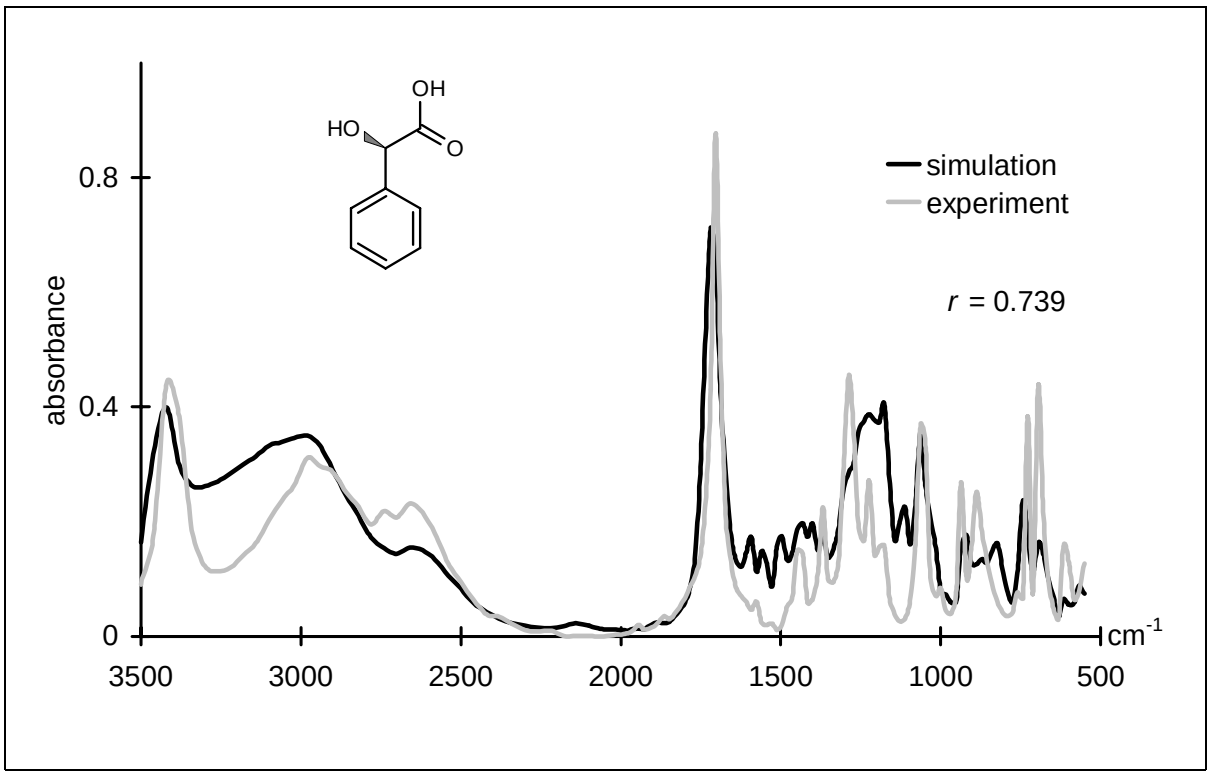


Abb. 7-9: Simulationsergebnis für Verbindung dresden_9

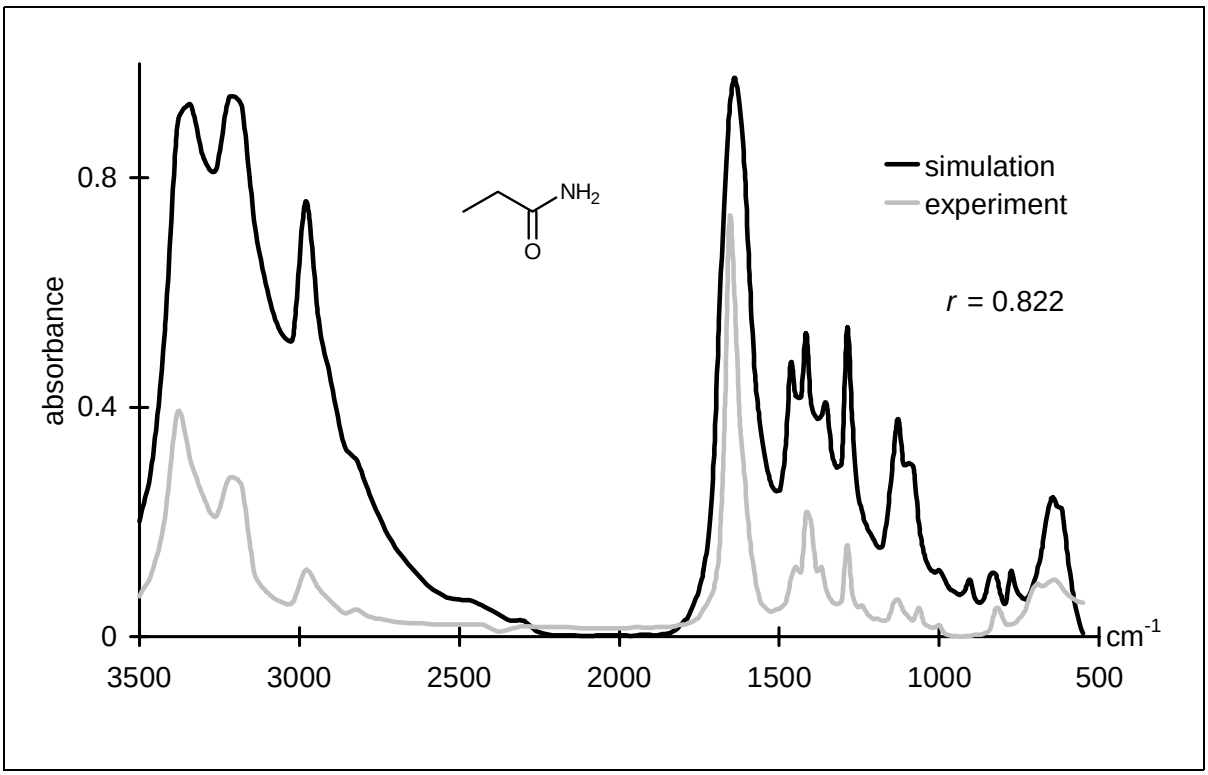


Abb. 7-10: Simulationsergebnis für Verbindung dresden_10

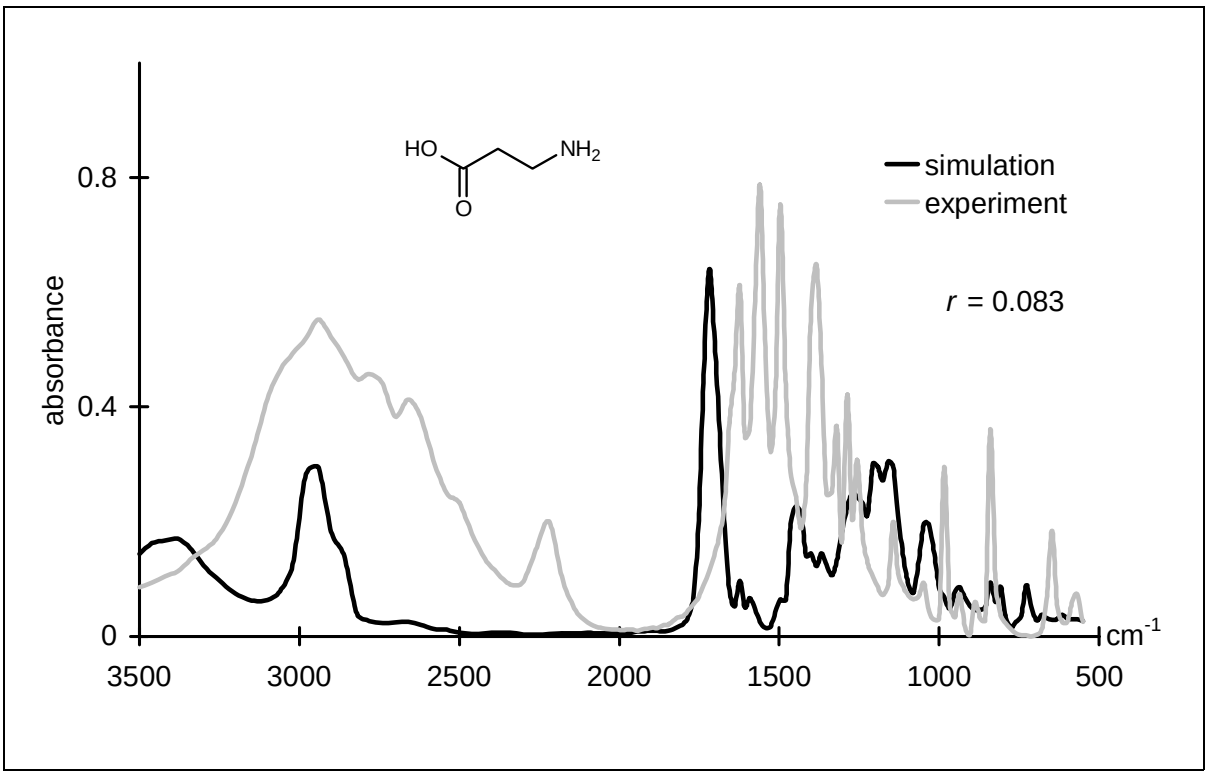


Abb. 7-11: Simulationsergebnis für Verbindung dresden_11

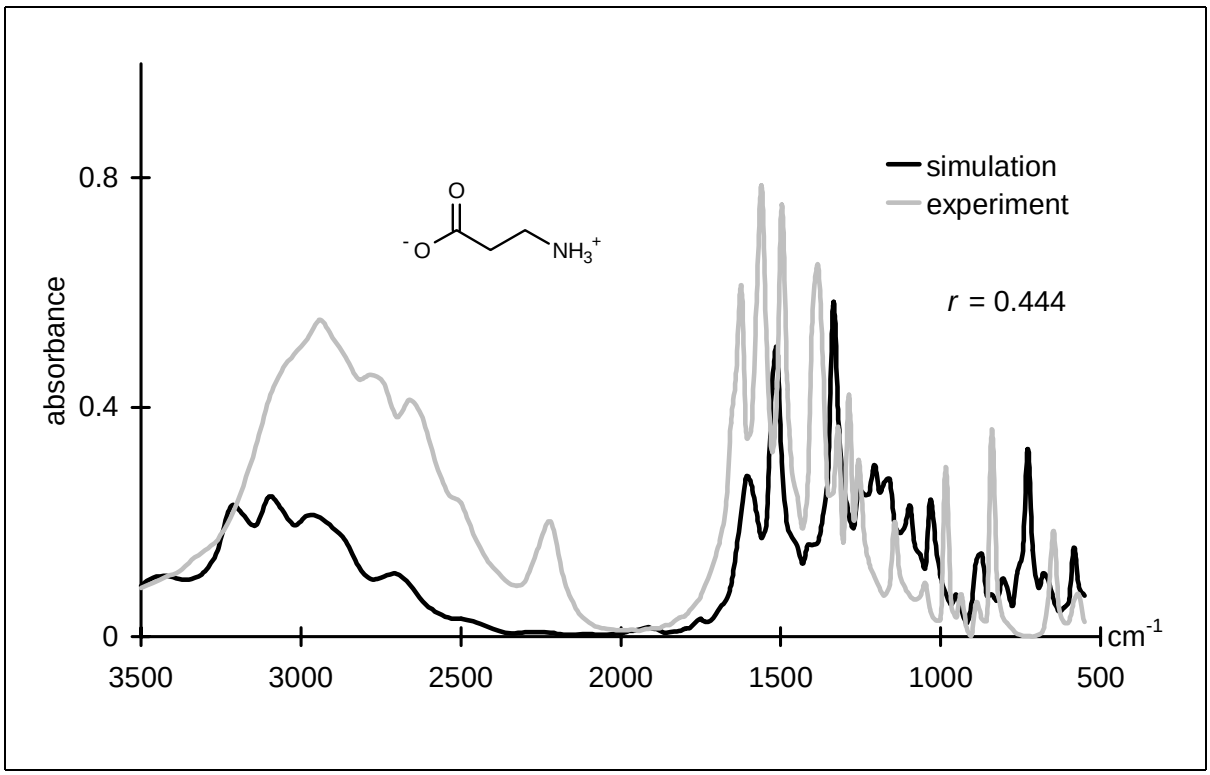


Abb. 7-12: Simulationsergebnis für Verbindung dresden_11 in der zwitterionischen Form

