

# Entwicklung neuer QSAR-Methoden und deren Anwendung an Dopaminrezeptorantagonisten

**Dissertation**

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

**Dipl. Pharm. Mathias Weigt**

aus

Halle

Bonn, Februar 2006





Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät  
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Prof. Dr. Michael Wiese
2. Referent: PD Dr. Matthias U. Kassack

Tag der Promotion: 16.8.2006

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn  
[http://hss.ulb.uni-bonn.de/diss\\_online](http://hss.ulb.uni-bonn.de/diss_online) elektronisch publiziert.

Erscheinungsjahr: 2006



# Verfassererklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Bonn, 16. Februar 2006

Mathias Weigt



Ich danke Prof. Wiese für die Möglichkeit, unter seiner Leitung diese Arbeit anfertigen zu können. Mit vielen Anregungen und steter konstruktiver Kritik nahm er entscheidend Einfluss auf ihren Erfolg.

Herrn PD Dr. Kassack danke ich sehr für die freundliche Übernahme des Koreferats.

Mein ganz besonderer Dank gilt außerdem Cristina Ferrari, Claudia Hadtstein, Alexandra Hamacher, Frau Prof. Ilza Pajeva und Christoph Globisch für die vielen fachlichen Beiträge und Diskussionen.

Nicht zuletzt danke ich allen, die durch Unterstützung und Anregung zum Gelingen dieser Dissertation beigetragen haben.

Für Melanie





# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
1.1	Zielsetzung der Arbeit . . . . .	8
<b>2</b>	<b>Datenauswahl</b>	<b>9</b>
2.1	Verfügbarkeit von Aktivitäts- und Affinitätsdaten . . . . .	10
2.2	Vergleichbarkeit von Aktivitäts- und Affinitätsdaten . . . . .	11
2.2.1	Vergleichbarkeit zwischen den Spezies Ratte u. Mensch . . . . .	13
2.2.2	Pufferabhängigkeit der Bindungskonstanten . . . . .	18
2.3	Begründung der Auswahl . . . . .	20
<b>3</b>	<b>Erweiterung der CoMFA-Methode mit automatischer PLS</b>	<b>23</b>
3.1	Möglichkeiten zur Konformationsanalyse . . . . .	24
3.1.1	Systematische Suche . . . . .	24
3.1.2	Stochastische Suche . . . . .	26
3.1.3	Entwicklung eines neuen Verfahrens zum Ähnlichkeitsclustering	27
3.2	CoMFA als Standardmethode der 3D-QSAR . . . . .	34
3.2.1	CoMFA Feldtypen . . . . .	35
3.2.2	Vorteile und Probleme des CoMFA-Verfahrens . . . . .	37
3.2.3	Verbesserung des CoMFA-Verfahrens mittels Automation . . .	39
3.3	Automatische PLS am Beispiel der D <sub>2</sub> - und D <sub>3</sub> -Rezeptorantagonisten	48
3.3.1	Konformationsanalyse der Grundstruktur . . . . .	52
3.3.2	Überlagerung und Optimierung der Einzelkonformationen . . .	54
3.3.3	CoMF-Analysen und Ergebnisse . . . . .	57
3.4	Auto-PLS zur Erstellung von CoMFA-Modellen für D <sub>1</sub> -Antagonisten .	68
3.4.1	Strukturen und biologische Daten . . . . .	68
3.4.2	Pharmakophorbasiertes Alignment als Ausgangspunkt für die Auto-PLS . . . . .	70
3.4.3	Initiale CoMFA-Modelle . . . . .	74
3.4.4	Optimierung der CoMFA-Modelle . . . . .	75

<b>4</b>	<b>COSMO</b>	<b>79</b>
4.1	Das COSMO-Solvatationsmodell . . . . .	80
4.1.1	Problem der Berücksichtigung der Solvation . . . . .	80
4.1.2	Solvatationsmodelle . . . . .	81
4.1.3	COSMO-Daten als Deskriptoren für die QSAR . . . . .	82
4.2	Berechnung der COSMO-Sigma-Profile . . . . .	84
4.3	Sigma-Profile als Moleküldeskriptoren . . . . .	87
4.3.1	Vergleichsverfahren . . . . .	87
4.3.2	Vergleich von Konformationen eines Moleküls . . . . .	88
4.3.3	Vergleich von verschiedenen Molekülen . . . . .	90
4.4	QSAR-Anwendung der COSMO-Deskriptoren . . . . .	91
4.4.1	Verwendete Strukturen und Konformationen . . . . .	92
4.4.2	PLS-Analyse der COSMO-Daten . . . . .	95
<b>5</b>	<b>Zusammenfassung und Ausblick</b>	<b>99</b>
<b>A</b>	<b>Anhang – Tabellen zu den D<sub>2</sub>- und D<sub>3</sub>-Rezeptorantagonisten</b>	<b>103</b>
<b>B</b>	<b>Anhang – NMR-Konformationsuntersuchungen</b>	<b>117</b>
<b>C</b>	<b>Anhang – Daten der Dopamin-D<sub>1</sub>-, D<sub>2</sub>-, D<sub>4</sub>-, D<sub>5</sub>-Antagonisten</b>	<b>127</b>
<b>D</b>	<b>Anhang Programme und Skripte</b>	<b>143</b>
D.1	Das Programm PLS-Toolbox . . . . .	143
D.2	Das SYBYL-Skript match_all . . . . .	149
D.3	Das Programm conf_elecT . . . . .	150
D.4	Das SYBYL-Skript auto_pls . . . . .	154
D.5	Das Programm plsreport . . . . .	157
D.6	Das SYBYL-Skript random_groups_pls . . . . .	159
D.7	Das Programm cosmo_anA . . . . .	161
	<b>Literatur</b>	<b>165</b>

# Abbildungsverzeichnis

1.1	Kreislauf der Leitstrukturoptimierung . . . . .	6
1.2	Verzahnung der QSAR mit anderen Molecular-Modelling-Methoden .	7
2.1	Vergleich: Calcium-Assay vs. Radioligand-Bindung . . . . .	12
2.2	Korrelation der pK <sub>i</sub> -Werte am Dopamin D <sub>1</sub> -Rezeptor: Ratte vs. Mensch	16
2.3	Sequenzvergleich des Dopamin D <sub>1</sub> -Rezeptors von Mensch und Ratte .	17
2.4	Vergleich der pK <sub>i</sub> -Werte für D <sub>1</sub> - und D <sub>5</sub> -Rezeptoren . . . . .	19
2.5	LE 403 und ähnliche Strukturen . . . . .	20
3.1	Torsionswinkeldiagramme in Abhängigkeit von der Konformation . .	25
3.2	Prozess des Simulated Annealing . . . . .	27
3.3	Diagrammausgabe von <code>conf_elecT</code> (Molekül: LE 404) . . . . .	31
3.4	Konformationen der Verbindung LE 404 nach dem Simulated Annealing.	33
3.5	3D-Strukturen selektiert durch <code>conf_elecT</code> (Molekül: LE 404) . . . .	33
3.6	Das CoMFA-Gitter . . . . .	34
3.7	Sterische und elektrostatische Potenzialfunktion . . . . .	36
3.8	Orientierung im Gitter . . . . .	40
3.9	Die Verteilung der q <sup>2</sup> -Werte in Abhängigkeit von der Gittergröße . . .	41
3.10	Blockschema der Forward-Optimierung . . . . .	45
3.11	Verteilung der q <sup>2</sup> -Werte aller SAMPLS-Analysen . . . . .	46
3.12	Verteilung der Vorhersagefehler durchgeführter SAMPLS-Analysen . .	47
3.13	Verteilung der q <sup>2</sup> -Werte nach 200 „Random-Groups“-PLS-Analysen .	48
3.14	Diversitätsverteilung und selektierte Repräsentanten der Benzazepine	53
3.15	Ausgewählte Konformationen des Grundgerüsts der Benzazepine . . .	53
3.16	Bei den Benzazepinen veränderte Torsionswinkel . . . . .	55
3.17	Überlagerung der Benzazepine vor und nach der Winkelanpassung . .	55