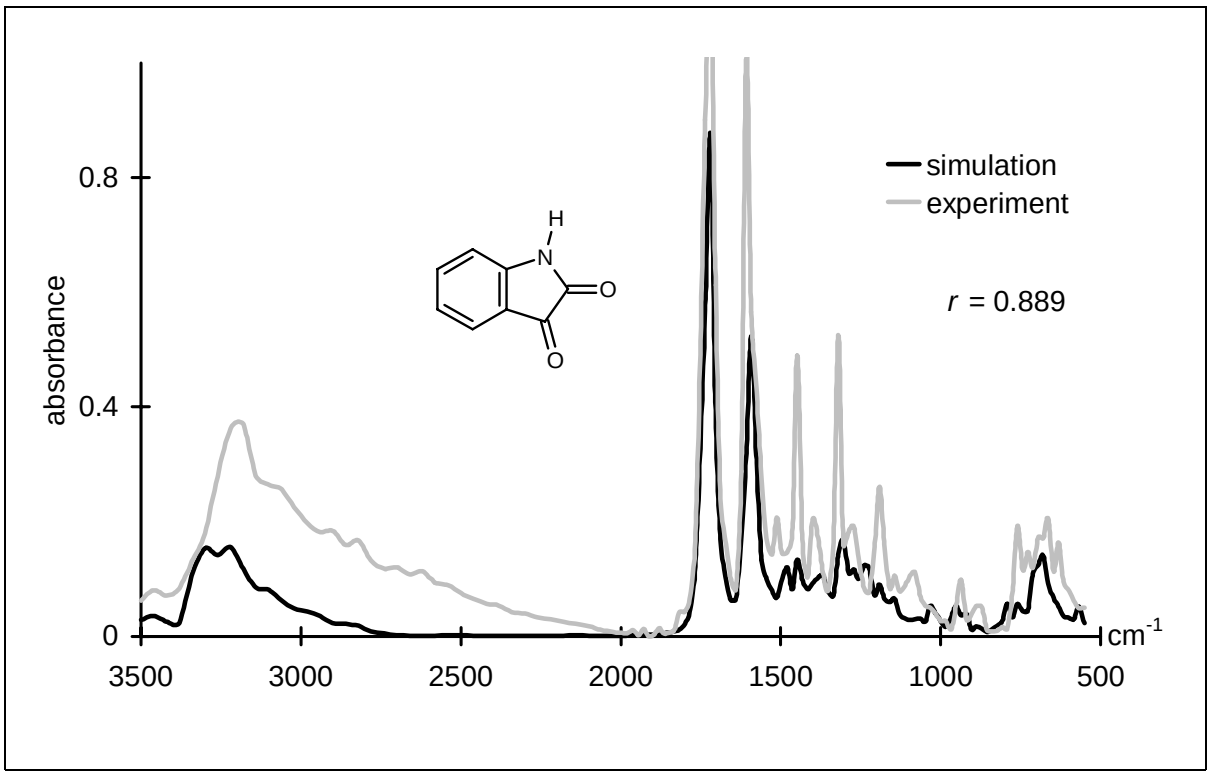


Abb. 7-13: Simulationsergebnis für Verbindung dresden_12

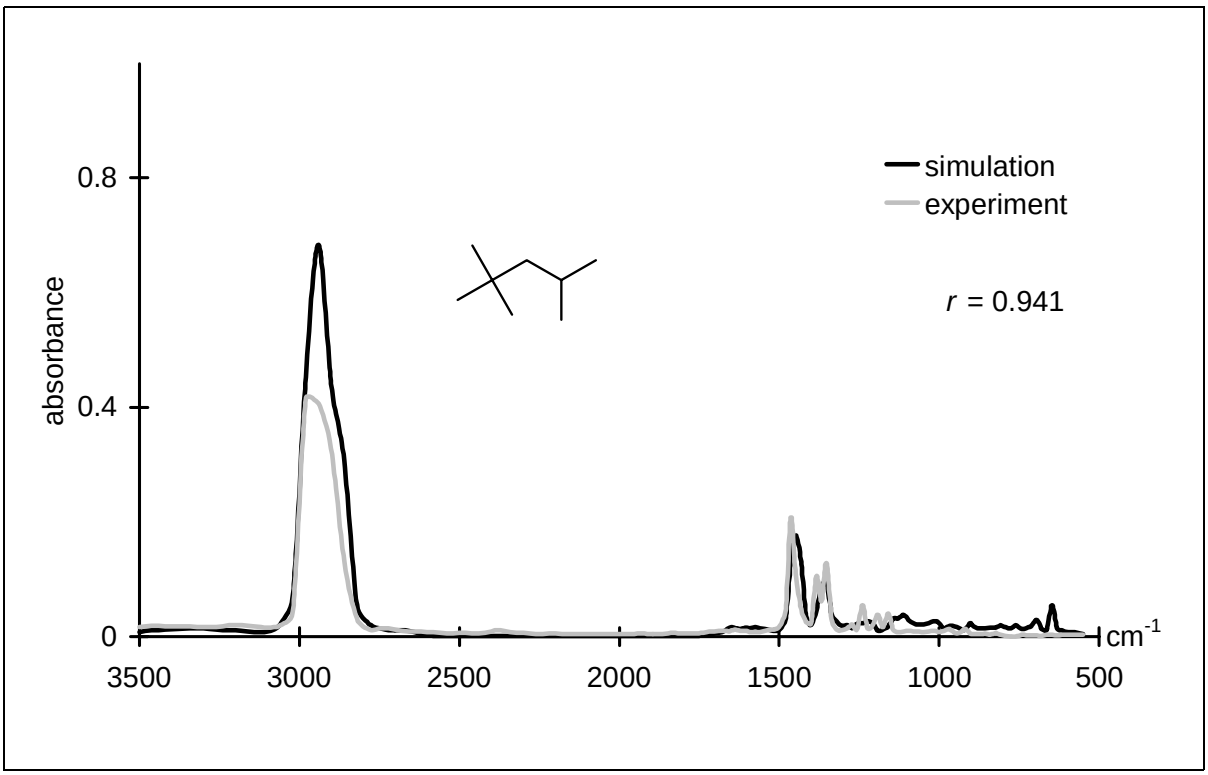


Abb. 7-14: Simulationsergebnis für Verbindung dresden_13

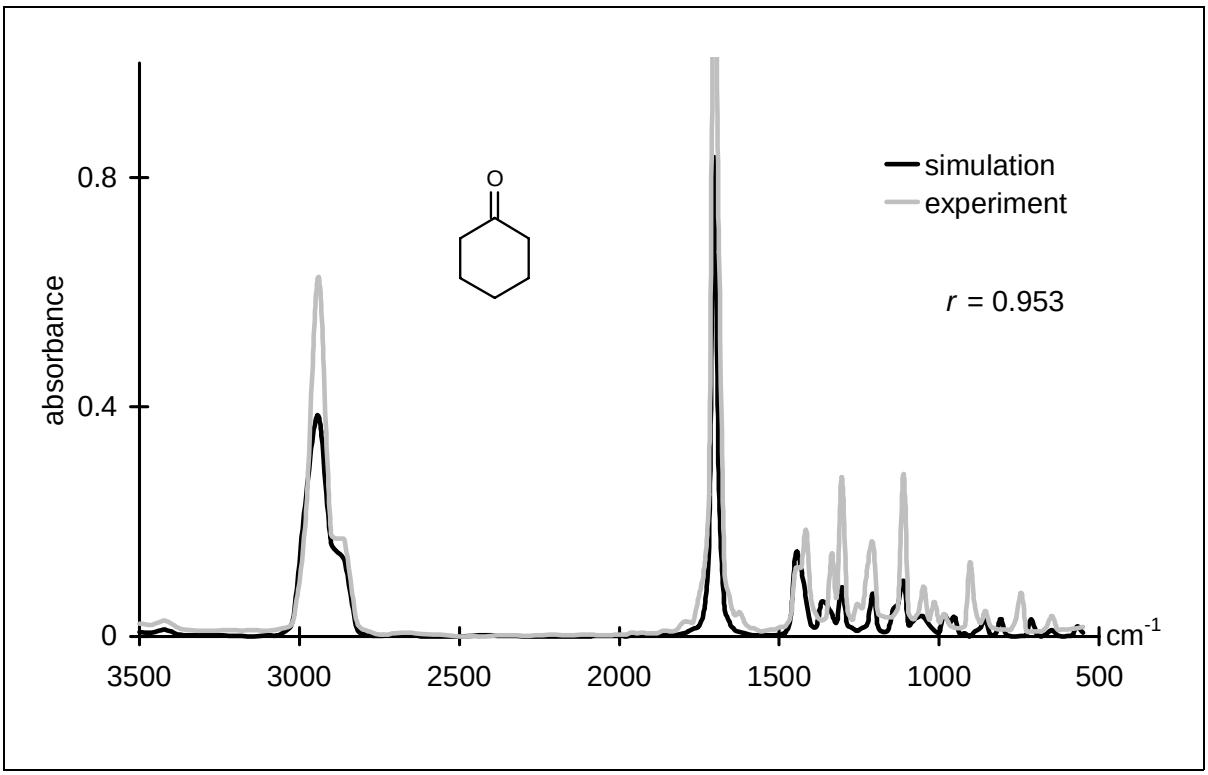


Abb. 7-15: Simulationsergebnis für Verbindung dresden_14

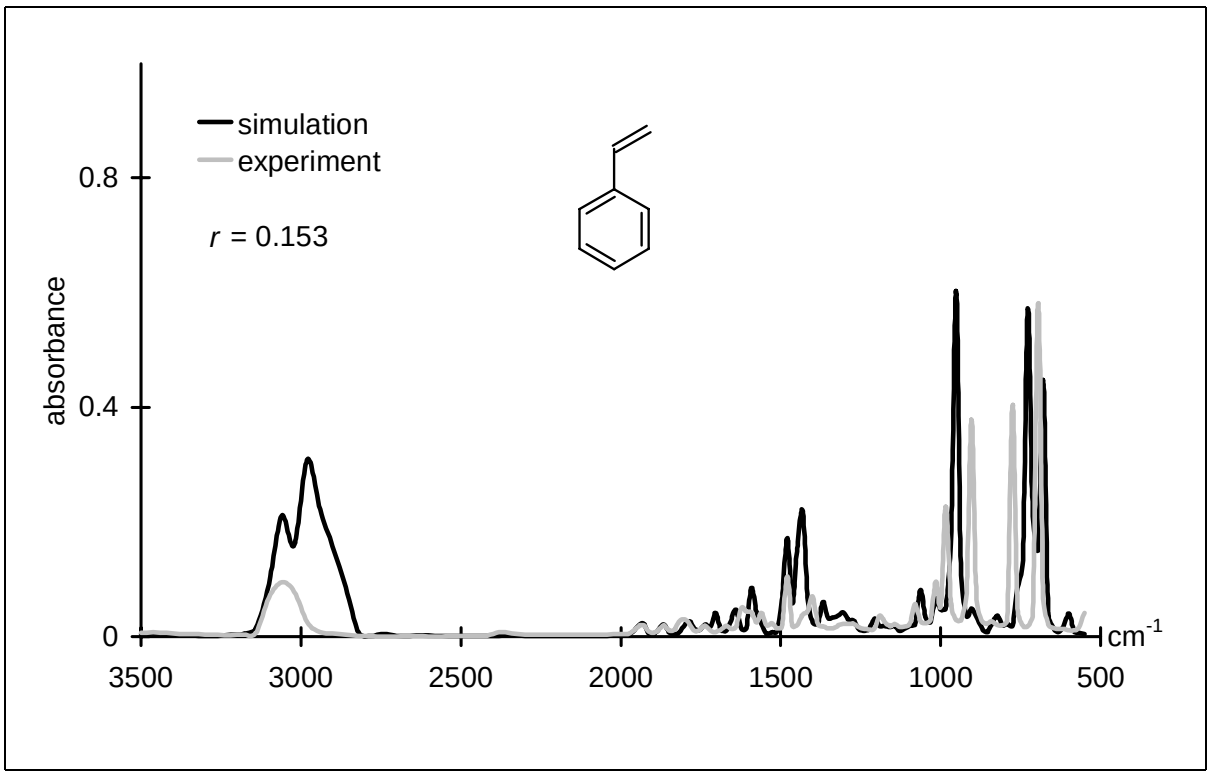


Abb. 7-16: Simulationsergebnis für Verbindung dresden_15

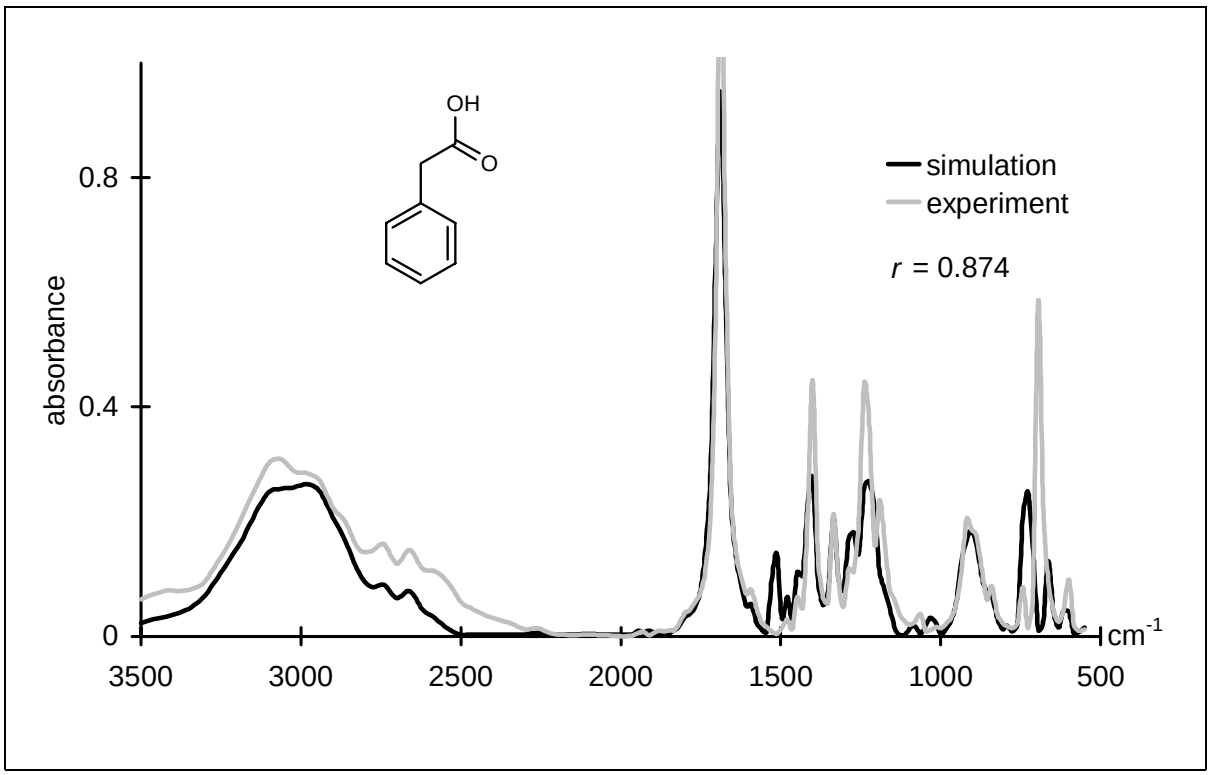
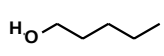
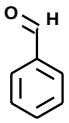
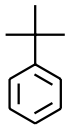
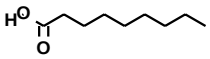
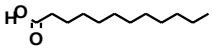
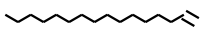
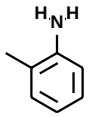
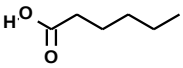
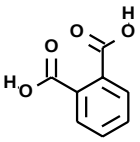
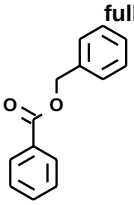
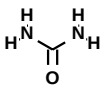
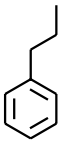
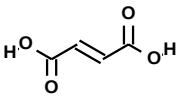
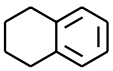
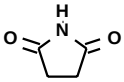
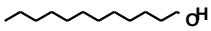
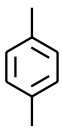
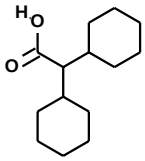
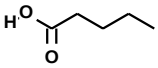
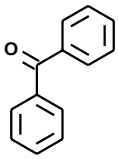
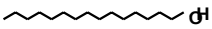
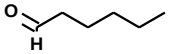
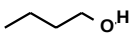
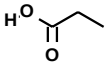
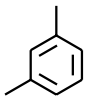
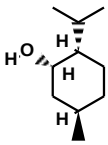
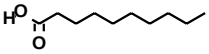
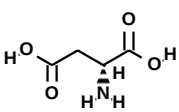
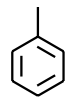
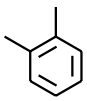
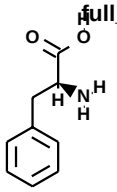
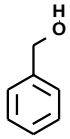
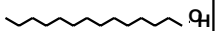
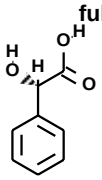
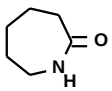
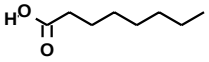
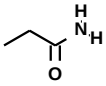
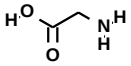
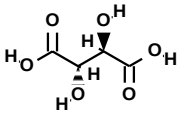
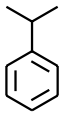
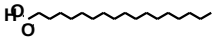
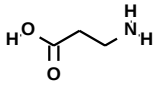
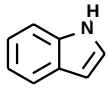
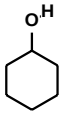
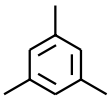
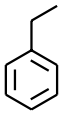
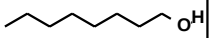
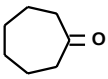
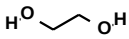
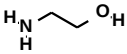
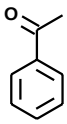
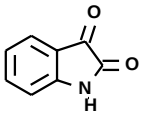
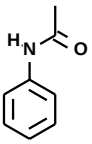
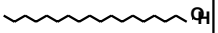
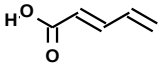
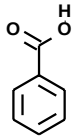
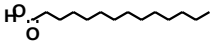
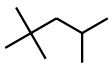
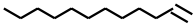
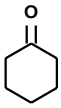
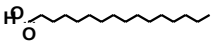
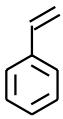
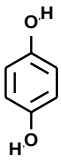
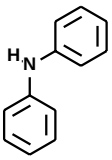
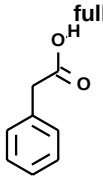


Abb. 7-17: Simulationsergebnis für Verbindung dresden_16

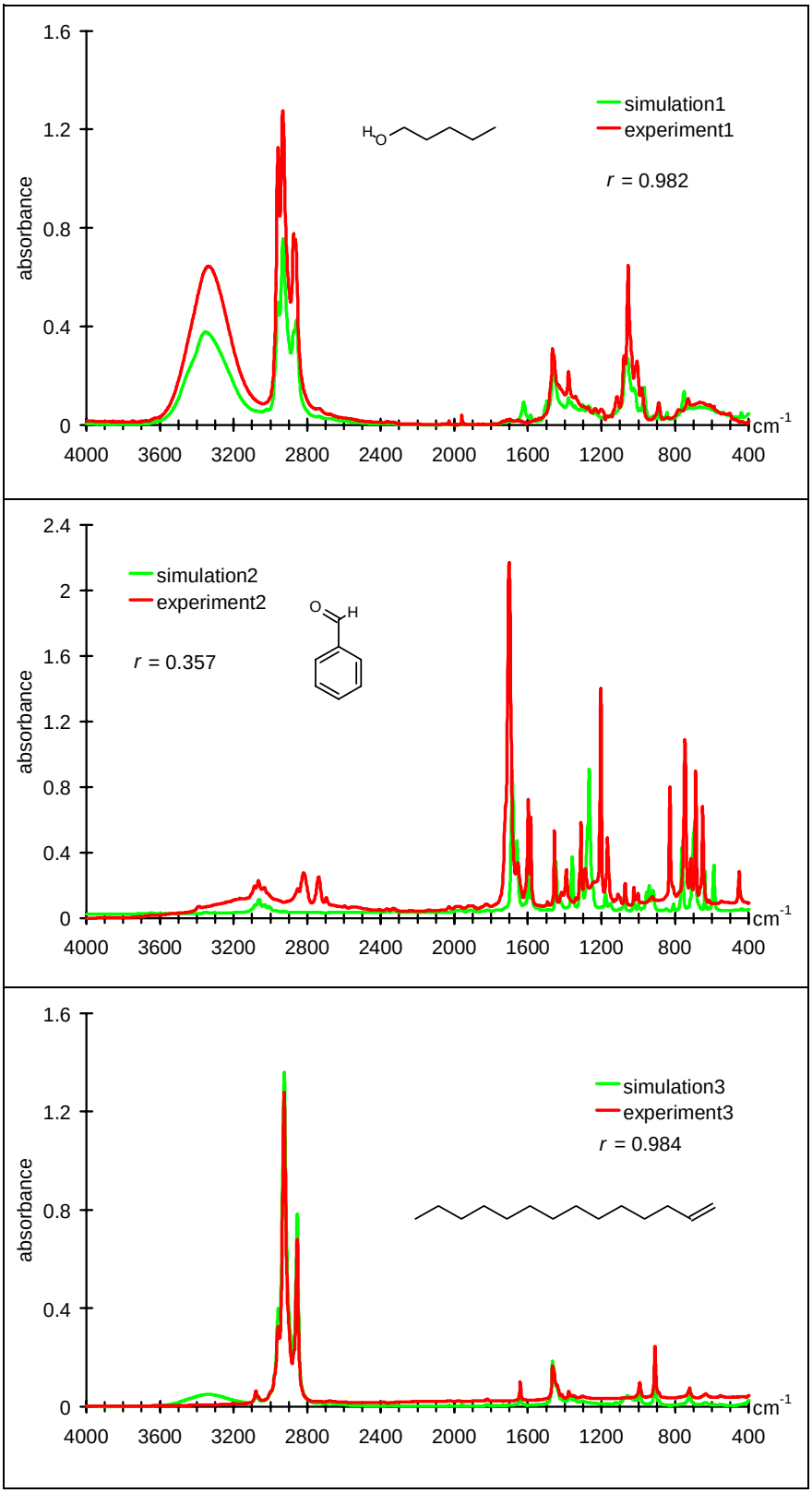
A.4 Datensatz mit nicht-datenreduzierten Spektren

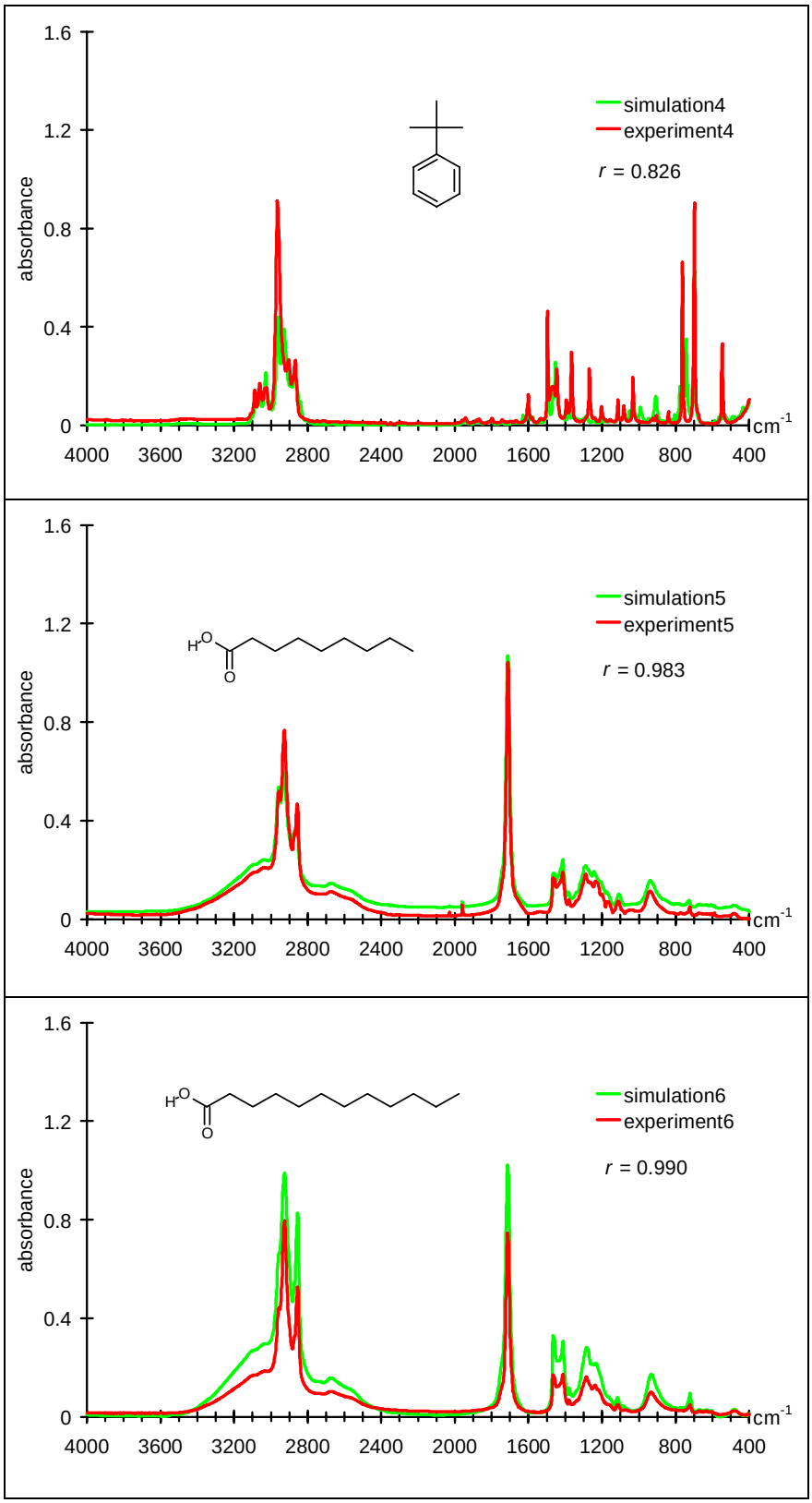
full_1 	full_2 	full_3 	full_4 	full_5 
full_6 	full_7 	full_8 	full_9 	full_10 
full_11 	full_12 	full_13 	full_14 	full_15 
full_16 	full_17 	full_18 	full_19 	full_20 
full_21 	full_22 	full_23 	full_24 	full_25 
full_26 	full_27 	full_28 	full_29 	full_30 
full_31 	full_32 	full_33	full_34	full_35 

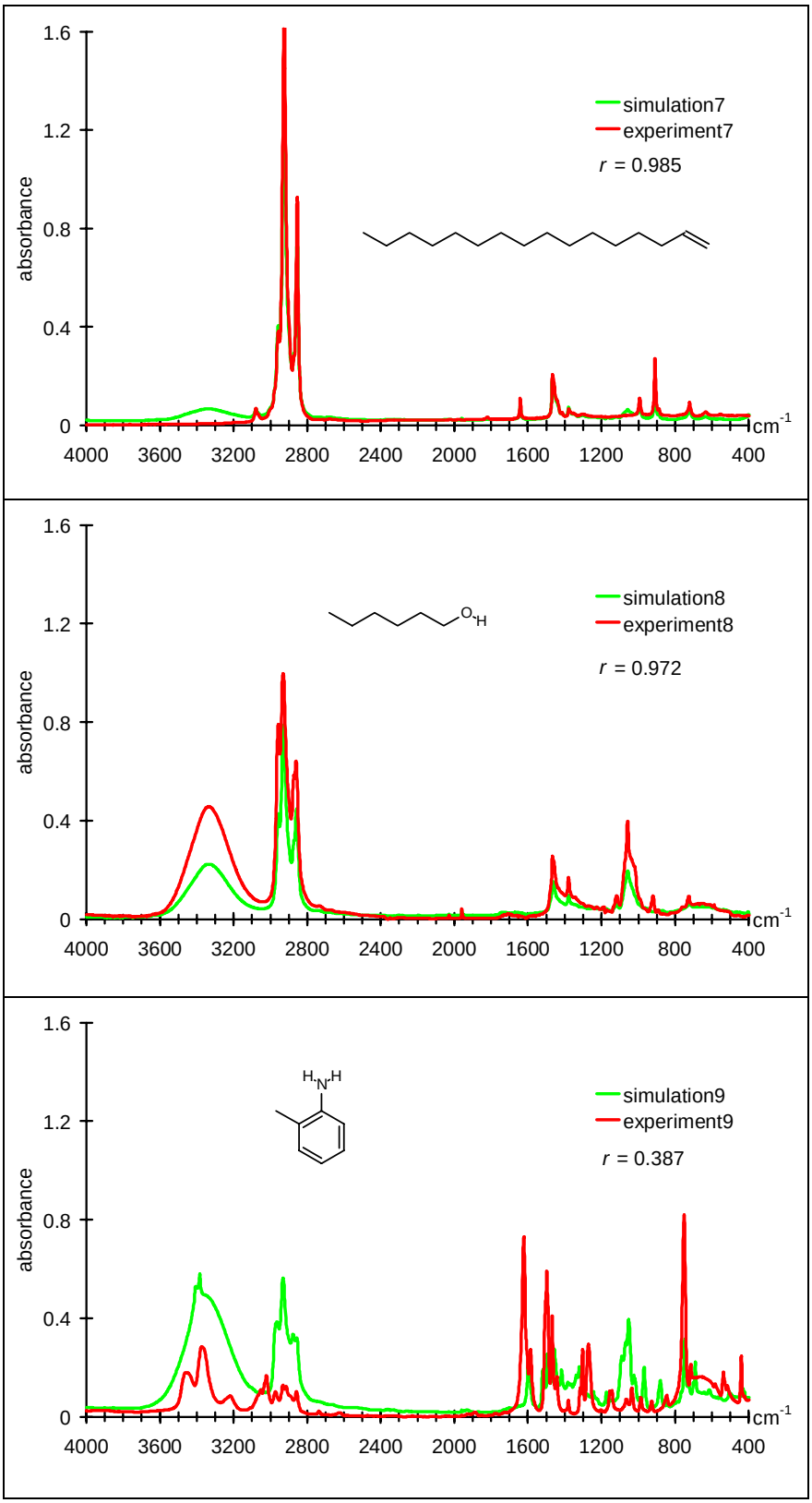
full_36 	full_37 	full_39 	full_40 	full_41 
full_42 	full_43 	full_44 	full_45 	full_46 
full_47 	full_48 	full_49 	full_50 	full_51 
full_52 	full_53 	full_54 	full_55 	full_56 
full_57 	full_58 	full_59 	full_60 	full_61 
full_62 	full_63 	full_64	full_65 	full_66 
full_67 	full_68 	full_69 	full_70 	full_71 

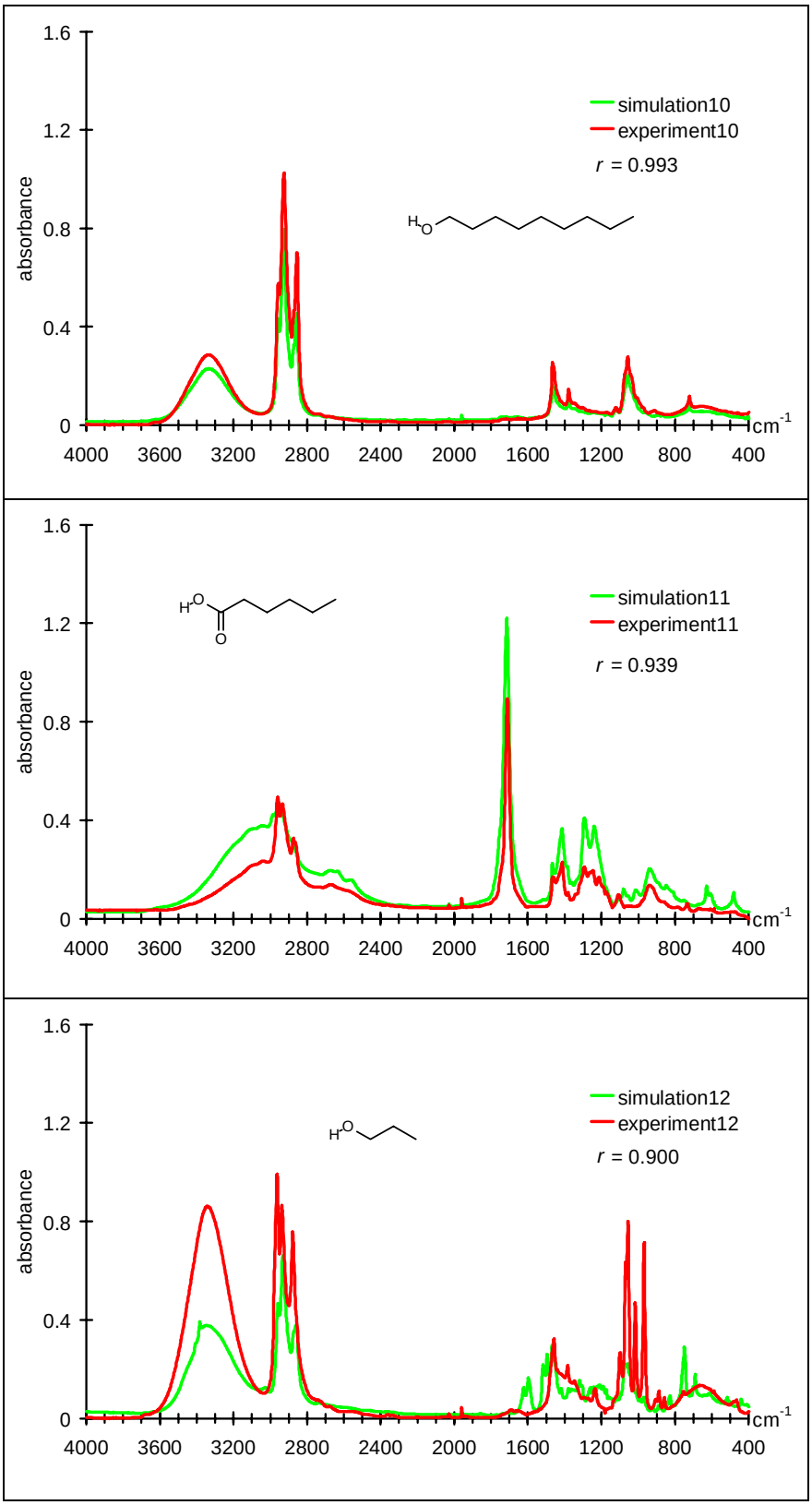
full_72 	full_73 	full_74 	full_75 	full_76 
full_77 	full_78 	full_79 	full_80 	full_82 
full_83 				

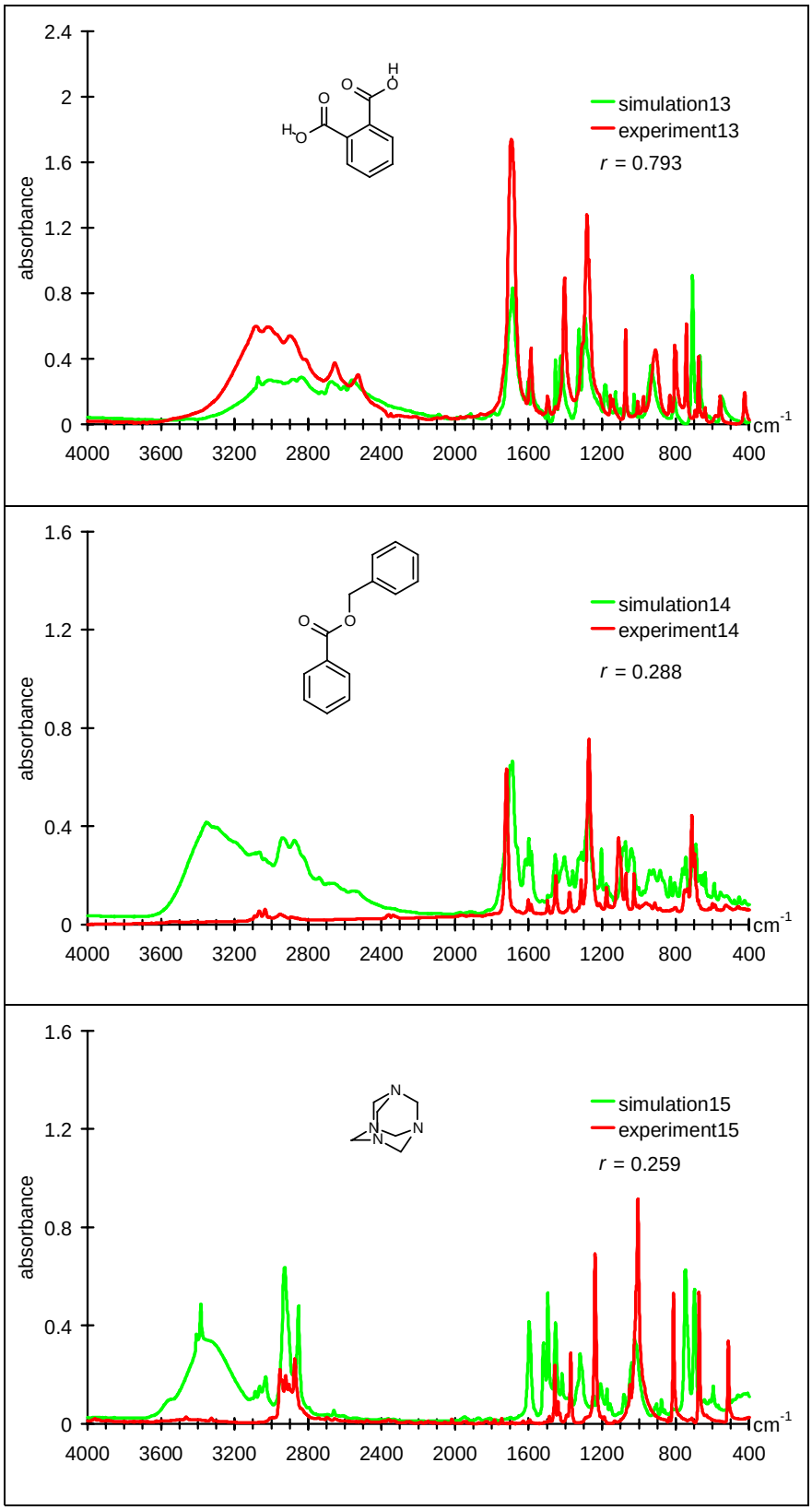
A.5 Simulationsergebnisse mit nicht-datenreduzierten Spektren