3.18	Vier verschiedene initiale Konformationsauswahlen . . . . .	59
3.19	Histogramme der $q^2$ -Werte der Optimierung für die $D_2/D_3$ -Selektivität	65
3.20	$Q^2$ -Wert-Histogramme der Optimierung für die $D_2$ - und $D_3$ -Bindung .	65
3.21	Verteilung der $q^2$ -Werte von 200 „Random-Groups“-PLS-Analysen für die $D_2$ - und $D_3$ -Modelle . . . . .	66
3.22	2D-Strukturen der ausgewählten Dopamin $D_1$ -Antagonisten . . . . .	69
3.23	Minimum-Energie-Konformationen der Stereoisomere von SCH 39166	71
3.24	Überlagerung von (-)-2b-SCH 39166 und (R)-(+)-SCH 23390 . . . . .	72
3.25	Die für die Überlagerung verwendeten Pharmakophormerkmale . . . . .	73
3.26	Überlagerung aller 44 anfänglichen Konformationen . . . . .	74
3.27	Bestes CoMFA-Modell des anfänglichen Alignments . . . . .	75
3.28	Alignment der finalen CoMFA- und CoMSIA-Modelle . . . . .	77
3.29	$Q^2$ -Wertverteilung nach Stabilitätstest der endgültigen Modelle . . . . .	78
4.1	COSMO-Sigma-Profile der Verbindungen Wasser, Chloroform, Aceton	83
4.2	COSMO-Sigma-Profile erstellt mit der Histogrammmethode . . . . .	85
4.3	Berechnung der Parzen-Window-Dichte aus einzelnen Gaußfunktionen	86
4.4	COSMO-Sigma-Profile der Konformationen von (R)-(+)-SCH 23390 .	89
4.5	COSMO-Sigma-Profile unterschiedlich ähnlicher Moleküle . . . . .	91
4.6a	COSMO-Sigma-Profile der Dopamin $D_1$ -Antagonisten . . . . .	93
4.6b	COSMO-Sigma-Profile der Dopamin $D_1$ -Antagonisten . . . . .	94
4.7	$Q^2$ -Werte mit zunehmendem Ausschluss von X-Variablen . . . . .	96
4.8	Vergleich der gemessenen mit der berechneten Aktivität . . . . .	98
4.9	Vom Modell benutzte Bereiche der Ladungsdichte . . . . .	98
B.1	Die für die NMR-Berechnungen berücksichtigten Konformationen von (R)-(+)-SCH 23390 . . . . .	118
B.2	500 MHz $^1\text{H}$ -NMR Spektrum von (R)-(+)-SCH 23390 bei 298K: Über- sicht über den Aromatenbereich . . . . .	122
B.3	500 MHz $^1\text{H}$ -NMR Spektrum von (R)-(+)-SCH 23390 bei 298K . . . .	123
B.4	500 MHz $^1\text{H}$ -NMR Spektren von (R)-(+)-SCH 23390 bei 298 u. 280 K	124
B.5	500 u. 200 MHz $^1\text{H}$ -NMR Spektrum von (R)-(+)-SCH 23390 bei 298 K und 280 K überlagert . . . . .	125
C.1	Diversitätsverteilung und ausgewählte Repräsentanten von AHA D11	129

C.2	Globale Energieminimumkonformation von AHA D11 . . . . .	129
C.3	Diversitätsverteilung und Repräsentanten von (R)-(+)-SCH 23390 . .	130
C.4	Globale Energieminimumkonformation von (R)-(+)-SCH 23390 . . . .	130
C.5	Diversitätsverteilung und ausgewählte Repräsentanten von LE 300 . .	131
C.6	Globale Energieminimumkonformation von LE 300 . . . . .	131
C.7	Diversitätsverteilung und ausgewählte Repräsentanten von LE 404 . .	132
C.8	Globale Energieminimumkonformation von LE 404 . . . . .	132
C.9	Diversitätsverteilung und ausgewählte Repräsentanten von LE 410 . .	133
C.10	Globale Energieminimumkonformation von LE 410 . . . . .	133
C.11	Diversitätsverteilung und ausgewählte Repräsentanten von LE 420 . .	134
C.12	Globale Energieminimumkonformation von LE 420 . . . . .	134
C.13	Diversitätsverteilung und ausgewählte Repräsentanten von LERU 301	135
C.14	Globale Energieminimumkonformation von LERU 301 . . . . .	135
C.15	Diversitätsverteilung und ausgewählte Repräsentanten von SH 3 . . .	136
C.16	Globale Energieminimumkonformation von SH 3 . . . . .	136
C.17	Diversitätsverteilung und ausgewählte Repräsentanten von LE 403 . .	137
C.18	Globale Energieminimumkonformation von LE 403 . . . . .	137
C.19	Diversitätsverteilung und ausgewählte Repräsentanten von LE 400 . .	138
C.20	Globale Energieminimumkonformation von LE 400 . . . . .	138
C.21	Diversitätsverteilung und Repräsentanten von (-)2a SCH 39166 . . . .	139
C.22	Gefundene repräsentative Konformationen von (-)2a SCH 39166 . . .	139
C.23	Diversitätsverteilung und Repräsentanten von (-)2b SCH 39166 . . . .	140
C.24	Die häufigsten Konformationen von (-)2b SCH 39166 . . . . .	140
C.25	Diversitätsverteilung und Repräsentanten von (+)2a SCH 39166 . . .	141
C.26	Gefundene repräsentative Konformationen von (+)2a SCH 39166 . . .	141
C.27	Diversitätsverteilung und Repräsentanten von (+)2b SCH 39166 . . .	142
C.28	Die häufigsten Konformationen von (+)2b SCH 39166 . . . . .	142
D.1	Programmfenster der PLS-Toolbox . . . . .	144



# 1. Einleitung

QSAR ist die englischsprachige Abkürzung für Quantitative Struktur-Wirkungsbeziehungen (**Q**uantitative **S**tructure-**A**ctivity **R**elationships). QSAR-Methoden werden mittlerweile in vielen Gebieten der Chemie und Life Sciences angewandt. Gerade bei den letztgenannten Disziplinen sind QSAR-Verfahren zum unverzichtbaren Werkzeug für die Entwicklung und Optimierung neuer Arzneistoffe geworden. Bei QSAR geht es nicht nur darum, Zusammenhänge zwischen der Struktur einer chemischen Verbindung und ihrer (biologischen) Wirkung zu finden, sondern auch und hauptsächlich darum, diese zu quantifizieren.

Um diese Beziehungen aufzustellen, müssen sowohl die chemische Struktur der Verbindungen als auch ihre Aktivität (biologische Wirkung) in Form von Zahlen ausgedrückt werden. Für die Struktur bedient man sich physikochemischer Deskriptoren, welche Informationen über sterische und elektronische Eigenschaften aber auch über Lipophilie und Topologie der Moleküle enthalten können. Diese Parameter sind entweder experimentell bestimmt oder mit Computerprogrammen berechnet worden, wobei die letztere Variante zunehmend Verbreitung findet. Die Aktivitätsparameter entstammen zumeist biologischen Assays und beinhalten bedingt durch die schlechte Reproduzierbarkeit mehr oder weniger große experimentelle Fehler, die nicht zuletzt die mögliche Genauigkeit der QSAR limitieren.