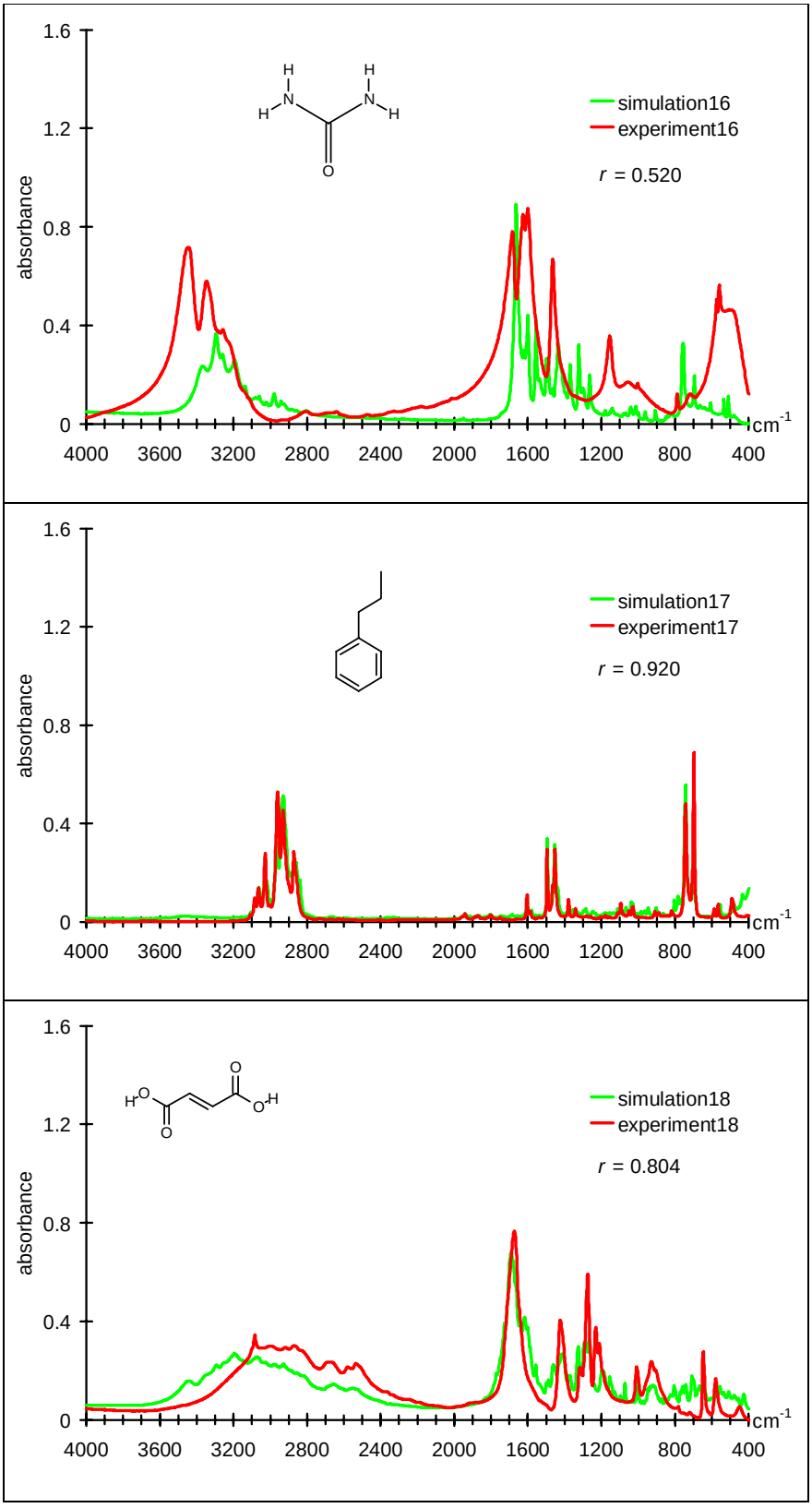


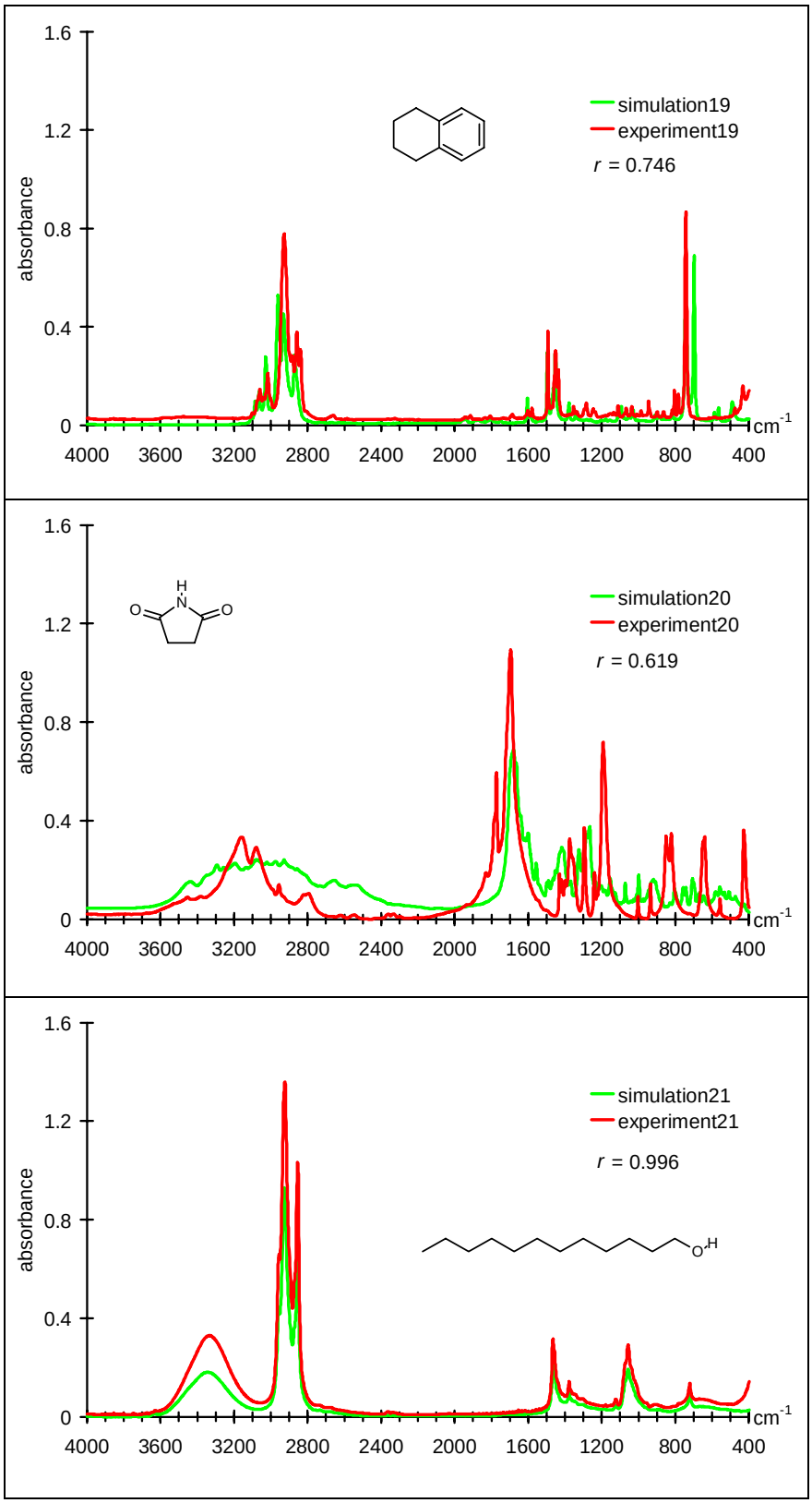

absorbance
1.6
1.2
0.8
0.4
0
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation4
experiment4
r = 0.826
simulation5
experiment5
r = 0.983
simulation6
experiment6
r = 0.990

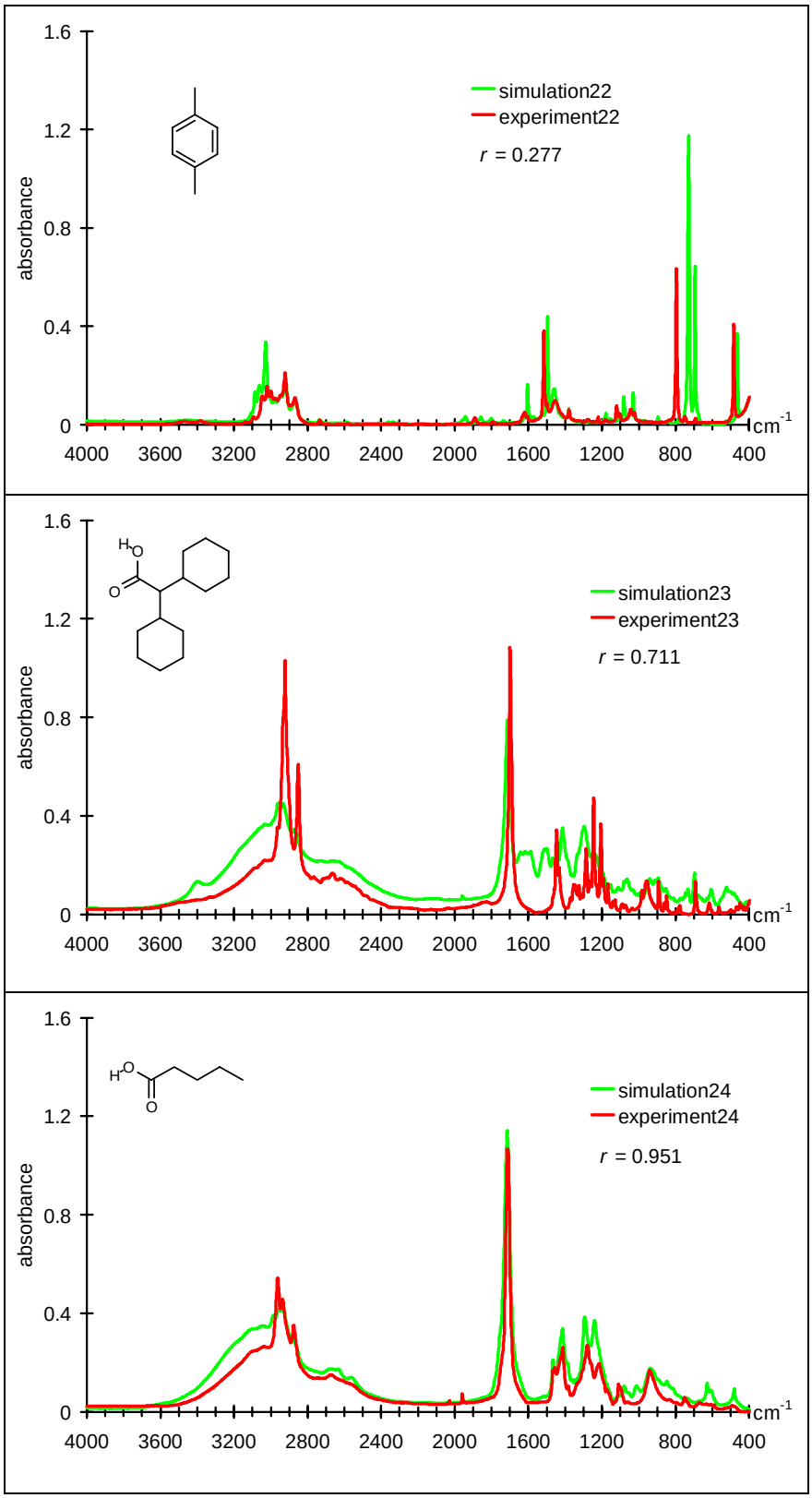

1.6
1.2
0.8
0.4
0
absorbance
simulation7
experiment7
r = 0.985
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation8
experiment8
r = 0.972
O
H
simulation9
experiment9
r = 0.387
H
N
H

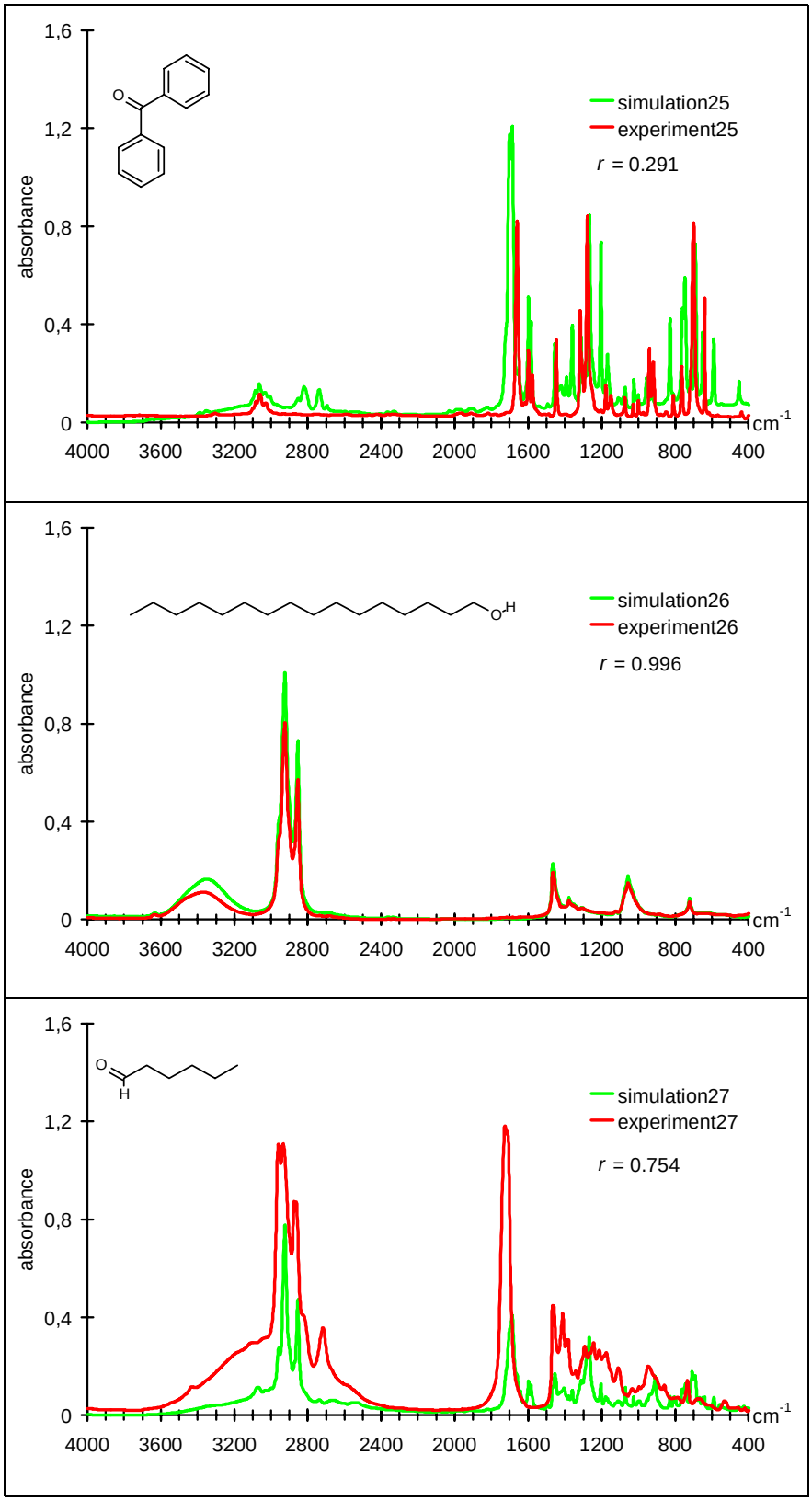

1.6
1.2
0.8
0.4
0
absorbance
simulation10
experiment10
r = 0.993
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation11
experiment11
r = 0.939
simulation12
experiment12
r = 0.900

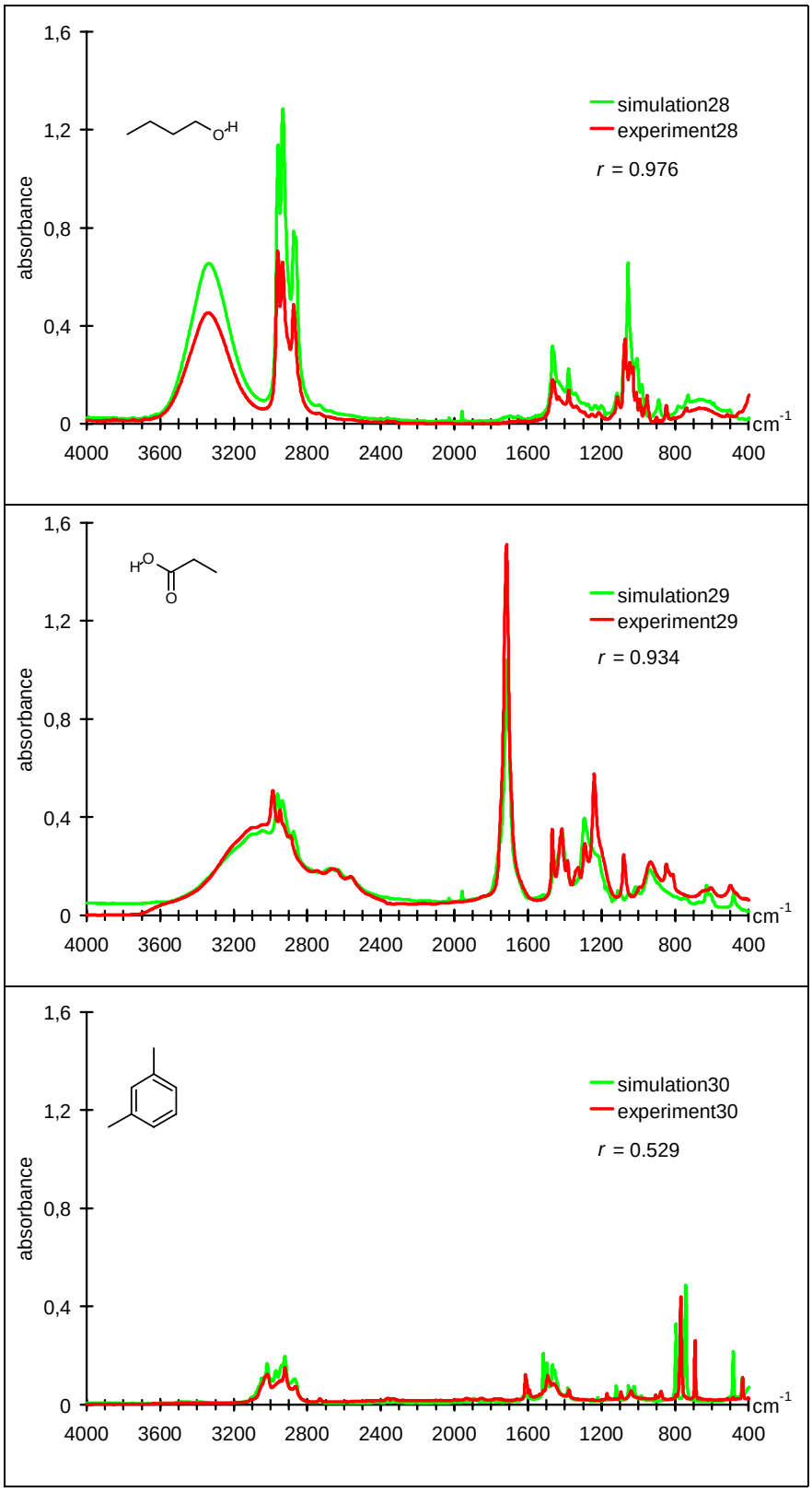

absorbance
cm-1
simulation13
experiment13
r = 0.793
simulation14
experiment14
r = 0.288
simulation15
experiment15
r = 0.259

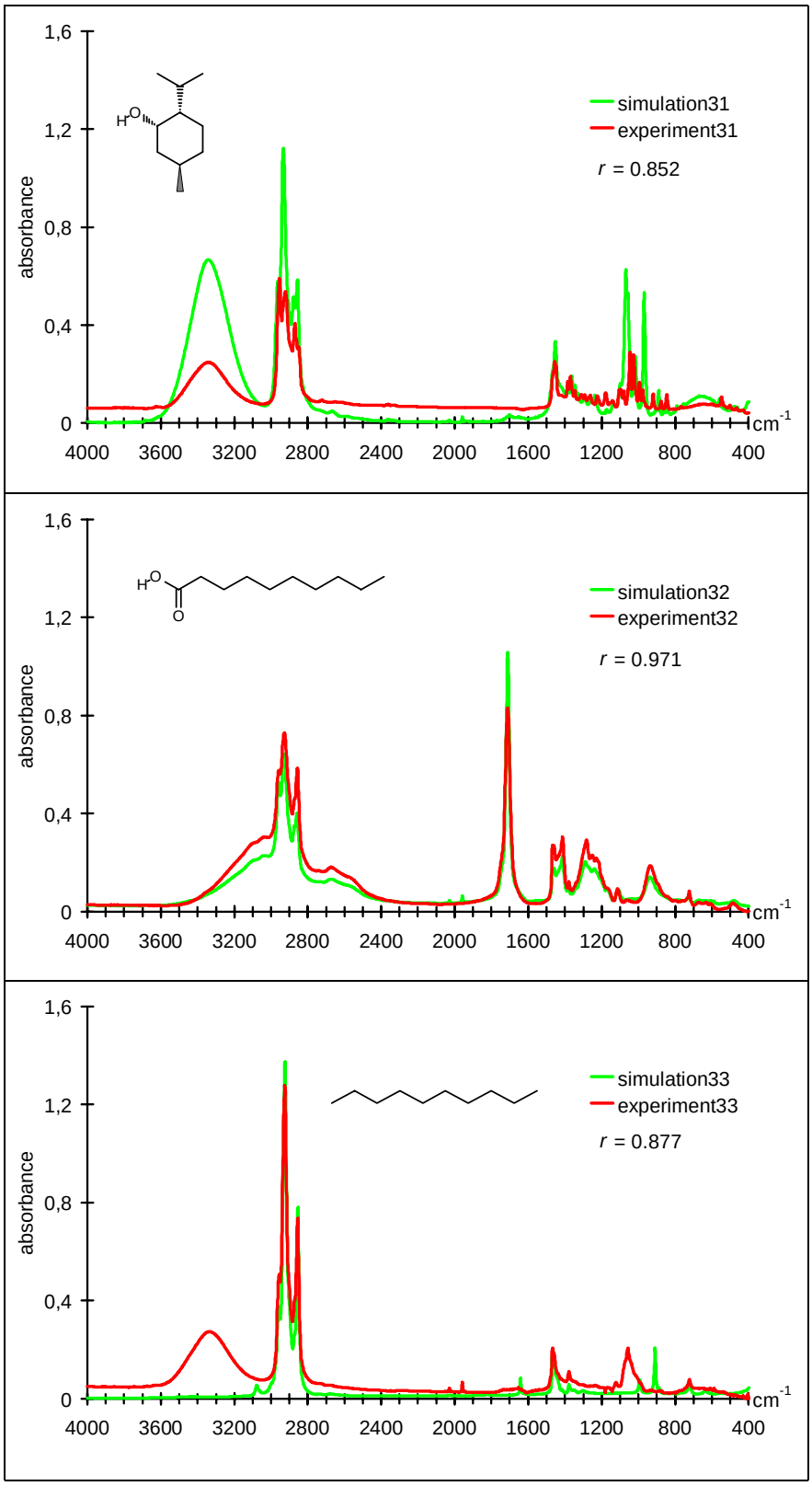

absorbance
1.6
1.2
0.8
0.4
0
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation16
experiment16
r = 0.520
simulation17
experiment17
r = 0.920
simulation18
experiment18
r = 0.804

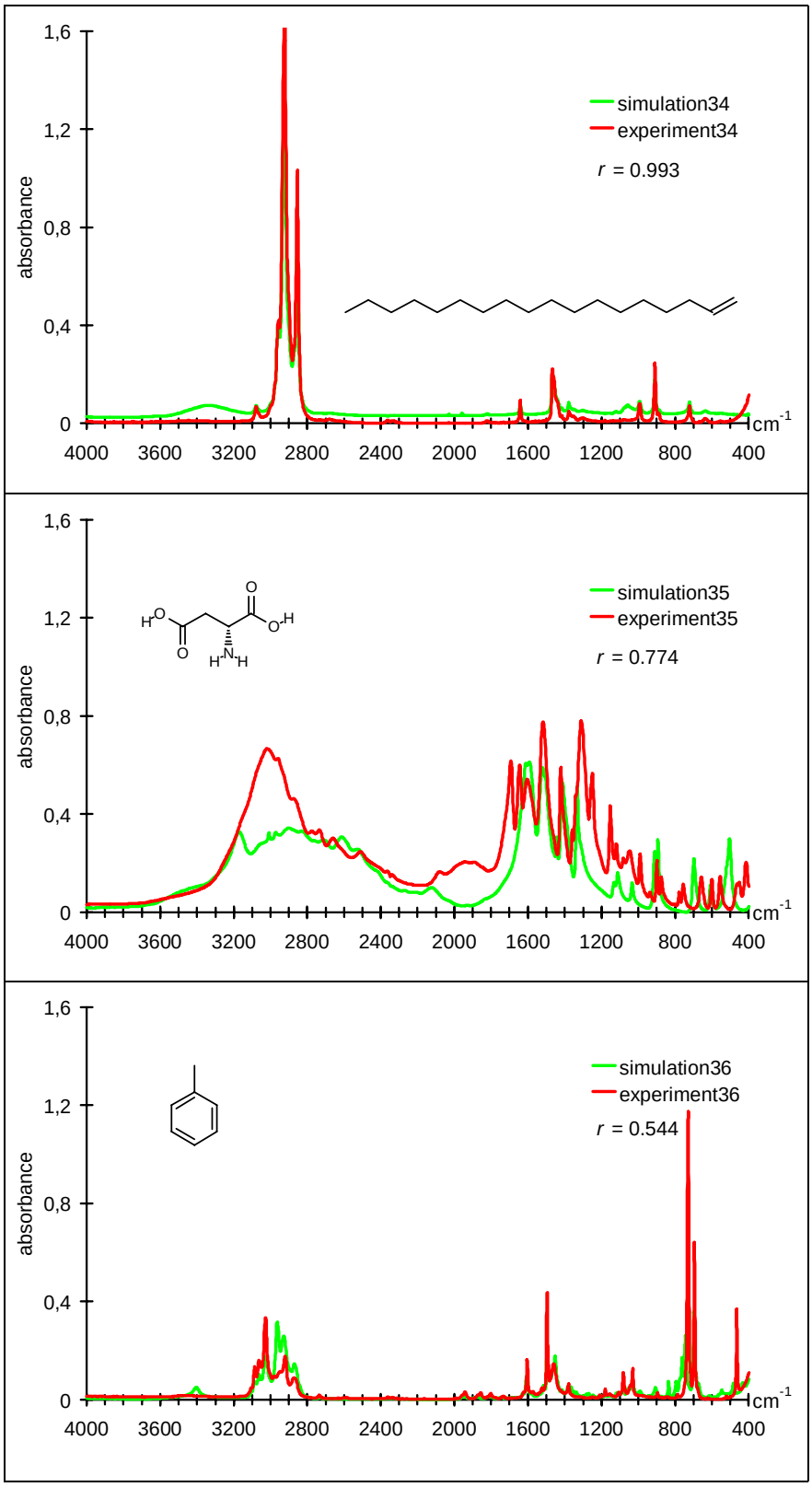

1.6
1.2
0.8
0.4
0
absorbance
4000
3600
3200
2800
2400
2000
1600
1200
800
400
cm-1
simulation19
experiment19
r = 0.746
simulation20
experiment20
r = 0.619
simulation21
experiment21
r = 0.996
H
N
O
O
O
H

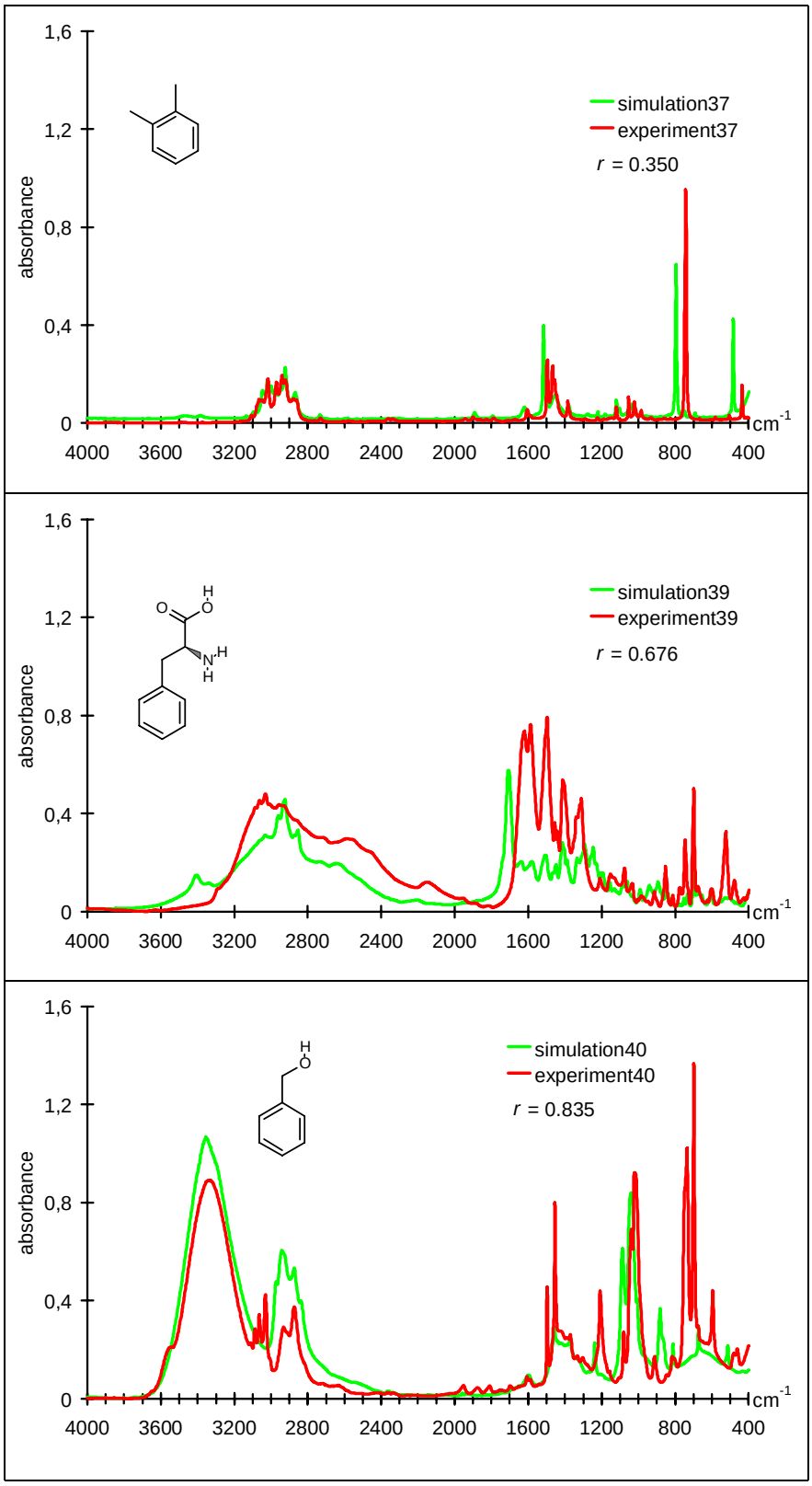

absorbance
1.6
1.2
0.8
0.4
0
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation22
experiment22
r = 0.277
simulation23
experiment23
r = 0.711
simulation24
experiment24
r = 0.951

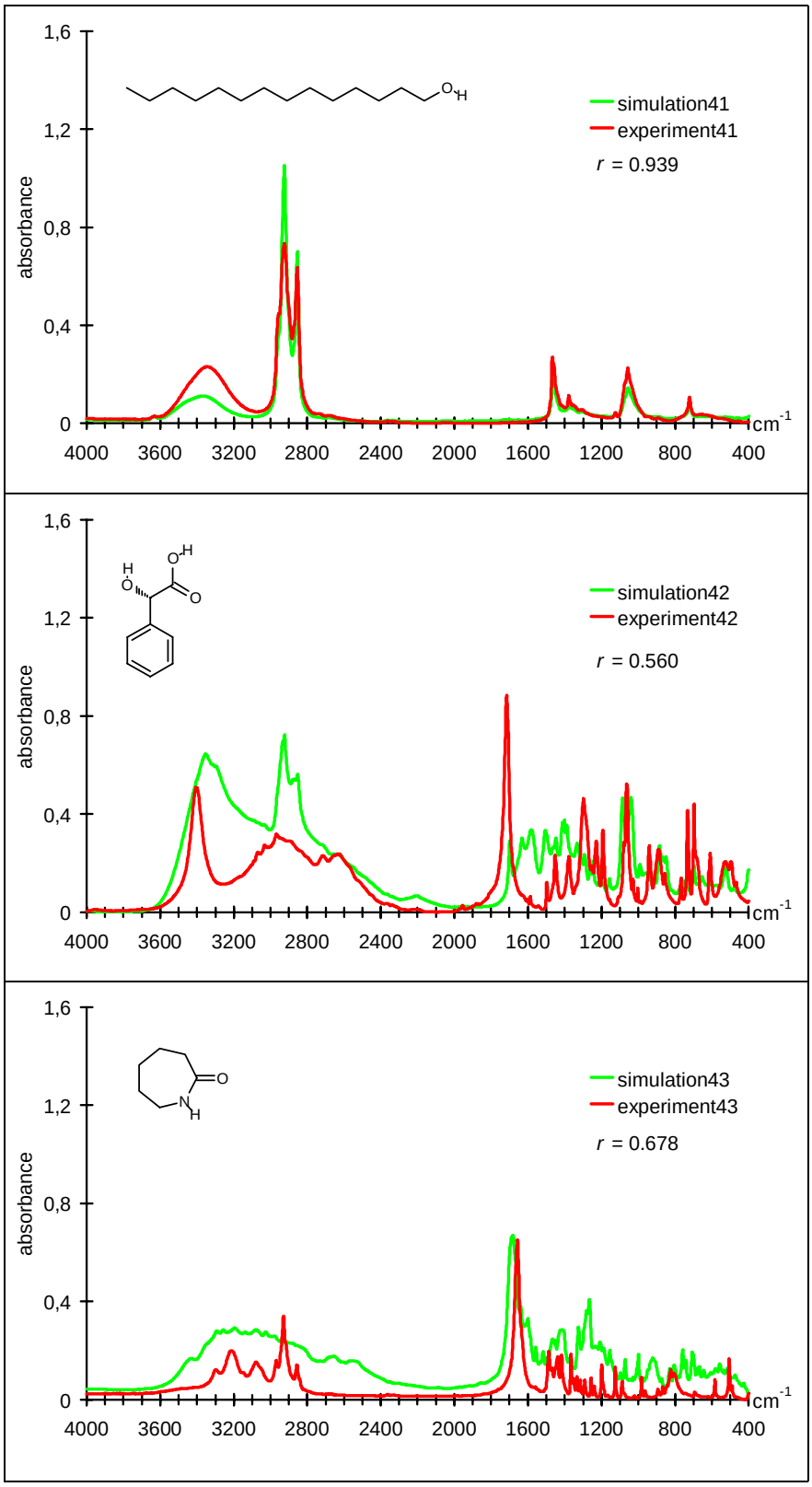

absorbance
1,6
1,2
0,8
0,4
0
4000
3600
3200
2800
2400
2000
1600
1200
800
400
cm-1
O
simulation25
experiment25
r = 0.291
simulation26
experiment26
r = 0.996
H
simulation27
experiment27
r = 0.754

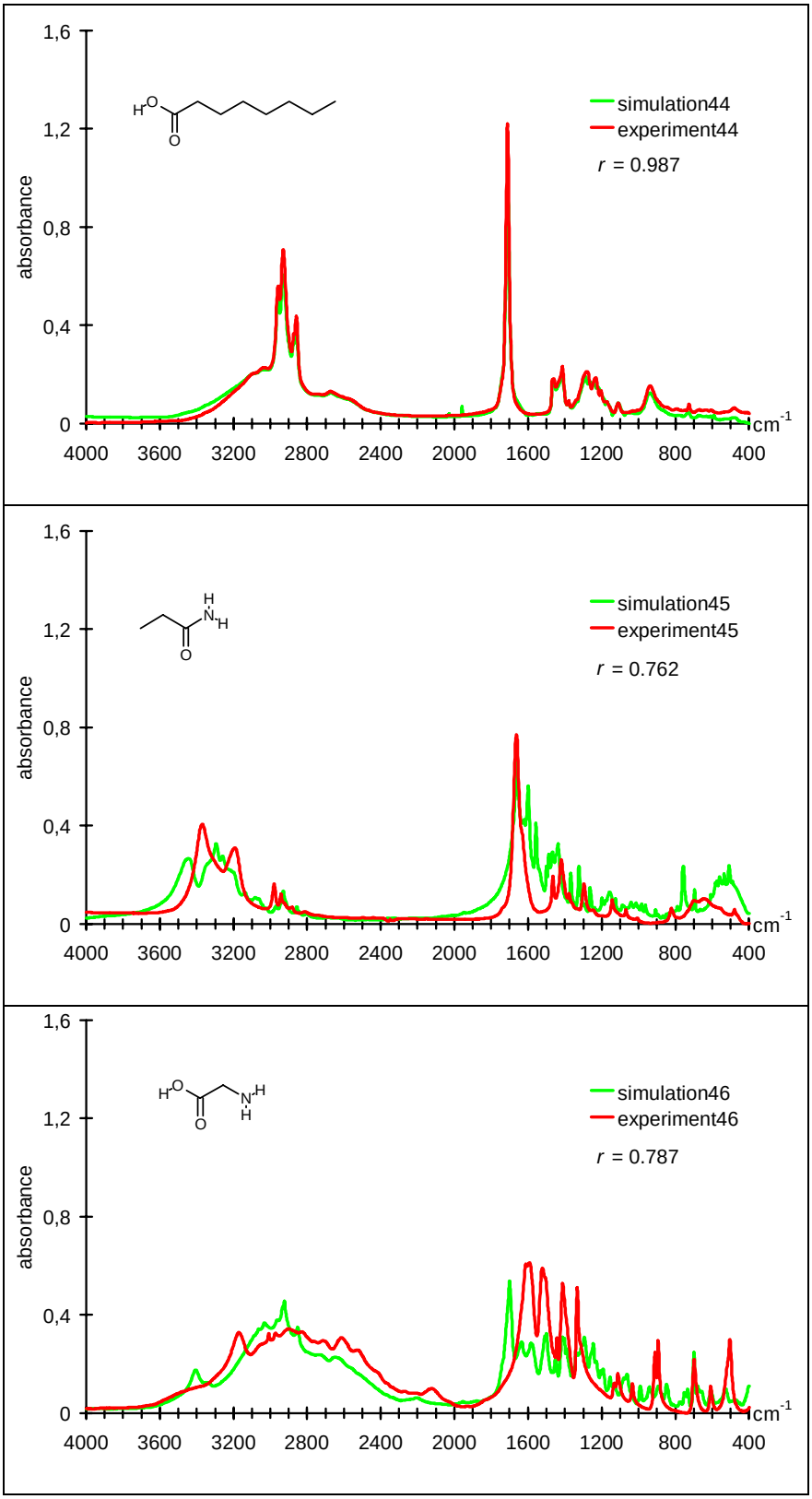

1,6
1,2
0,8
0,4
0
absorbance
simulation28
experiment28
r = 0.976
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation29
experiment29
r = 0.934
simulation30
experiment30
r = 0.529

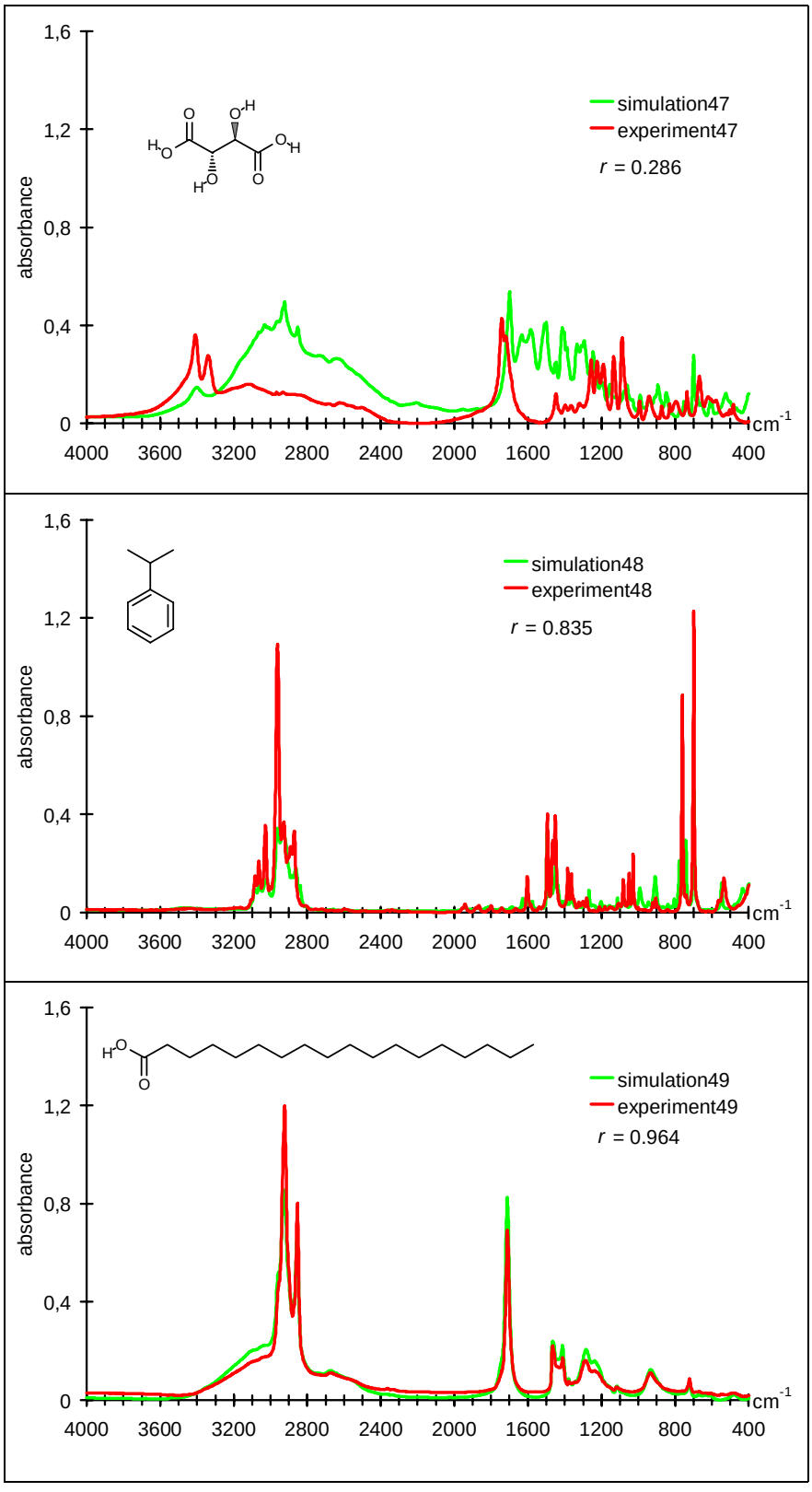

1,6
1,2
0,8
0,4
0
absorbance
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation31
experiment31
r = 0.852
simulation32
experiment32
r = 0.971
simulation33
experiment33
r = 0.877