Die Anfänge der QSAR liegen in der Toxikologie und können bis ins 19. Jahrhundert zurück datiert werden. 1863 beschrieb A.F.A. Crois in seiner Doktorarbeit an der Universität Straßburg, dass Alkohole mit abnehmender Wasserlöslichkeit um so toxischer auf Säugetiere wirken [1]. Bereits 1868 postulierten T. Fraser und A. Crum Brown anhand der biologischen Effekte von am basischen Stickstoffatom methylierten und nicht methylierten Alkaloiden, dass die „physiologische Aktivität“  $\Phi$  eine Funktion der chemischen Konstitution  $C$  sein muss (Gleichung 1.1).

$$\Phi = f(C) \quad (1.1)$$

Sie konnten damals ihre Funktion noch nicht anwenden, da sie keinen Weg fanden, die chemische Struktur ihrer Verbindungen quantitativ zu beschreiben. Außerdem waren ohnehin die meisten Strukturen organischer Verbindungen zu dieser Zeit noch unbekannt. Aber auch heute gibt es noch keinen Weg, aus der Struktur einer einzelnen Verbindung quantitativ ihre Aktivität abzuleiten. Das gelingt nur bei Betrachtung der Unterschiede sowohl in der Struktur als auch in der Aktivität von mehreren Verbindungen (s. Gleichung 1.2).

$$\Delta\Phi = f(\Delta C) \quad (1.2)$$

Im letzten Jahrzehnt des 19. Jahrhunderts arbeitete Charles Ernest Overton an der Universität Zürich an seinen „Studien zur Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie“ [2], worin erstmals quantitative Struktur-Toxizitätsbeziehungen abgeleitet wurden. Overton [3] und unabhängig von ihm zur selben Zeit Hans Horst Meyer an der Universität Marburg korrelierten die narkotische Aktivität von Verbindungen sowohl mit ihrer Struktur (Kettenlänge) als auch mit Verteilungskoeffizienten in verschiedenen lipophilen Phasen [4]. Sie sind somit zu den Pionieren der QSAR zu zählen.



Bis 1950 gab es kaum Versuche, Struktur und Eigenschaften von chemischen Verbindungen mit der biologischen Aktivität zu korrelieren, obwohl diese Charakteristika immer besser untersucht werden konnten. Um die elektronischen Substituenteneffekte zu beschreiben, war mittlerweile die Gleichung von Louis Hammett verfügbar [5]. Er führte die verschiedenen Dissoziationskonstanten unterschiedlich substituierter aromatischer Carbonsäuren auf die elektronischen Einflüsse der Substituenten (Parameter  $\sigma$ ) zurück. Mit der elektronischen Hammett-Gleichung allein ließen sich viele biologische Daten nur schlecht korrelieren. Aus diesem Grund entwickelte Corwin Hansch in den sechziger Jahren des 20. Jahrhunderts den Lipophilieparameter  $\pi$ . Aus der Differenz der LogP-Werte der Verbindungen mit und ohne Substituenten konnte er den lipophilen Einfluss des einzelnen Substituenten ablesen.

Damit schien nun der Durchbruch gelungen zu sein und Hansch und Fujita publizierten 1964 die bis heute wichtigste Methode der klassischen QSAR [6]. Im gleichen Jahr beschrieben Spencer M. Free und James W. Wilson ein noch einfacheres aber dennoch effektives Verfahren, welches ebenso bis heute angewandt wird [7].

Erst mit der Verfügbarkeit schneller Computer wurde eine neue QSAR-Methode möglich, welche die 3D-Struktur der Moleküle berücksichtigte. Die vergleichende molekulare Feldanalyse (CoMFA) korreliert die Unterschiede von verschiedenen vom Molekül ausgehenden Eigenschaftsfeldern mit der biologischen Aktivität (s. Kapitel 3.2, Seite 34). Eine ähnliche Methode stellt das CoMSIA-Verfahren dar. Mit dem CoMFA/CoMSIA-Verfahren war man nicht mehr ausschließlich auf ein gemeinsames Grundgerüst der Moleküle beschränkt. Außerdem näherte man sich auf räumlicher Ebene dem Schlüssel-Schloss-Prinzip der Rezeptorbindung an, da man anhand der Eigenschaftsfelder auf komplementäre Eigenschaften des Rezeptors zu schließen versuchte.

Allerdings traten auch neue Probleme auf, wie z. B. das Finden der korrekten Überlagerung und der richtigen Konformationen. Das gilt auch für die Pseudorezeptormodelle - ein Versuch mittels QSAR-Methoden ausgehend von den Liganden mehr über den komplementären Rezeptor zu erfahren. Schon bei der CoMFA-Methode konnten

über die Betrachtung der Konturdiagramme Aussagen über räumliche Eigenschaften des Rezeptors abgeleitet werden. Mit Hilfe von Pseudorezeptorgeneratoren [8, 9] wurden die QSAR-Informationen mit Fokus auf den (bis dahin) unbekannten Rezeptor betrachtet. Neueste Entwicklungen unterscheiden sich nicht mehr grundsätzlich von den genannten Verfahren, auch wenn sie laut Bezeichnungen wie „4D-QSAR“ [10] oder „5D-QSAR“ [11] in neue Dimensionen vorzustößen scheinen.

Mit jedem Fortschritt, der bei der Entwicklung von QSAR-Methoden gemacht wurde, traten auch neue Probleme zu Tage. In der Folge von Hanschs Entwicklung wurden in Ergänzung der klassischen Hansch-Parameter eine Vielzahl weiterer Molekül- und Substituenten-Deskriptoren entwickelt. Dies führte zu mathematisch statistischen Problemen, welche besonders die klassische lineare Regression betrafen. Die Anzahl der eingesetzten Deskriptorvariablen ist durch die Anzahl der vermessenen Verbindungen begrenzt. Außerdem werden bei der multiplen linearen Regression (MLR) alle Variablen als linear unabhängig voneinander betrachtet, was sie aber oft nicht sind.

Mit der Anzahl der verwendeten Variablen steigt die Chance, einen Zusammenhang zwischen Deskriptoren und Aktivität zu finden. Es steigt aber ebenso die Gefahr, dass dieser Zusammenhang nicht der ursächliche ist. Dieses Phänomen nennt man Zufallskorrelation. Weiterhin kommt es zu einer Anpassung des Modells an die Messfehler, welche bei Daten aus biologischen Systemen sehr groß sein können. Dieses „Overfitting“ führt zu einer Verschlechterung bei der Anwendung des QSAR-Modells zur Vorhersage der Aktivität neuer Verbindungen.

Natürlich wurden Strategien entwickelt, diese Beeinträchtigungen einzudämmen bzw. zu umgehen. Mit Hilfe der Variablenselektion wird die Zahl der Deskriptoren verringert, indem wenig deskriptive Variablen entfernt werden. Da eine systematische Vorgehensweise dabei nicht in angemessener Zeit durchführbar ist, werden dabei allerlei interessante Suchstrategien, wie z. B. genetische Algorithmen [9] oder Tabu-Search [12, 13], angewandt. Ohne den Einsatz geeigneter Validierungsverfahren sind die resultierenden Modelle jedoch meistens wenig prediktiv [14, 15].

Das Problem der Kolinearität und Parametervielzahl wurde durch Einsatz der neuen Regressionsmethoden PCR (principle components regression - Hauptkomponentenregression) und PLS (partial least squares) [16] bewältigt, mit denen man auch hunderte von interkorrelierten Variablen, wie sie bei der CoMFA-Methode vorkommen, mit der Aktivität weniger Verbindungen in Zusammenhang bringen konnte.

Ein weiterer Ansatz war der Einsatz künstlicher neuronaler Netze, welche die Fähigkeit besitzen, nahezu jeden gewünschten Zusammenhang zwischen Struktur und Aktivität (auch nichtlineare Funktionen) bei entsprechendem Training zu adaptieren. Diese Fähigkeit macht sie jedoch besonders anfällig für das Overfitting. Erst durch die Kombination mit Validierungsmethoden wie der Kreuzvalidierung oder dem Bootstrapping kam man zu Modellen, die sowohl den gegebenen Zusammenhang so gut wie möglich abbildeten, als auch bei der Vorhersage nicht versagten. Mit Hilfe der Kreuzvalidierung ist es möglich, QSAR-Modelle hinsichtlich ihrer Prädiktivität zu beurteilen. Dabei wird ein kreuzvalidierter Regressionskoeffizient  $q^2$  (s. Gleichung 3.7 auf Seite 42) angegeben, welcher allerdings wenig über die externe Vorhersagekraft aussagt. Letztere kann nur über die Verwendung von externen Testdatensätzen, welche nicht in die Modellbildung einbezogen werden, bestimmt werden.