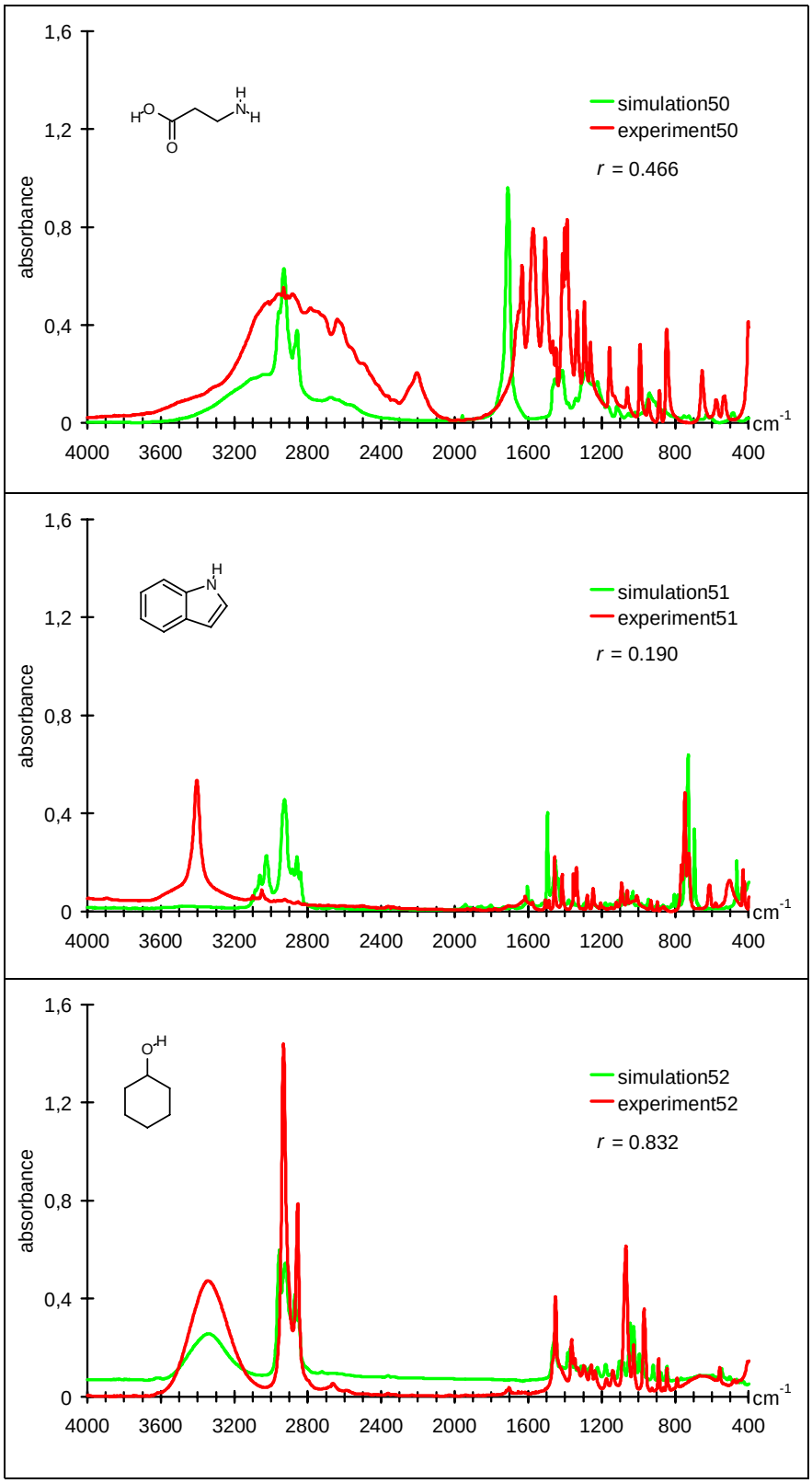

absorbance
1,6
1,2
0,8
0,4
0
4000
3600
3200
2800
2400
2000
1600
1200
800
400
cm-1
simulation34
experiment34
r = 0.993
simulation35
experiment35
r = 0.774
simulation36
experiment36
r = 0.544

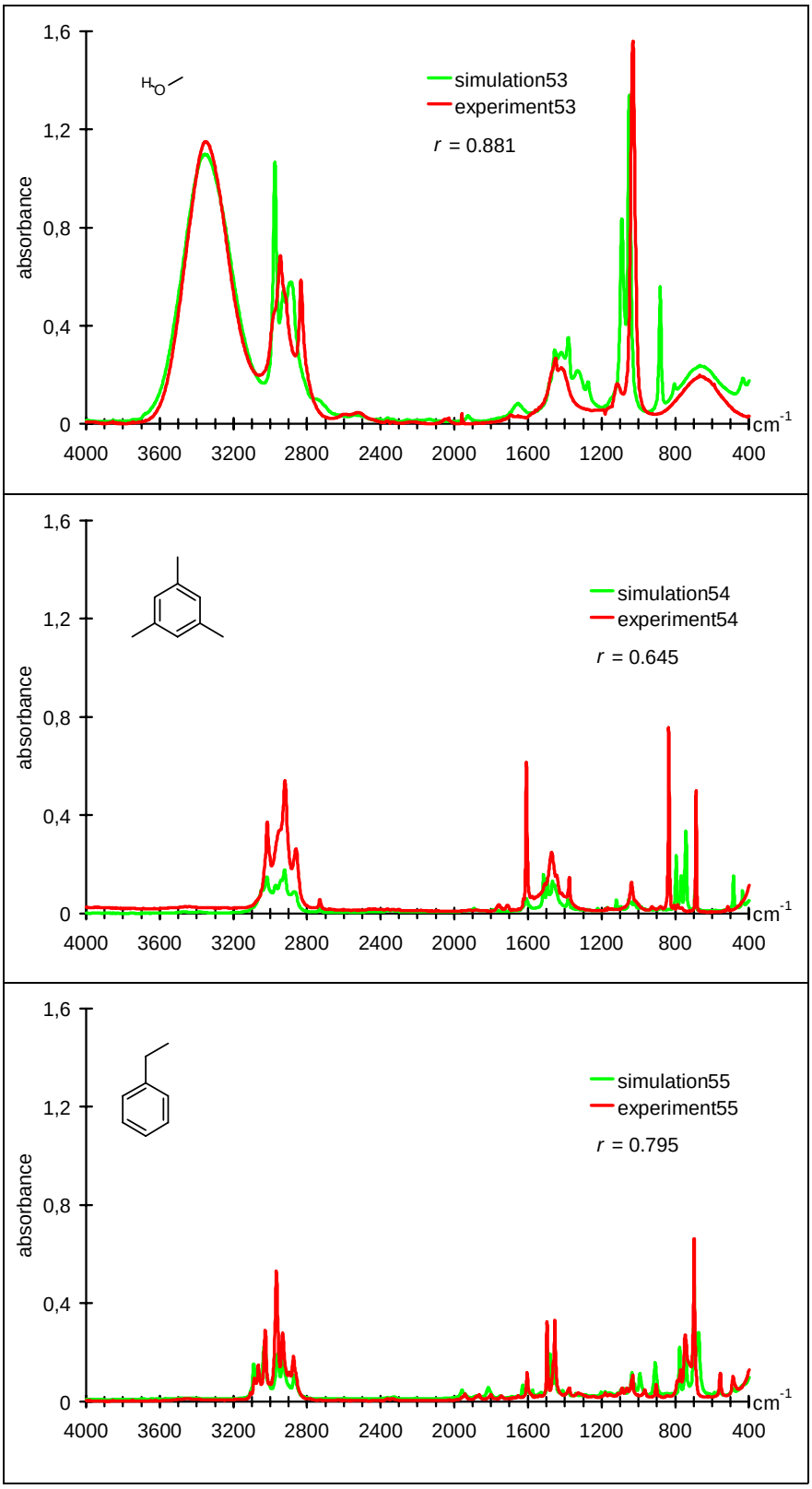

simulation37
experiment37
r = 0.350
absorbance
cm-1
simulation39
experiment39
r = 0.676
absorbance
cm-1
simulation40
experiment40
r = 0.835
absorbance
cm-1

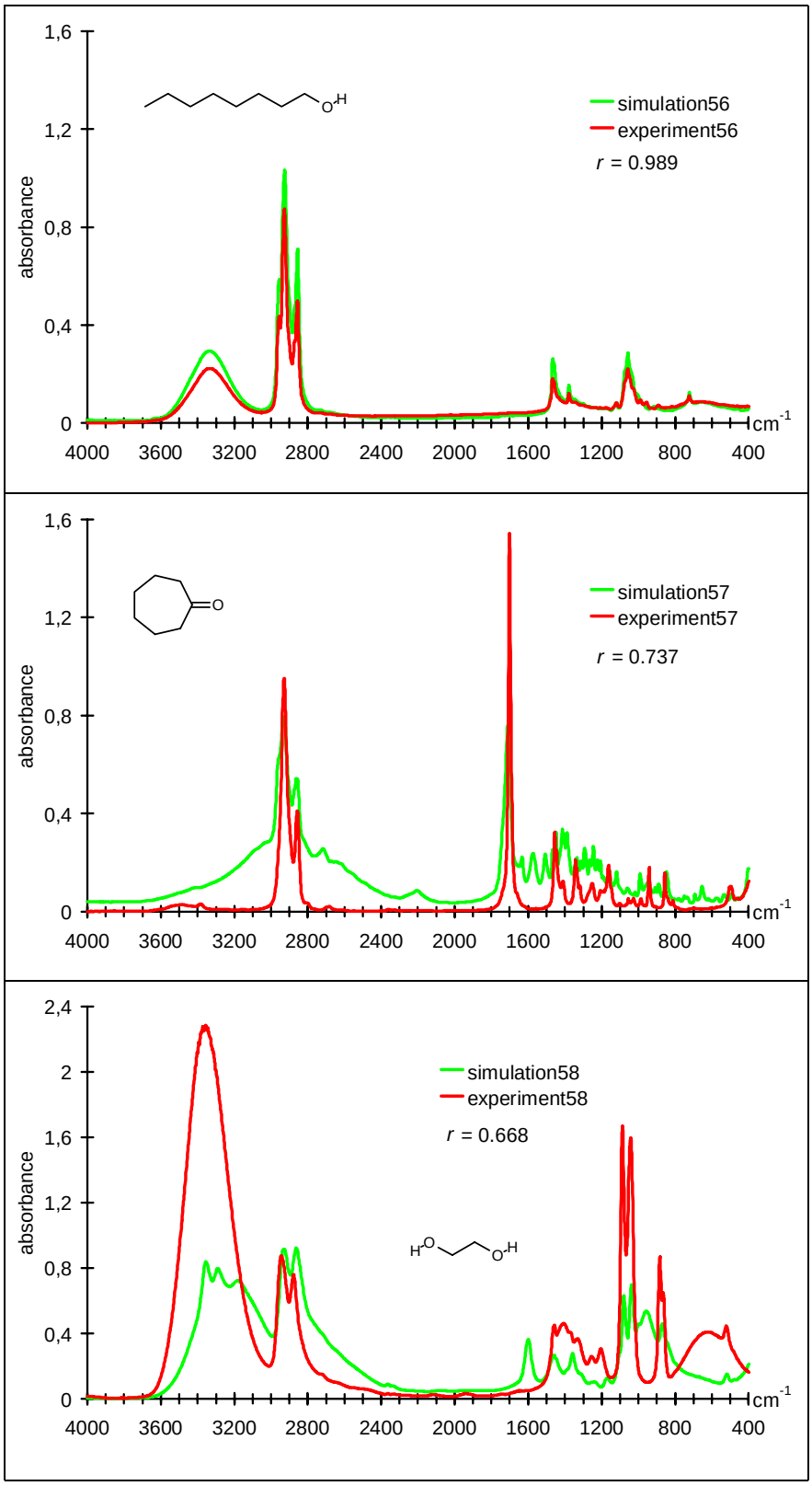

1,6
1,2
0,8
0,4
0
absorbance
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation41
experiment41
r = 0.939
simulation42
experiment42
r = 0.560
simulation43
experiment43
r = 0.678

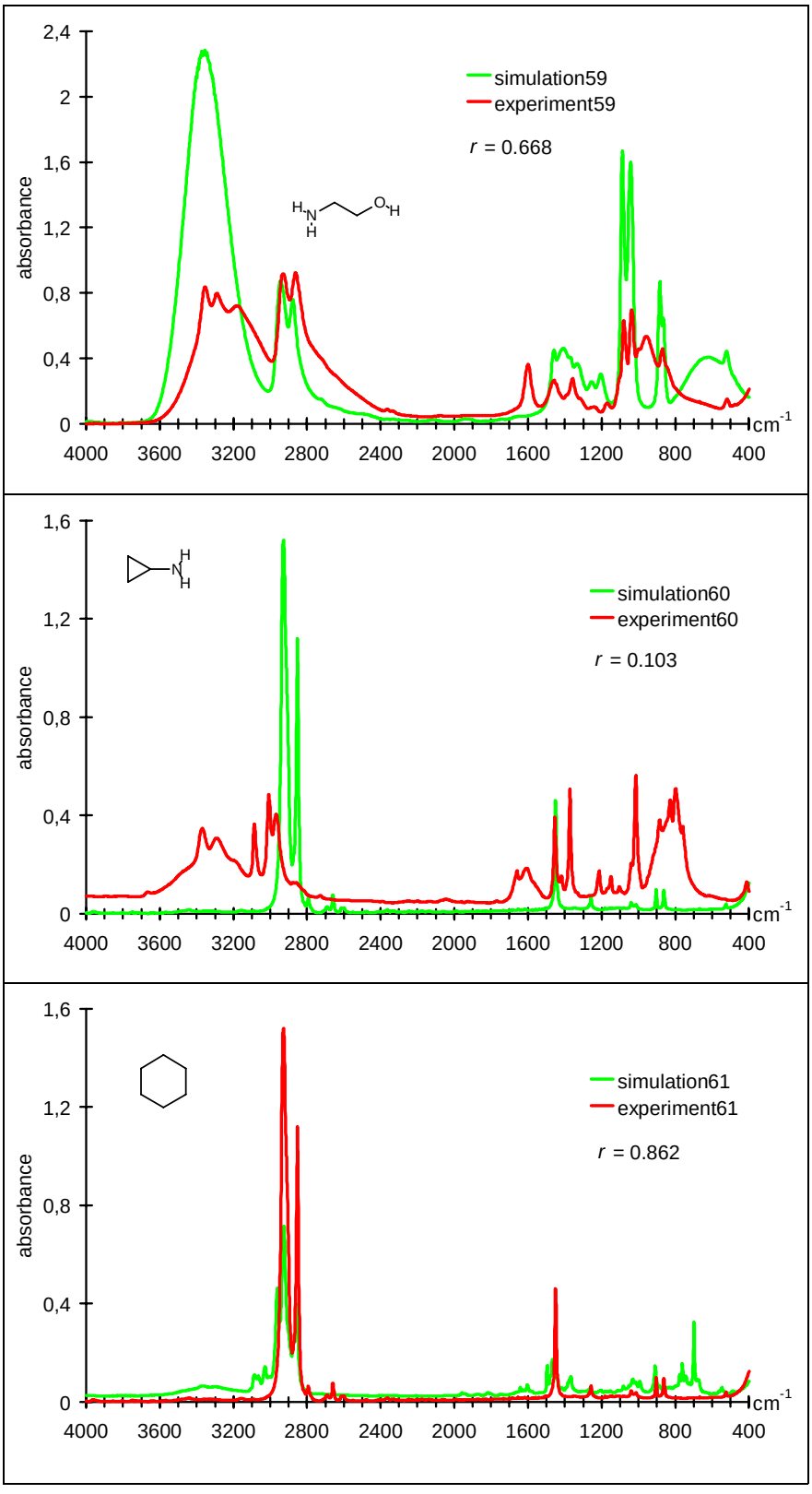

1,6
1,2
0,8
0,4
0
absorbance
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation44
experiment44
r = 0.987
simulation45
experiment45
r = 0.762
simulation46
experiment46
r = 0.787