QSAR hat heute einen festen Platz im Kreislauf der Leitstrukturoptimierung eingenommen. Durch die Einbeziehung von QSAR-Modellen soll die Leitstrukturoptimierung zielgerichteter ablaufen und schneller bessere (höhere Aktivität, geringere Toxizität usw.) Derivate gefunden werden (s. Abbildung 1.1). Das steht wiederum in engem Zusammenhang mit dem eigentlichen Verständnis der Struktur-Wirkungsbeziehungen. Häufig wird durch QSAR-Untersuchungen eine Pharmakophorhypothese aufgestellt bzw. eine solche bestätigt oder verworfen. Auf der anderen Seite sind Pharmakophorhypothesen gute Startpunkte für den Entwurf eines QSAR-Modells.

Ebenso verhält es sich mit der Erkennung unterschiedlicher Bindungsmodi von Liganden. Hier kann ligandenbasiertes Design ebenfalls erfolgreich sein, wie sich an

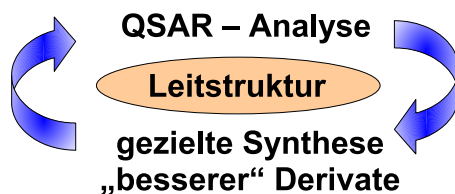


Abbildung 1.1: Kreislauf der Leitstrukturoptimierung

einem sehr eindrucksvollen Beispiel zeigen lässt. Der verschiedenartige Bindungsmodus von Methothrexat und Dihydrofolat in der Dihydrofolatreduktase wurde durch Betrachtung der Eigenschaftsfelder einige Zeit vor der röntgenkristallographischen Aufklärung postuliert [17].

Natürlich ist das rein ligandenbasierte Design immer durch die Unkenntnis des wahren Aufbaus des Rezeptors limitiert. Sind diese Informationen oder Teile davon jedoch vorhanden, können sie auch zur Verbesserung von QSAR-Modellen genutzt werden. Das läuft auf eine Annäherung an rezeptorbasierte Methoden hinaus, welche bislang als Alternativen zur QSAR angesehen werden. Dazu gehören Verfahren wie das Docking und das darauf aufbauende Virtual Screening. Dabei wird versucht, die Aktivität von Verbindungen aufgrund der Interaktionen mit der Rezeptortasche vorherzusagen. In einer Bewertungsfunktion werden verschiedene Interaktionsbeiträge aufsummiert und so eine der Bindungsenergie proportionale Kennzahl berechnet. Die freie Enthalpie der Bindung korreliert nach Gleichung 3.3 (Seite 34) mit dem Logarithmus der Gleichgewichtskonstante der Ligand-Rezeptorbindung. Die Gleichgewichtskonstante selbst oder eine proportionale Größe (z. B.  $IC_{50}$ ) ist in den meisten Fällen die der Messung zugängliche Bioaktivität.

Das Docking ist also strenggenommen die Aktivitätsvorhersage mit rezeptorbasierten QSAR-Modellen. Leider ist es bisher noch niemandem gelungen, die enthalpischen und entropischen Vorgänge im Rezeptor bei der Bindung eines Liganden quantitativ korrekt zu beschreiben. Vergrößert wird dieses Problem, wenn keine originale 3D-Struktur der Rezeptortasche vorhanden ist, sondern diese durch Homologie-Modellierung in Anlehnung an einen strukturell verwandten Rezeptor erstellt wurde.

Das trifft auf alle G-Protein gekoppelten Rezeptoren zu. Somit ist die Aktivitätsvorhersage ohne Einbeziehung von gemessenen Daten bislang auch nicht besonders erfolgreich im Vergleich zur ligandenbasierten QSAR.

Methoden der klassischen und vor allem der 3D-QSAR werden deshalb auch in naher Zukunft unverzichtbar bleiben und natürlich kontinuierlich weiterentwickelt werden. Sicherlich wird die Konvergenz zu Methoden des rezeptorbasierten Designs zunehmen und beide Ansätze weiter gegenseitig Ideen inkorporieren (s. Abbildung 1.2).

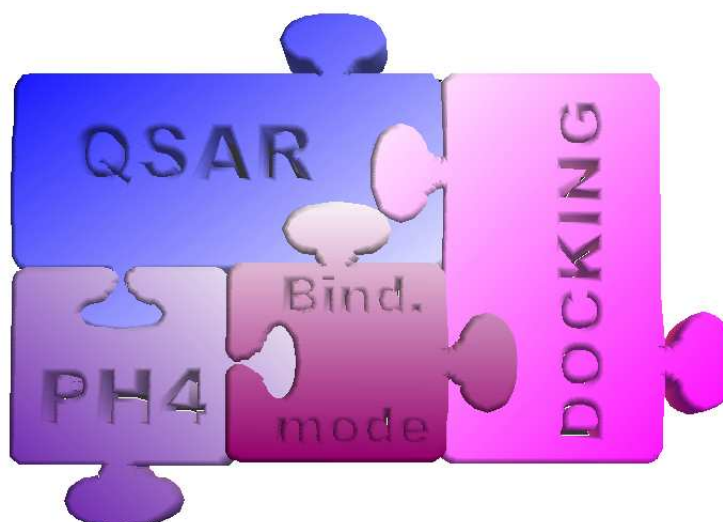


Abbildung 1.2: Verzahnung der QSAR mit verwandten bzw. alternativen Molecular-Modelling-Methoden; PH4 = Pharmakophor

## 1.1 Zielsetzung der Arbeit

Im Rahmen dieser Promotion sollten für eine Serie von Dopaminrezeptorantagonisten, welche im AK Prof. Lehmann synthetisiert und deren Aktivität an verschiedenen Dopaminrezeptorsubtypen (vorrangig D<sub>1</sub>) vermessen worden waren, prediktive QSAR-Modelle abgeleitet werden.

Die Dopaminrezeptoren gehören zu den G-Protein gekoppelten Rezeptoren, die an der Zelloberfläche exprimiert werden. Bisher ist es lediglich im Fall des Rinderrhodopsin gelungen, die Struktur mit kristallographischen Methoden aufzuklären [18]. Insofern sind rezeptorbasierte Verfahren wenig erfolgversprechend. Erschwerend kommt hinzu, dass die Dopaminrezeptorantagonisten eine strukturell sehr heterogene Substanzklasse mit nur wenigen gemeinsamen Merkmalen umfassen. Bisher veröffentlichte QSAR- und Pharmakophormodelle sind daher speziell auf homologe Verbindungsserien zugeschnitten und damit nicht allgemein anwendbar.

Die pharmakologische Forschung (außerhalb der Universität Bonn) konzentrierte sich hauptsächlich auf Verbindungen, die an D<sub>2</sub>-artige Rezeptorsubtypen (D<sub>2</sub>, D<sub>3</sub> und D<sub>4</sub>) binden, da diese zusammen mit den Serotoninrezeptoren als Hauptangriffspunkte bei der Behandlung der Schizophrenie betrachtet wurden. Die meisten der vorliegenden pharmakologischen Daten existieren für die Dopaminrezeptorantagonisten am D<sub>1</sub>-Rezeptorsubtyp.

Bei der Arbeit mit diesen Strukturen trat immer wieder das sogenannte Alignmentproblem (s. Abschnitt 3.2.2) in den Vordergrund, was die Ableitung von QSAR-Modellen erheblich erschwerte. Zur Lösung dieses Problems benötigt man effektive, neuartige Methoden, welche nicht nur im Falle der Dopaminantagonisten helfen können, quantitative Struktur-Wirkungsbeziehungen abzuleiten. Die während dieser Arbeit entwickelten Methoden wurden an Dopaminantagonisten des D<sub>1</sub>-, D<sub>2</sub>- und D<sub>3</sub>-Rezeptorsubtyps angewandt und evaluiert.

## 2. Datenauswahl

Manchmal ist im Gespräch von QSAR-Analytikern mit den Verantwortlichen für Synthese und Testung die Rede von für die QSAR-Analyse „untauglichen“ Daten. Das ist selbstverständlich übertrieben. Es sind höchstens nicht ausreichend viele Daten vorhanden. Anforderungen an „ideale“ chemische und biologische QSAR-Daten wären:

- ein möglichst gleiches Grundgerüst der Strukturen (um dem Alignmentproblem aus dem Weg zu gehen)
- möglichst viele Variationen an mehreren verschiedenen Molekülteilen und damit:
- so viele Verbindungen wie möglich
- Testung aller Verbindungen an der gleichen Zielstruktur mit dem gleichen Testsystem

Den idealen QSAR-Datensatz gibt es nicht. Erfolgreich aussagekräftige QSAR-Modelle abzuleiten stellt somit immer wieder eine Herausforderung dar. Dennoch ist es von Vorteil, wenn wenigstens einige der Anforderungen zum Teil erfüllt werden können, was durch ein entsprechendes Synthese- und Testdesign erreicht werden

kann. Zu diesem Zweck ist eine enge Kooperation zwischen den drei Zweigen der medizinischen Chemie – Synthese, Testung, Modelling/QSAR – unabdingbar. Hin und wieder gibt es einige „Universalgelehrte“, die eine solche Verzahnung durch Personalunion erreichen. Durch zunehmende Spezialisierung, wachsendes Wissen und zunehmende Möglichkeiten auf jedem einzelnen dieser Gebiete wird es für den medizinischen Chemiker aber immer schwieriger und vom Zeitaufwand fast unmöglich, persönlich auf all diesen Gebieten gleichzeitig zu arbeiten.

## 2.1 Verfügbarkeit von Aktivitäts- und Affinitätsdaten

Ein Schwerpunkt der Forschung am Institut für Pharmazeutische Chemie der Universität Bonn sind Antagonisten für Dopaminrezeptoren. Es wurden eine Reihe von Dopaminantagonisten synthetisiert [19–22].

Die Strukturen weisen zwar die vom Dopaminrezeptor bekannten pharmacophoren Gemeinsamkeiten auf (ein bis zwei aromatische Systeme und ein protonierbarer Stickstoff), sind jedoch strukturell sehr heterogen. Die Strukturen und ihre Konformationsanalysen sind im Anhang C ab Seite 127 beschrieben.

Die Verbindungen wurden sowohl mit Radioligandbindungsuntersuchungen von Michael Decker als auch mittels des Calcium-Assays von Barbara Hoefgen [23] an verschiedenen Rezeptorsubtypen des humanen Dopaminrezeptors charakterisiert. Die Daten sind ebenfalls im Anhang C tabelliert. Für die Daten von Michael Decker sind keine Standardabweichungen angegeben, da hier nur jeweils ein Versuch durchgeführt wurde (bei höchstaffinen Verbindung wurde der Mittelwert aus zwei Versuchen angegeben). Die meisten Daten liegen jeweils für den Dopamin D<sub>1</sub>-Rezeptor vor.

Weitere Daten zu diesen Verbindungen liegen für die Rezeptoren D<sub>1</sub>-D<sub>5</sub> von Alexandra Hamacher vor. Für diese Daten wurde allerdings ein anderes Zell- und Puffersystem verwendet, um die Vergleichbarkeit mit den Daten des Calcium-Assays zu



gewährleisten. Außerdem standen verschiedene der Literatur entnommene Daten zur Verfügung, darunter ein relativ großer und homogener Datensatz von Antagonisten des Dopamin D<sub>2</sub>- und D<sub>3</sub>-Rezeptors [24].

## 2.2 Vergleichbarkeit von Aktivitäts- und Affinitätsdaten

Da eine Vergrößerung des Datensatzes für QSAR-Zwecke grundsätzlich sinnvoll erscheint und auch kontinuierlich Inhibitionskonstanten für Dopaminantagonisten in einer Datenbank veröffentlicht werden [25], stellt sich die Frage nach der Vergleichbarkeit der von verschiedenen Arbeitsgruppen veröffentlichten Daten. Grundsätzlich sollten Daten, die an der gleichen Spezies gewonnen wurden, miteinander vergleichbar sein. Häufig werden jedoch an Stelle von K<sub>i</sub>-Werten IC<sub>50</sub>-Werte angegeben. Diese sind abhängig von der Konzentration und Art des Radioliganden und somit nur bei Einhaltung gleicher Versuchsbedingungen vergleichbar. In einigen Fällen sind diese IC<sub>50</sub>-Werte korrigiert worden, auf welche Weise, bleibt jedoch auch nach genauem Studium dieser Veröffentlichungen unklar. Hemmkonstanten in Form von K<sub>i</sub>-Werten werden aus IC<sub>50</sub>-Werten nach der Gleichung von Cheng und Prusoff [26] berechnet. Nach dieser Gleichung müssen zusätzlich die Konzentration des Radioliganden [L] und seine Gleichgewichtsdissoziationskonstante K<sub>D</sub> bekannt sein:

$$K_i = \frac{IC_{50}}{1 + \frac{[L]}{K_D}} \quad (2.1)$$

Vergleicht man Radioligandbindungsdaten mit funktionellen Daten des Calcium-Assays, so sind Abweichungen der Zahlenwerte aufgrund der verschiedenen Messsignale zu erwarten. Es sollte aber eine lineare Abhängigkeit der logarithmierten K<sub>i</sub>-Werte, zumal sie im selben Labor gemessen wurden, bestehen. Diese Korrelation ist aber nur für einige Verbindungen gegeben, wie die Abbildung 2.1 zeigt. Zwischen den beiden Assays bestehen fundamentale Unterschiede hinsichtlich der experimentellen Bedingungen, die zu diesen Abweichungen führen können.