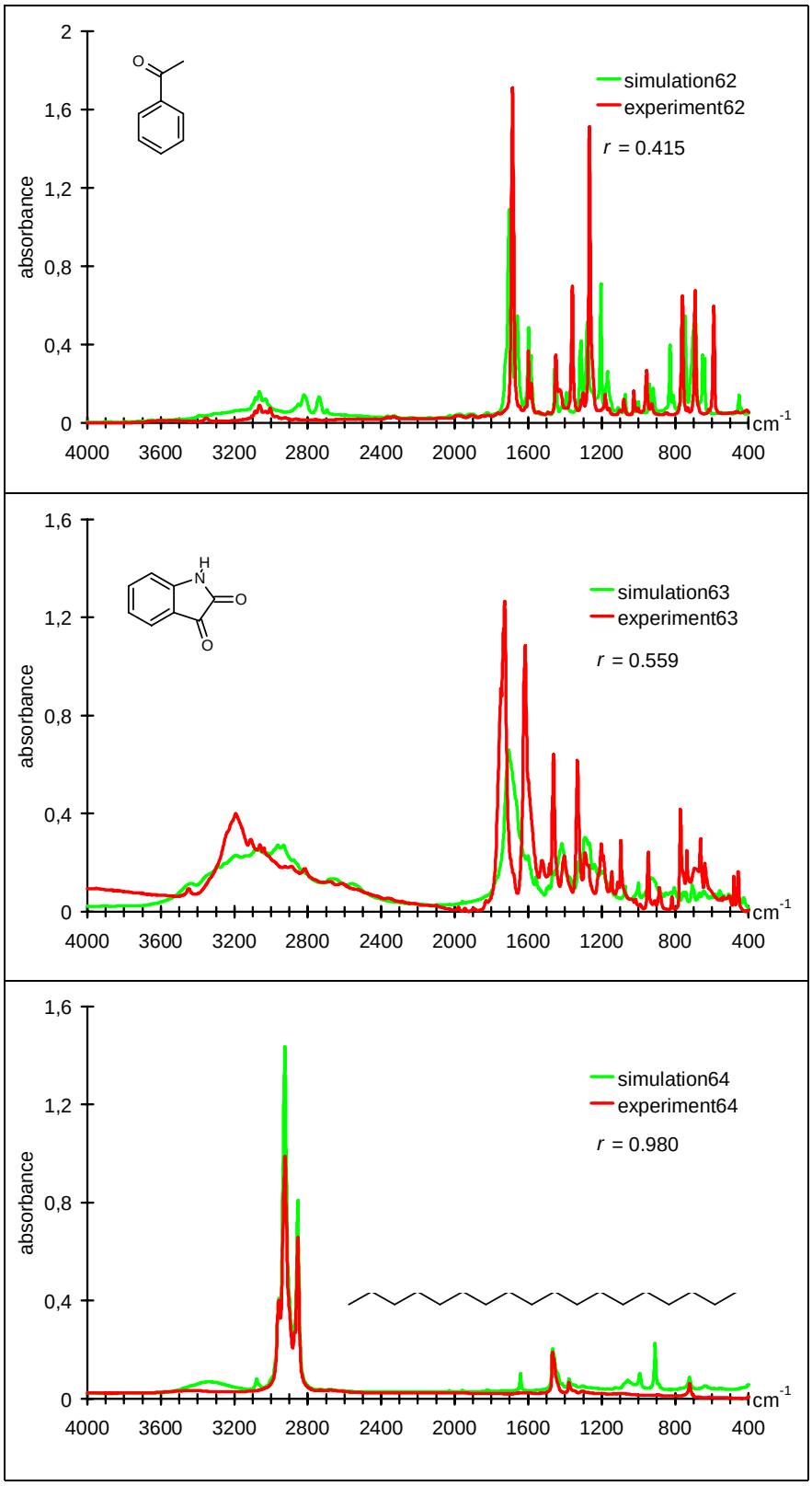

simulation47
experiment47
r = 0.286
absorbance
cm-1
simulation48
experiment48
r = 0.835
absorbance
cm-1
simulation49
experiment49
r = 0.964
absorbance
cm-1
1,6
1,2
0,8
0,4
0
4000
3600
3200
2800
2400
2000
1600
1200
800
400

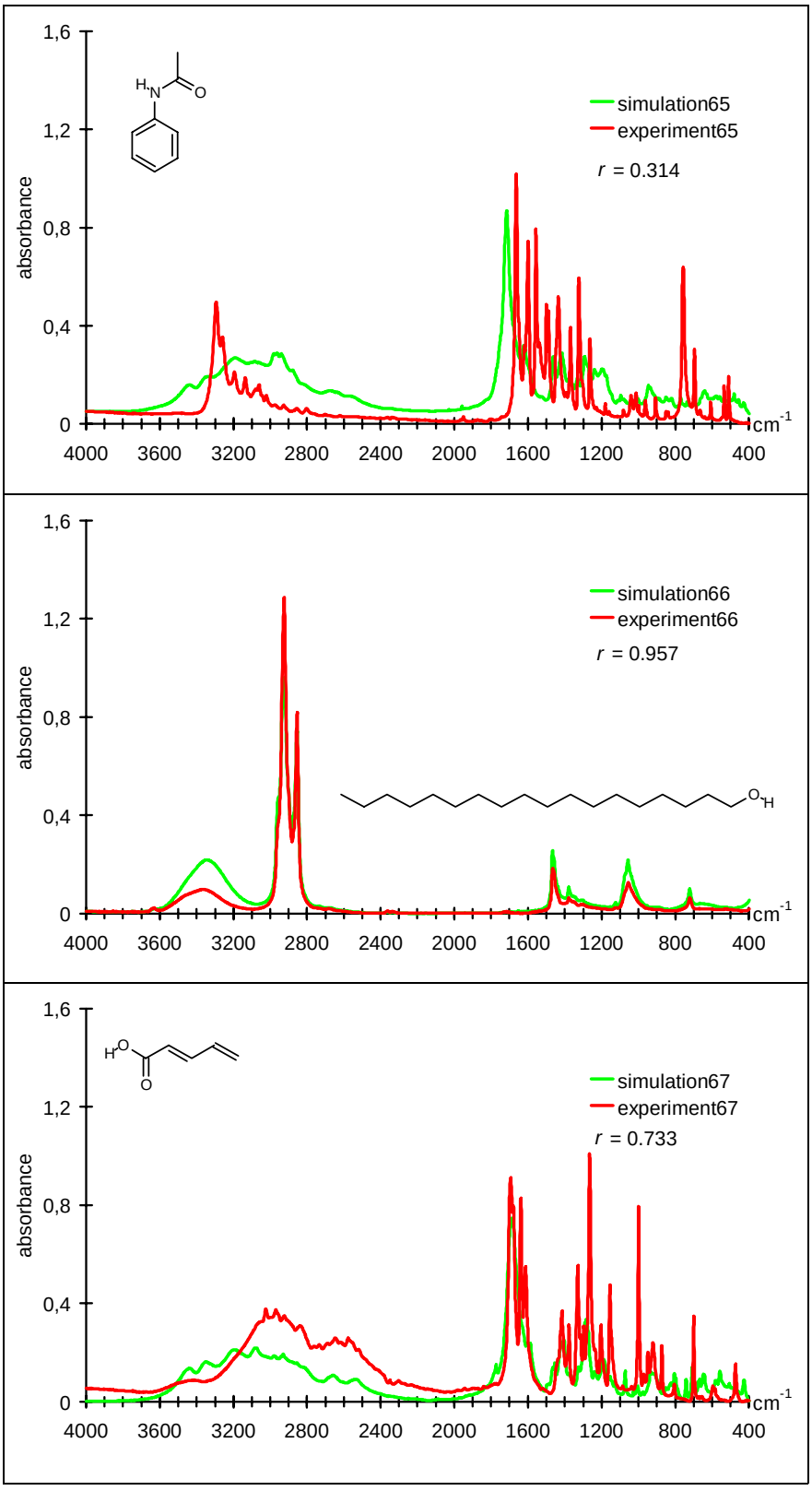

simulation50
experiment50
r = 0.466
simulation51
experiment51
r = 0.190
simulation52
experiment52
r = 0.832
absorbance
cm-1
1,6
1,2
0,8
0,4
0
4000
3600
3200
2800
2400
2000
1600
1200
800
400

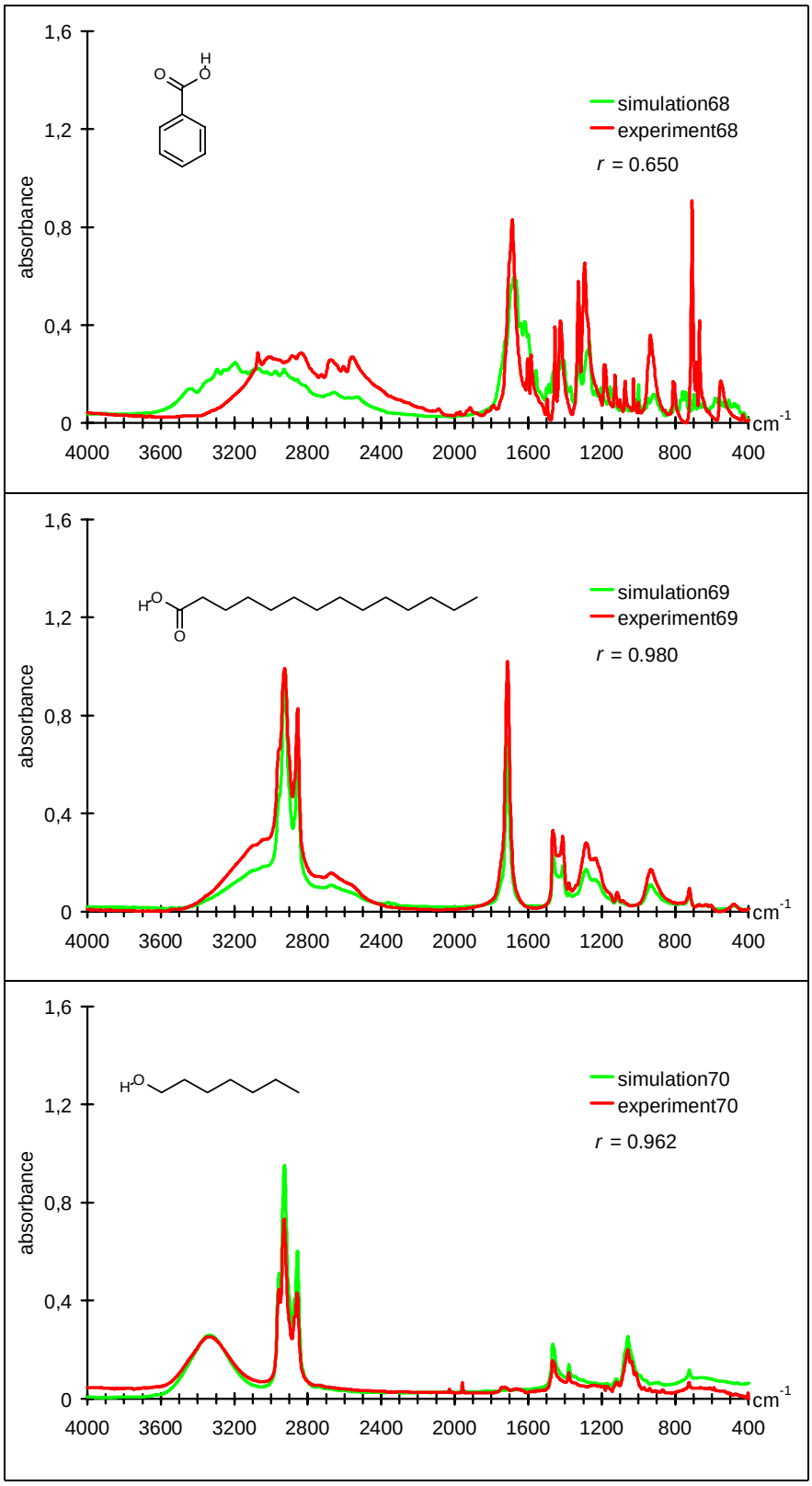

1,6
1,2
0,8
0,4
0
absorbance
simulation53
experiment53
r = 0.881
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation54
experiment54
r = 0.645
simulation55
experiment55
r = 0.795

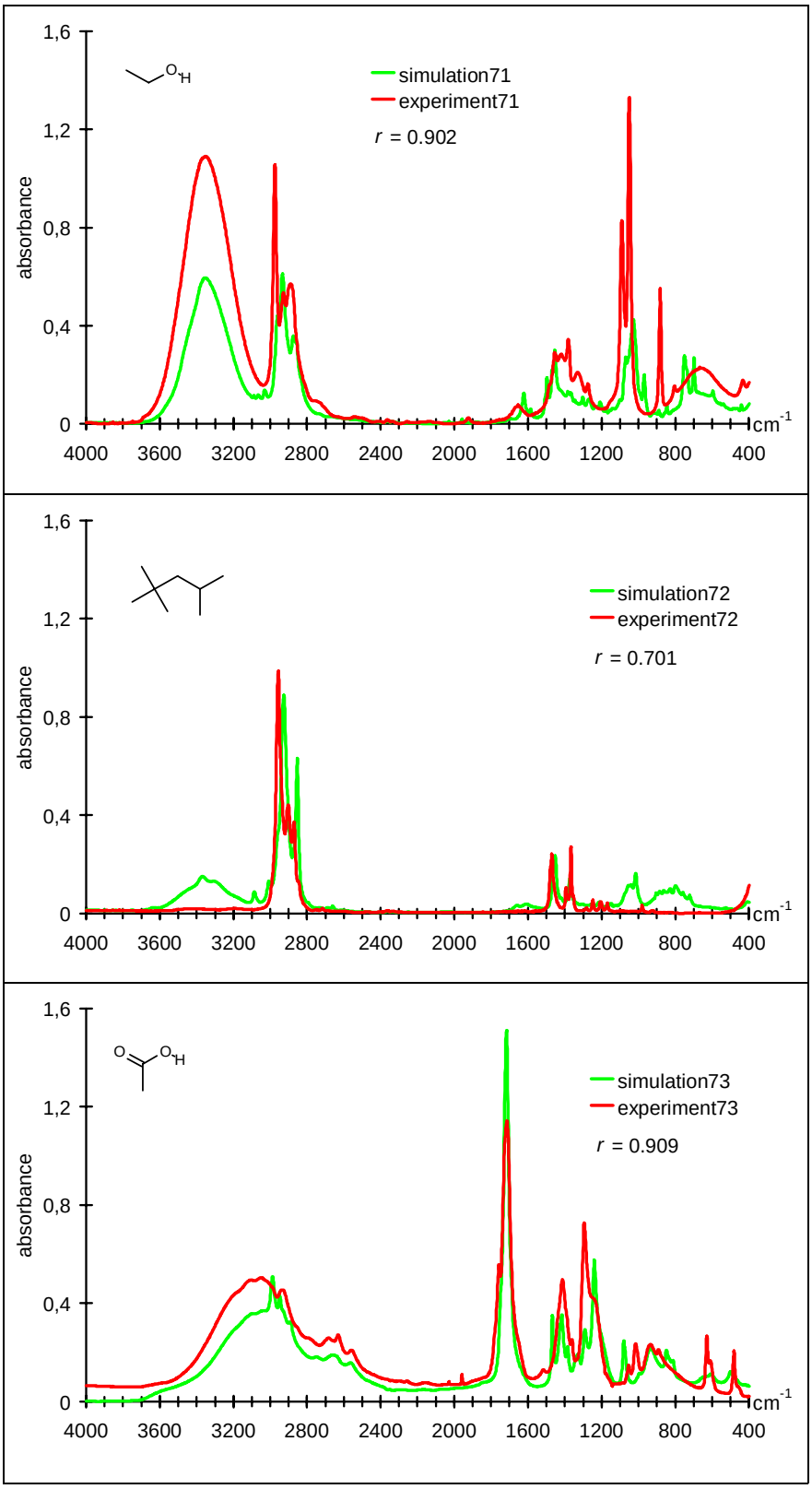

absorbance
cm-1
simulation56
experiment56
r = 0.989
simulation57
experiment57
r = 0.737
simulation58
experiment58
r = 0.668

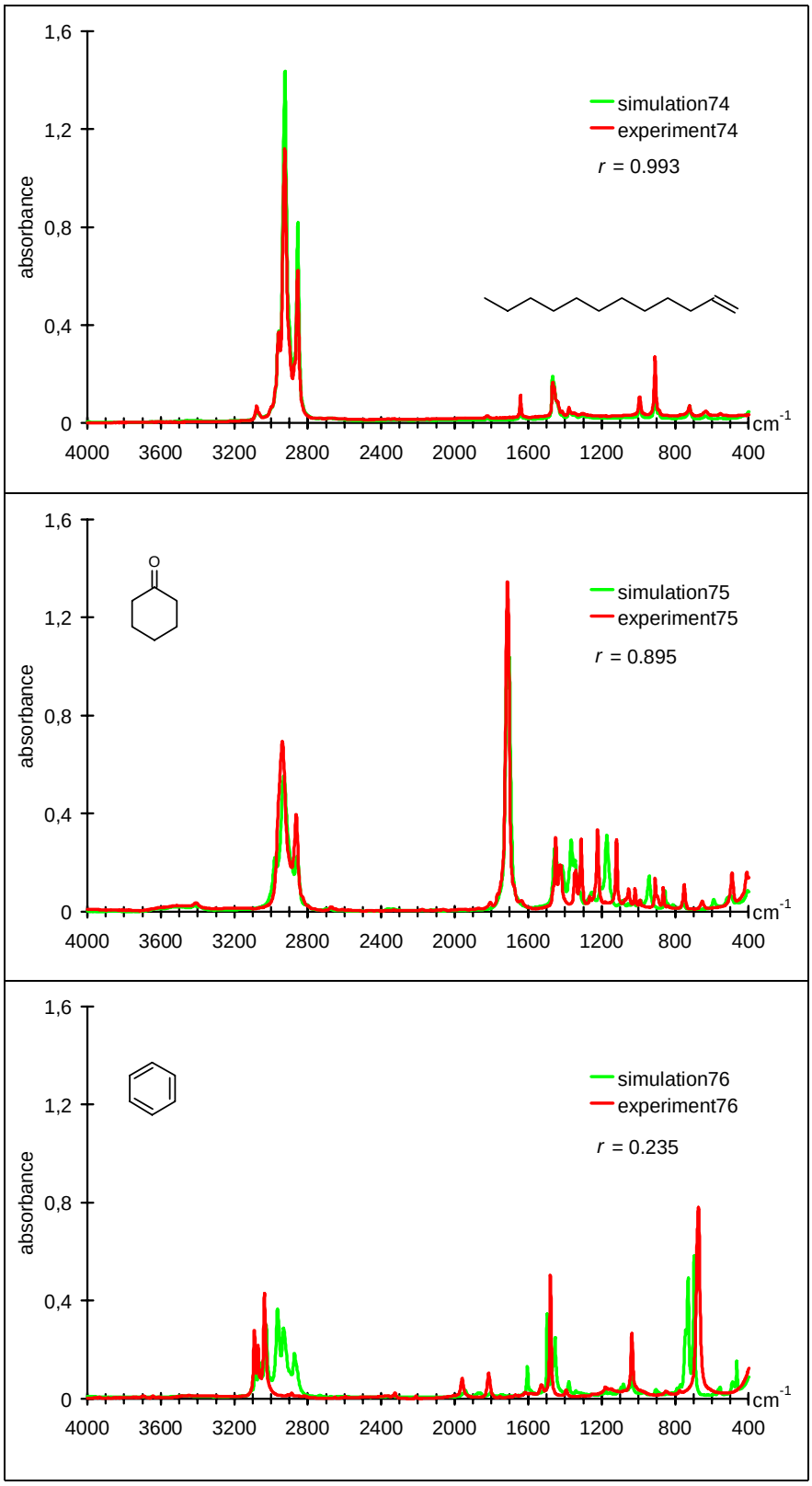

absorbance
2,4
2
1,6
1,2
0,8
0,4
0
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation59
experiment59
r = 0.668
simulation60
experiment60
r = 0.103
simulation61
experiment61
r = 0.862