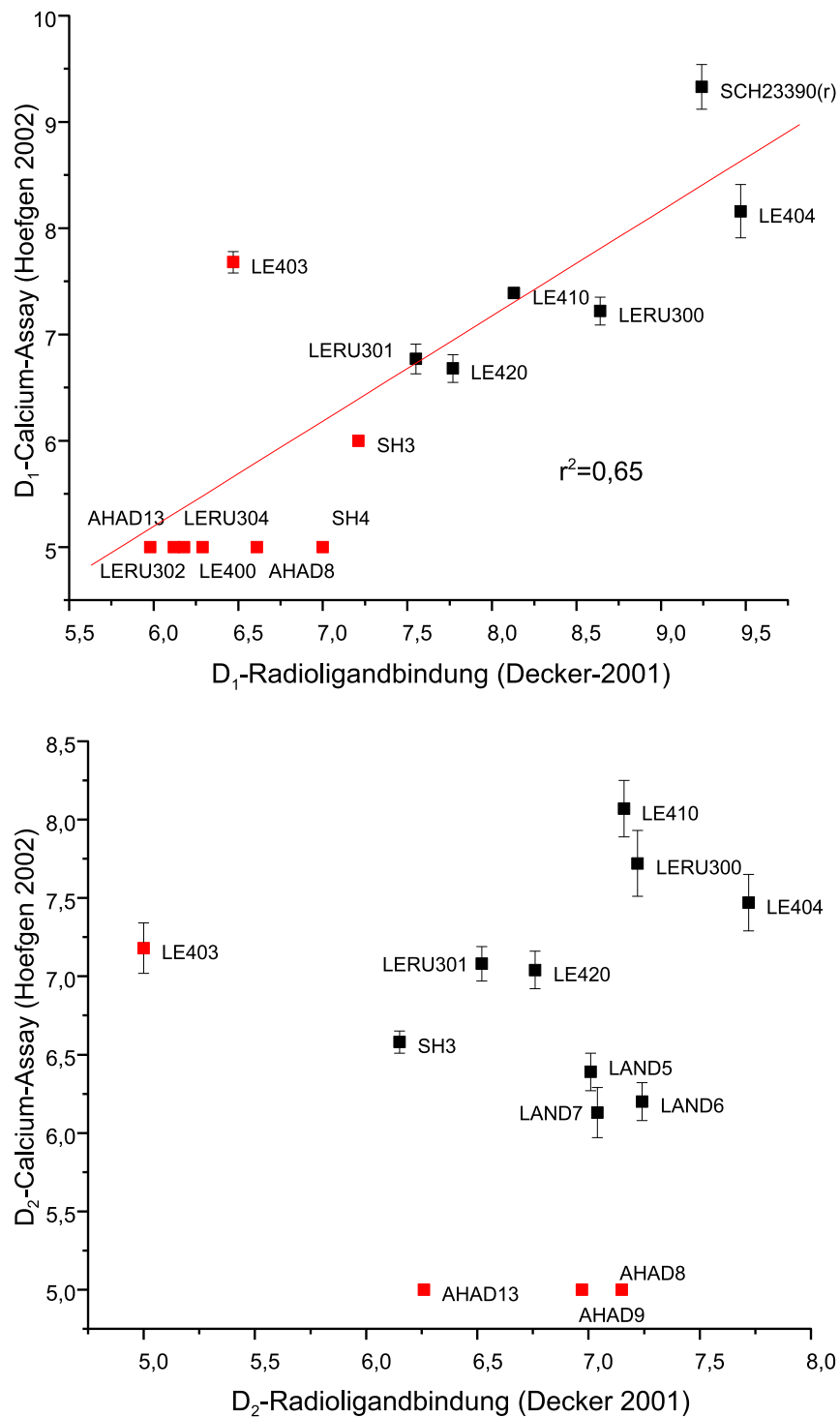


Abbildung 2.1: Vergleich der (apparenten)  $pK_i$ -Werte ermittelt mit Hilfe des Calcium-Assays und mit Radioligand-Bindung für D<sub>1</sub> (oben) und D<sub>2</sub> (unten). Bei rot markierten Werten konnte der genaue  $K_i$ -Wert entweder im Calcium-Assay oder mittels Radioligandbindung nicht bestimmt werden.

Beim Radioligandbindungsassay wird die die Rezeptoren enthaltende Membranpräparation mit dem Radioliganden und der Testsubstanz mehrere Stunden inkubiert, um die Ausbildung eines Gleichgewichts zu ermöglichen. Dabei findet eine Kompetition des Radioliganden mit der Testverbindung um die gleiche Bindungsstelle statt. Anschließend wird der Ligand-Rezeptor-Komplex durch Filtration abgetrennt. Das Messsignal ist die durch Szintillatoren in Licht umgewandelte radioaktive Strahlung des gebundenen Radioliganden. Es lässt (unter Berücksichtigung der Gesamt- und der unspezifischen Bindung) einen direkten Rückschluss auf die Affinität der Testsubstanz zu.

Beim Calcium-Assay ist das betrachtete Signal die Fluoreszenz des Calcium komplexierenden Farbstoffs Oregon Green 488 BAPTA-1. Dieser fluoresziert bei Bindung von  $\text{Ca}^{2+}$ -Ionen, welche aus dem endoplasmatischen Retikulum ausgeschüttet werden. Dem geht die Bindung eines Agonisten am G-Protein gekoppelten Rezeptor voraus, welcher den Antagonisten teilweise verdrängt. Es schließt sich nach Interaktion mit dem G-Protein eine Signaltransduktionskaskade an, die je nach Rezeptorsubtyp ( $G_s$ ,  $G_i$  usw.) sehr unterschiedlich ablaufen kann. Einen Überblick über die Komplexität der Calcium-Signaltransduktion geben Berridge et al. in [27]. Die Hemmkonstante  $K_i$  eines Antagonisten wird durch eine modifizierte Cheng-Prusoff-Gleichung (Gleichung 2.2) ermittelt, bei der statt des  $K_D$ -Werts des Radioliganden der  $EC_{50}$ -Wert des eingesetzten Agonisten verwendet wird:

$$K_i = \frac{IC_{50}}{1 + \frac{[L]}{EC_{50}}} \quad (2.2)$$

### 2.2.1 Vergleichbarkeit zwischen den Spezies Ratte u. Mensch

In der Literatur finden sich eine Vielzahl von Radioligandbindungsdaten von interessanten Verbindungen, welche ausschließlich an Dopaminrezeptoren der Ratte (Sprague Dawley, *Rattus norvegicus* L.) gemessen wurden. Obwohl schon 1991 hu-

mane Dopaminrezeptoren von Sunahara et al. kloniert wurden [28], benutzten die Forscher weltweit lange Zeit weiterhin Dopaminrezeptoren aus dem Rattenstriatum. Um die Verwendbarkeit von Bindungsdaten, welche an Dopaminrezeptoren der Ratte gemessen wurden, zu überprüfen, wurde exemplarisch die Literatur nach Hemmkonstanten für den Dopamin D<sub>1</sub>-Rezeptor durchsucht, welche sowohl an Rattenstriata, als auch an humanen klonierten Dopamin D<sub>1</sub>-Rezeptoren gemessen wurden [28–53]. In Tabelle 2.1 sind diese Werte zusammengestellt. Für die Ratte finden sich sowohl Werte, welche am Striatum als auch an klonierten Rezeptoren gemessen wurden. Da für letztere leider nur wenige Werte gefunden wurden, konnten diese nicht separat mit den Werten der humanen klonierten Rezeptoren verglichen werden. Wenn mehr als zwei Werte zur Verfügung standen ( $n > 2$ ), wurden jeweils Mittelwert (MW) und Standardabweichung (SD) angegeben. Bei lediglich zwei Werten wurde die Abweichung vom Mittelwert angegeben.

Anschließend wurde eine Regression der pK<sub>i</sub>-Werte durchgeführt, die in Abbildung 2.2 zu sehen ist. Die blau markierten Verbindungen Sulpirid, Norepinephrin (Noradrenalin) und Methysergid wurden aufgrund der großen Abweichungen nicht in die Regression einbezogen. Die Korrelation ist nur dann ausreichend, wenn diese Ausreißer nicht berücksichtigt werden. Deshalb sollten Verbindungen, für die ausschließlich Daten vom Rattenstriatum zugänglich sind, für die Erstellung eines QSAR-Modells für humane Rezeptoren nicht benutzt werden. Eine Verwendung innerhalb eines Testdatensatzes zur externen Vorhersage ist jedoch sehr wohl denkbar.

Die Abweichungen sind wahrscheinlich zum Teil auf die geringe Sequenzhomologie der extrazellulären Loops der Dopaminrezeptoren zurückzuführen. Die transmembranären Domänen weisen zwischen den Spezies Ratte und Mensch dagegen eine recht hohe Homologie auf. Dies führt zur Ausbildung einer ähnlichen Rezeptortasche, so dass die Erkennung des relativ kleinen (und unselektiven) Dopaminmoleküls gewährleistet ist. Für die Bindung anderer selektiver Agonisten und Antagonisten spielen die extrazellulären Loops eine größere Rolle. Zur Veranschaulichung ist in Abbildung 2.3 ein Vergleich der Sequenzen der Dopamin D<sub>1</sub>-Rezeptoren von Mensch

und Ratte dargestellt. Verschiedene Aminosäuren sind dabei blau und ähnliche pink gekennzeichnet. Die „Ähnlichkeit“ (besser: Austauschbarkeit) wurde vom Programm CLUSTALW [54] mit Standardeinstellungen anhand statistischer Substitutionsmuster<sup>1</sup> (in diesem Fall Gonnet250) berechnet.

Spezies	Ratte				Mensch	
Rezeptor	striatum		striatum + klon.		kloniert	
Verbindung	MW	SD/ABW	MW	SD/ABW	MW	SD/ABW
Clozapin	534	307(4)	469	275(5)	175	29,94(3)
Apomorphin	432	(1)	432	(1)	484	170(3)
Butaclamol(+)	14,6	(1)	14,6	(1)	3,6	0,6(2)
Chlorpromazin	74	(1)	74	(1)	76	34,12(3)
Dopamin	13876	12426(2)	13876	12426(2)	3405	1065(2)
Flupentixol	4,3	(1)	4,3	(1)	3,5	0,5(2)
Fluphenazin	11,2	(1)	11,2	(1)	17,33	9,07(3)
Haloperidol	290,18	323(5)	405	303(6)	62,5	26,34(4)
Iloperidon	546	(1)	546	(1)	29	87(2)
Methysergid	217	(1)	217	(1)	10000	(1)
Norepinephrin	12000	(1)	12000	(1)	10000	(1)
Olanzapin	250	(1)	175	75(2)	46,5	11,5(2)
Quetiapin	4240	(1)	3425	815(2)	995	283(2)
Risperidon	620	(1)	585	35(2)	395	128(2)
S33084	1000	(1)	1000	(1)	501,2	(1)
SCH23388(s)	192	(1)	192	(1)	41	(1)
SCH23390(r)	0,25	0,05(2)	0,31	0,09(3)	0,58	0,23(2)
Spiperon	8400	(1)	8400	(1)	399	178(2)
Sulpirid	30000	(1)	30000	(1)	36000	(1)
Thioridazin	59	(1)	59	(1)	94,5	5,5(2)
Zotepin	84	(1)	84	(1)	71	(1)

Tabelle 2.1: Vergleich von  $K_i$ -Werten an Dopamin  $D_1$ -Rezeptoren der Ratte (striatal und kloniert) und des Menschen (kloniert), Anzahl der Werte in Klammern, Werte extrahiert aus [28–53], MW: Mittelwert, SD: Standardabweichung, ABW: Abweichung vom Mittelwert (bei zwei Werten)

<sup>1</sup>Beim Vergleich von (sehr vielen) Proteinsequenzen wurden die häufig gegeneinander ausgetauschten Aminosäuren in Substitutionsmustern zusammenfasst, wobei auch eine Gewichtung nach Abstand im phylogenetischen Baum erfolgt.

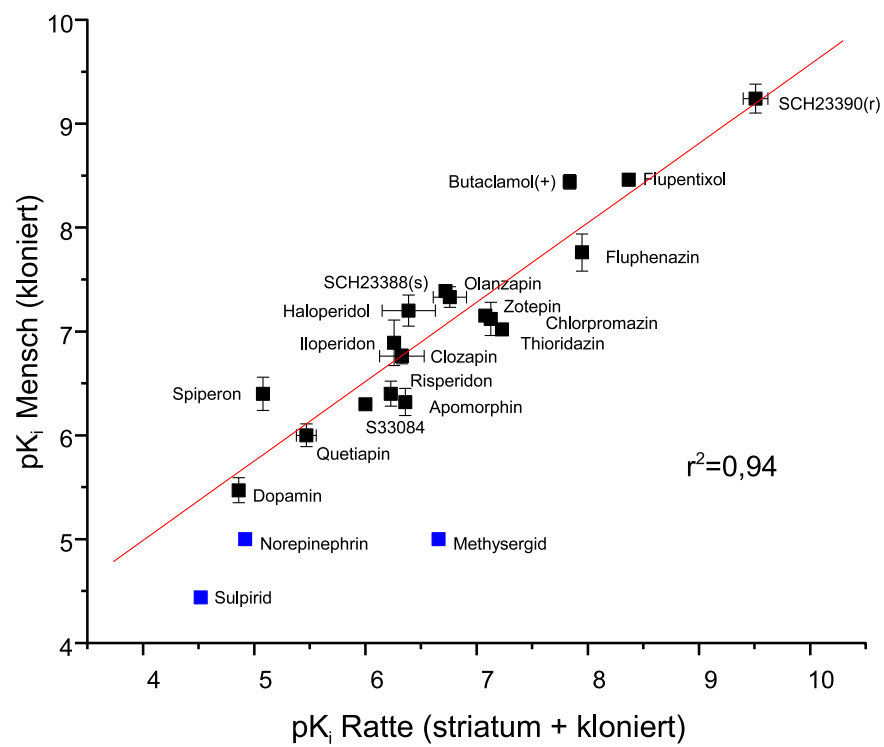


Abbildung 2.2: Korrelation der  $pK_i$ -Werte am Dopamin  $D_1$ -Rezeptor: Ratte vs. Mensch

Abbildung 2.3: Sequenzvergleich des Dopamin D<sub>1</sub>-Rezeptors von Mensch und Ratte, ■: verschieden, ■: ähnlich, □: identisch

Abbildung 2.3: Sequenzvergleich des Dopamin D<sub>1</sub>-Rezeptors von Mensch und Ratte, ■: verschieden, ■: ähnlich, □: identisch

### 2.2.2 Pufferabhängigkeit der Bindungskonstanten

Beim Vergleich von Bindungskonstanten tritt ein weiteres Problem auf, welches eher geringe Aufmerksamkeit genießt, aber für die Auswahl der Daten entscheidende Bedeutung erlangt. Die von M. Decker durchgeführten Messungen wurden mit CHO-Zellen (Chinesische Hamster-Ovarialkarzinom-Zellen) in einem Tris/MgCl<sub>2</sub>-Puffer durchgeführt. Die Daten des Calcium-Assays wurden von Frau Hoefgen mit HEK293 Zellen in KHP (Krebs-Hepes-Puffer) erhoben. Die Radioligandbindungsuntersuchungen wurden von Frau Hamacher fortgesetzt. Es kamen hierbei die bereits für den Calcium-Assay etablierten Zellen und das gleiche KHP-Puffersystem zum Einsatz, um diese Werte besser vergleichen zu können. Die HEK-Zellen wurden außerdem von den Experimentatoren vorgezogen, da es sich um menschliche Zellen handelt.

Im Vergleich der Radioligandbindungsdaten aus beiden Assays zeigte sich dabei eine systematische Differenz von ca. einer Log-Einheit des pK<sub>i</sub>-Wertes. Bisher konnten nur die Daten für D<sub>1</sub>- und D<sub>5</sub>-Rezeptoren verglichen werden, da für D<sub>2L</sub>-, D<sub>3</sub>- und D<sub>4</sub>-Rezeptoren nicht genügend Werte zur Verfügung standen. Die Diagramme in Abbildung 2.4 zeigen die Korrelation der Werte für die verschiedenen Zell- und Puffersysteme. Das Phänomen der Pufferabhängigkeit wird auch von Strange bestätigt [55, 56]. Für D<sub>1</sub>-Rezeptoren ergibt sich ein quadrierter Regressionskoeffizient von  $r^2=0,96$ , für D<sub>5</sub>-Rezeptoren von  $r^2=0,89$ .

Die Verbindung LE 403 stellt einen Ausreisser dar. Ihr pK<sub>i</sub>-Wert wurde von M. Decker zwei Log-Einheiten niedriger bestimmt als von Frau Hoefgen und Frau Hamacher. Da dies bei drei Rezeptoren (D<sub>1</sub>, D<sub>2</sub> und D<sub>3</sub>) auftritt, liegt ein systematischer Fehler vor. Die von M. Decker in Voruntersuchungen ermittelten Rezeptorsättigungswerte<sup>2</sup> lassen für LE 403 einen Wert in der Größenordnung des von LE 404 erwarten. Betrachtet man qualitativ die strukturellen Unterschiede von LE 403, LE 404 und LE 410 (s. Abbildung 2.5), so ist unwahrscheinlich, dass die Anwesenheit einer zweiten Hydroxylgruppe zu einem Aktivitätsabfall von zwei Log-Einheiten führt.