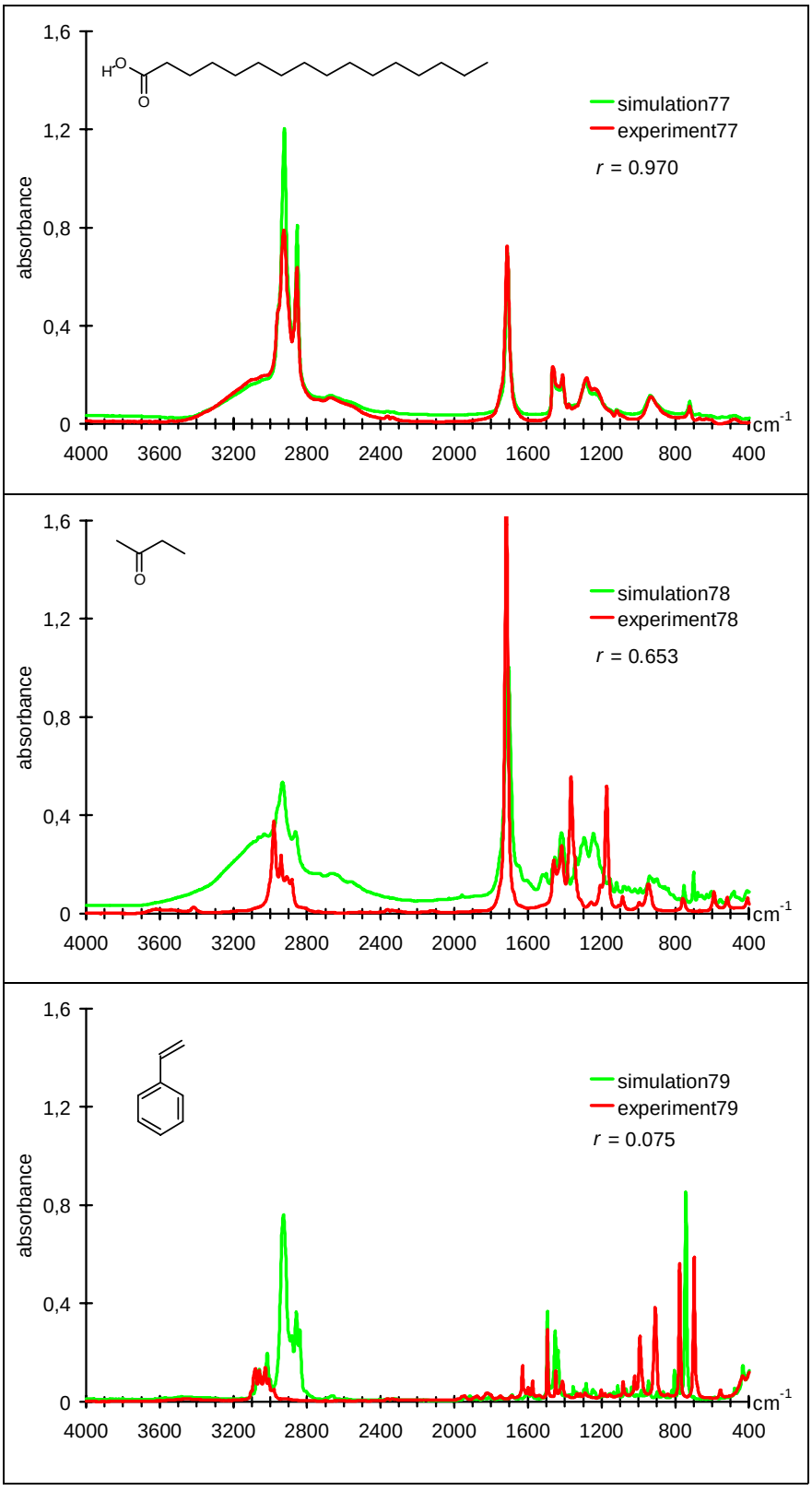

simulation62
experiment62
r = 0.415
absorbance
cm-1
simulation63
experiment63
r = 0.559
absorbance
cm-1
simulation64
experiment64
r = 0.980
absorbance
cm-1

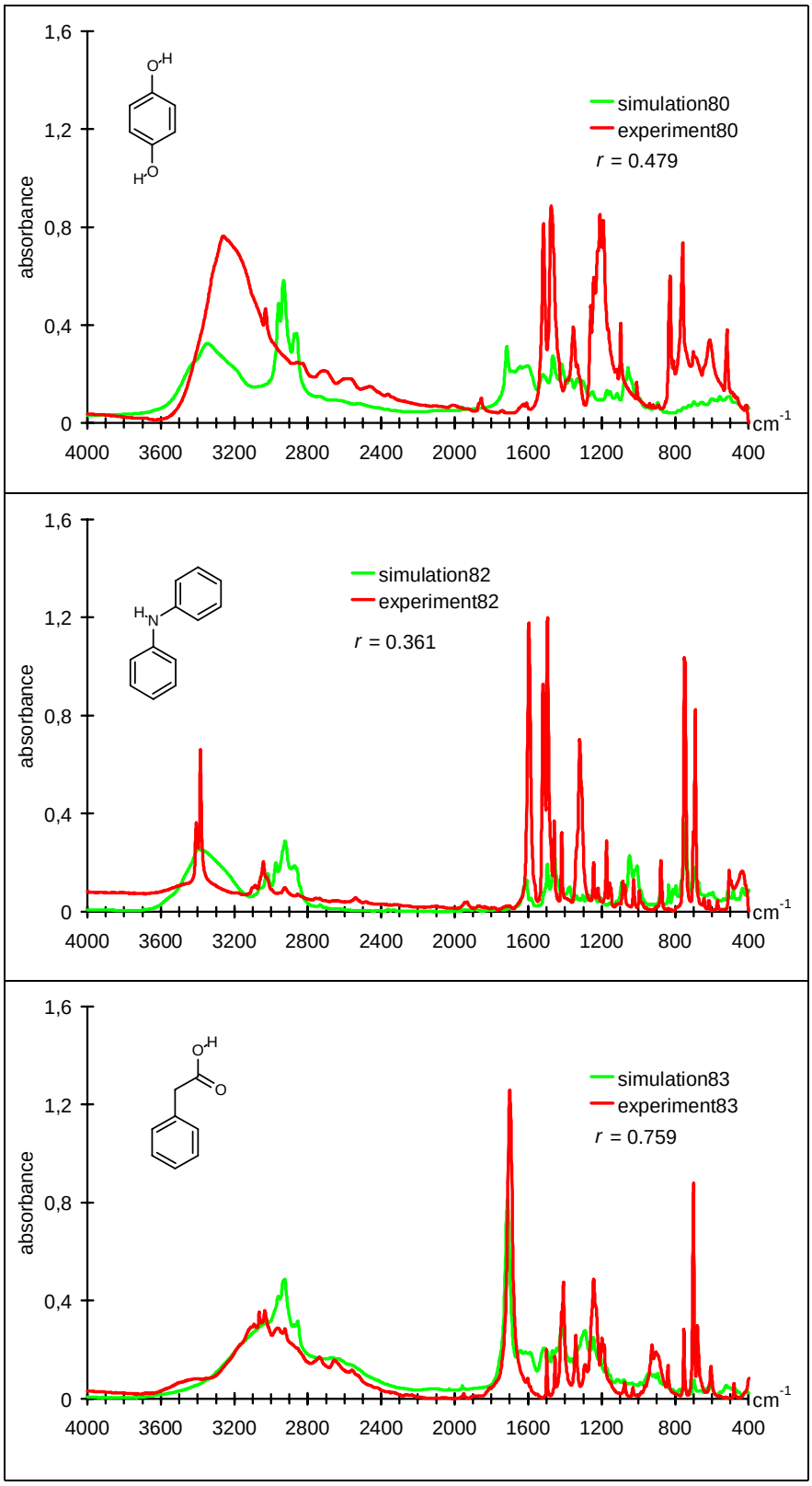

1,6
1,2
0,8
0,4
0
absorbance
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation65
experiment65
r = 0.314
simulation66
experiment66
r = 0.957
simulation67
experiment67
r = 0.733


1,6
1,2
0,8
0,4
0
absorbance
simulation68
experiment68
r = 0.650
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation69
experiment69
r = 0.980
simulation70
experiment70
r = 0.962


simulation71
experiment71
r = 0.902
absorbance
cm-1
simulation72
experiment72
r = 0.701
absorbance
cm-1
simulation73
experiment73
r = 0.909
absorbance
cm-1
1,6
1,2
0,8
0,4
0
4000
3600
3200
2800
2400
2000
1600
1200
800
400

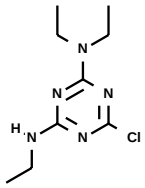
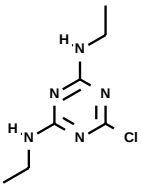
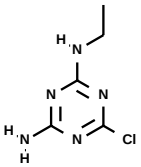
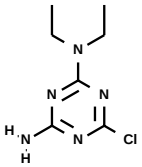
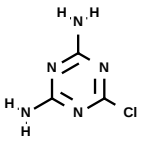
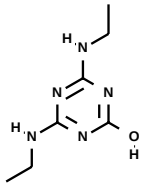
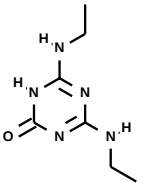
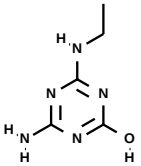
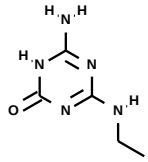
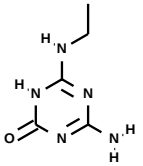
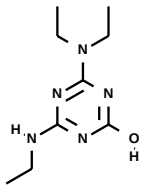
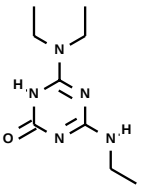
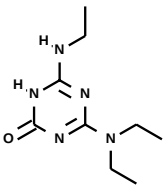
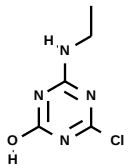
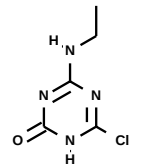
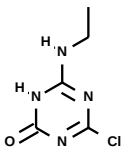
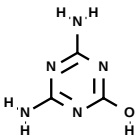
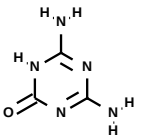
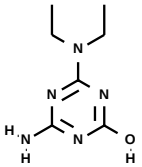
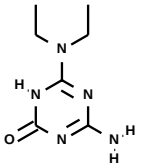
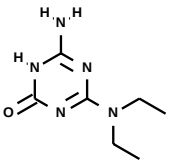
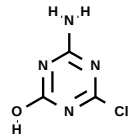
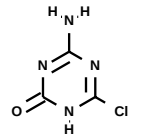
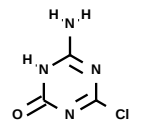
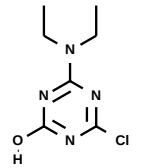
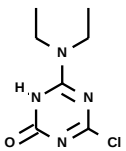
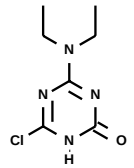
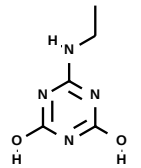
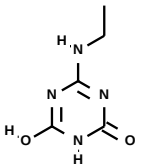
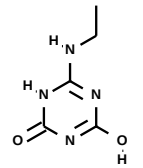
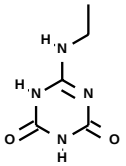
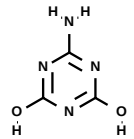
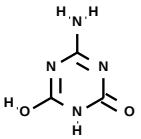
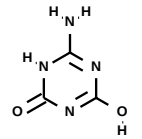
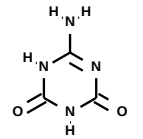

1,6
1,2
0,8
0,4
0
absorbance
simulation74
experiment74
r = 0.993
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
O
simulation75
experiment75
r = 0.895
simulation76
experiment76
r = 0.235


1,6
1,2
0,8
0,4
0
absorbance
4000 3600 3200 2800 2400 2000 1600 1200 800 400
cm-1
simulation77
experiment77
r = 0.970
simulation78
experiment78
r = 0.653
simulation79
experiment79
r = 0.075

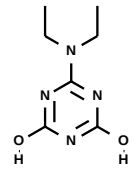
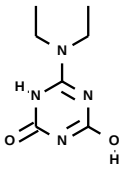
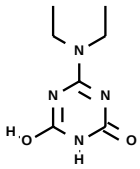
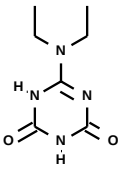
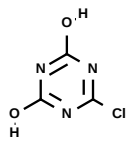
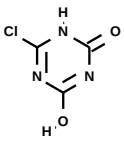
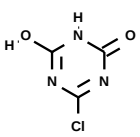
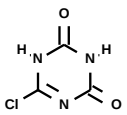
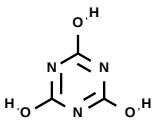
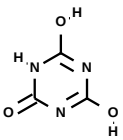
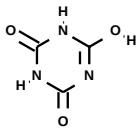
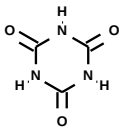

simulation80
experiment80
r = 0.479
simulation82
experiment82
r = 0.361
simulation83
experiment83
r = 0.759
absorbance
cm-1
1,6
1,2
0,8
0,4
0
4000
3600
3200
2800
2400
2000
1600
1200
800
400

A.6 Abbauprodukte des Trietazins

#1 /home/pauls/simu/kmap/dyn/radial/trietazin/trietazin_newname.ctx [1 – 35]

Edukt 	**I_1** 	**II_1** 	**I_2** 	**III_3** 
II_2 	**II_2_t1** 	**III_2** 	**III_2_t1** 	**III_2_t2** 
I_3 	**I_3_t1** 	**I_3_t2** 	**III_3** 	**III_3_t1** 
III_3_t2 	**IV_1** 	**IV_1_t1** 	**II_3** 	**II_3_t1** 
II_3_t2 	**IV_2** 	**IV_2_t1** 	**IV_2_t2** 	**II_4** 
II_4_t1 	**II_4_t2** 	**IV_3** 	**IV_3_t1** 	**IV_3_t2** 
IV_3_t3 	**V_1** 	**V_1_t1** 	**V_1_t2** 	**V_1_t3** 

#2 /home/pauls/simu/kmap/dyn/radial/trietazin/trietazin_newname.ctx [1 − 35]

III_4 	**III_4_t1** 	**III_4_t2** 	**III_4_t3** 	**V_2** 
V_2_t1 	**V_2_t2** 	**V_2_t3** 	**VI_1** 	**VI_1_t1** 
VI_1_t2 	**V_1_t3** 			

A.7 Publikationen

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. Selzer, P.; Schuur, J.H.; Gasteiger, J. Simulation of IR Spectra with Neural Networks Using the 3D-MoRSE Code, in *Software Development in Chemistry 10*; J. Gasteiger (Ed.); Gesellschaft Deutscher Chemiker: Frankfurt am Main, **1996**; p. 293.
3. 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.
4. Selzer, P. Excel im Accord - Chemieorientierte Tabellenkalkulation, *Nachr. Chem. Tech. Lab.*, **1997**, *45*, 296-303.
5. Schuur, J.; Selzer, P.; Steinhauer, V.; Gasteiger, J. Kooperative, rechnergestützte IR-Spektreninterpretation - neue Wege für die Infrarotspektroskopie, *GIT Labor-Fachzeitschrift*, **1997**, *3*, 283-286.
6. 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.
7. Schuur, J.; Selzer, P.; Steinhauer, V.; Gasteiger, J. 3D Structure Coding Opens New Applications for IR Spectroscopy, *Linking and Interpreting Spectra through Molecular Structures LISMS*, Chichester, **1997**, 359, 15-28.
8. Kostka, T.; Selzer, P.; Gasteiger, J. Computer-Assisted Prediction of the Degradation Products and Infrared Spectra of s-Triazine Herbicides, in *Software Development in Chemistry 11*; G. Fels, V. Schubert (Eds.); Gesellschaft Deutscher Chemiker: Frankfurt am Main, **1997**; p. 226.
9. Gasteiger, J.; Kostka, T.; Selzer, P.; Bauerschmidt, S.; Höllering, R.; Steinhauer, L. Computer Methods for the Prediction and Identification of Degradation Products of Chemicals Using IR Spectra Simulation, *Proceed. ECO-INFORMA'97, Eco-Informa Press*, **1997**, 509-513.
10. Kostka, T.; Selzer, P.; Gasteiger, J. Rapid Identification of Herbicide Degradation Products Using Reaction Prediction and Infrared Spectra Simulation Methods Proceed. *ECO-INFORMA'97, Eco-Informa Press*, **1997**, 514-516.
11. Selzer, P.; Hemmer, M.; Schuur, J.; Steinhauer, V.; Gasteiger, J. TeleSpek - Telekooperation in der Spektroskopie, *Nachr. Chem. Tech. Lab.*, **1998**, *46*, A78-A82.
12. Selzer, P. ModelMaker-Systemmodellierung auf dem PC, *Nachr. Chem. Tech. Lab.*, **1998**, 46, 652-656.
13. Selzer, P. Infrared Data Correlations with Chemical Structure, in *Encyclopedia of Computational Chemistry*, P. v. R. Schleyer (Ed.), Wiley, Chichester, 1997, in Druck.

A.8 Lebenslauf

Name	Paul Selzer
Geburtsdatum und -ort	09. Juni 1968 in Ostrau (Tschechien)
Eltern	Adolf Selzer und Edeltraud Selzer, geb. Malkrab
Staatsangehörigkeit	deutsch
Familienstand	ledig, keine Kinder

Schulbildung

09/1974 - 07/1978	Georg-Ledebour Grundschule Nürnberg
09/1978 - 06/1987	Pirckheimer-Gymnasium Nürnberg

Zivildienst

11/1987 - 06/1989	Rettungsdienst beim Bayerischen Roten Kreuz

Hochschulausbildung

10/1989 - 07/1994	Studium der Chemie an der Friedrich-Alexander Universität Erlangen-Nürnberg
08/1994 - 02/1995	Diplomarbeit bei Prof. Gasteiger am Computer-Chemie-Centrum des Instituts für Organische Chemie der Universität Erlangen-Nürnberg zu dem Thema „Strukturcodierung organischer Moleküle zur Simulation von Infrarotspektren“
seit 03/1995	Promotionsarbeit bei Prof. Gasteiger