---

<sup>2</sup>Die Abnahme der rezeptorgebundenen Radioaktivität durch eine 10 mikromolare Lösung am D<sub>1</sub>-Rezeptor beträgt für LE 403 99 %, für LE 404 99 % und für LE 410 97 %.



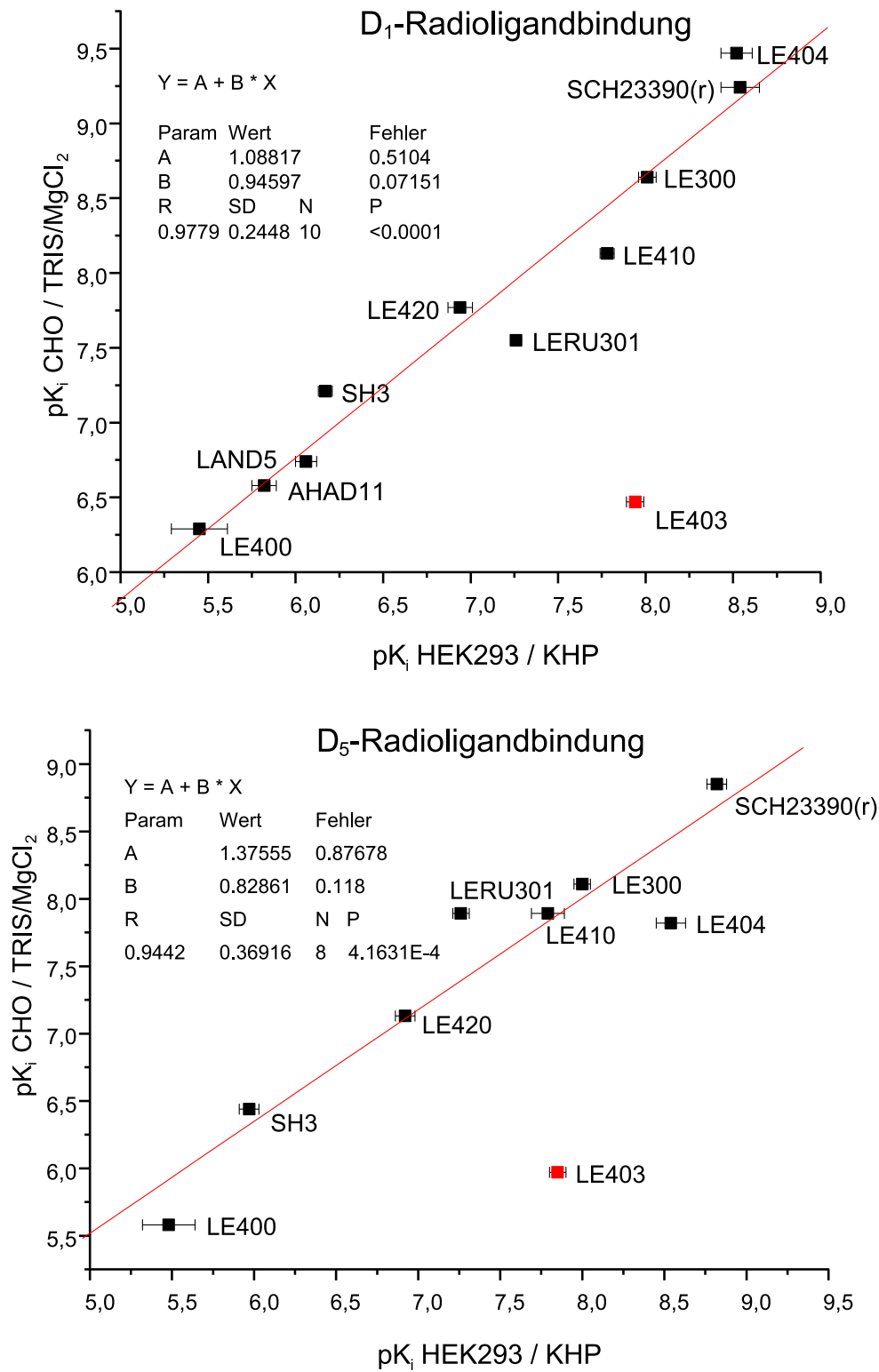


Abbildung 2.4: Vergleich der  $pK_i$ -Werte ermittelt mit Radioligand-Bindung für D<sub>1</sub>- (oben) und D<sub>5</sub>-Rezeptoren (unten). Die Verbindung LE 403 wurde nicht mit in die Regression aufgenommen.

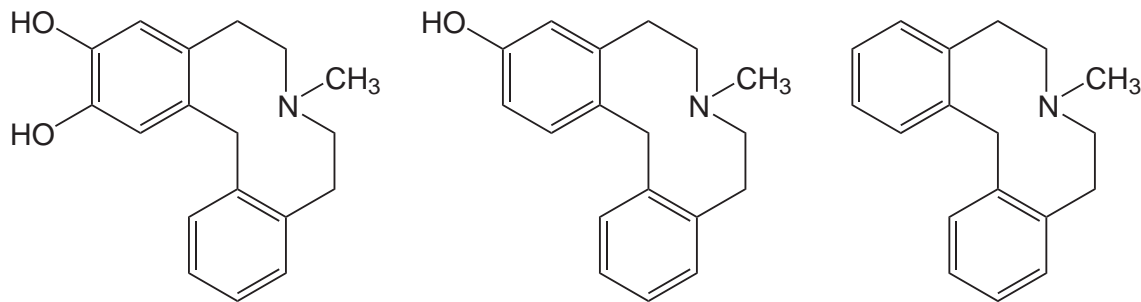


Abbildung 2.5: Vergleich von LE 403 (links,  $\text{pK}_{i,\text{Decker}}=6.47$ ) mit LE 404 (Mitte,  $\text{pK}_{i,\text{Decker}}=9.47$ ) und LE 410 (rechts,  $\text{pK}_{i,\text{Decker}}=8.13$ )

## 2.3 Begründung der Auswahl

Für die Entwicklung neuer QSAR-Methoden ist es vorteilhaft, zunächst mit möglichst „idealen“ Daten zu arbeiten. Häufig werden sogar künstlich erzeugte Datensätze verwendet. Ebenfalls weit verbreitet ist die Verwendung von Verbindungen und Aktivitätsdaten, für die bereits erfolgreich QSAR-Modelle erstellt worden sind.

Im Hinblick darauf erschienen die D<sub>1</sub>-Rezeptorantagonisten wenig geeignet. Die Strukturen waren sehr heterogen und die Daten inhomogen. Weiterhin waren nicht genügend Daten für einen der Rezeptorsubtypen vorhanden, um eine Aufteilung in Trainings- und Testdatensatz vorzunehmen.

Die in der Literatur [24] gefundenen Daten zu Antagonisten des Dopamin D<sub>2</sub>- und D<sub>3</sub>-Rezeptors konnten diese Lücke füllen. Ihre Strukturen waren bezüglich des Grundgerüsts sehr homogen (s. Tabelle 3.3 auf Seite 48). Auch die Rezeptorbindungsdaten entstammten alle derselben Quelle und wiesen keine Lücke auf. Bezüglich des Alignments waren somit deutlich weniger Probleme zu erwarten. Für die Etablierung der Auto-PLS-Methode wurden schließlich diese Daten verwendet (s. Kapitel 3.3 ab Seite 48).

Wie sich bei der Entwicklung der Auto-PLS-Methode herausstellte, sollte dieses Verfahren in der Lage sein, mit der strukturellen Heterogenität der Dopamin D<sub>1</sub>-Antagonisten besser zurechtzukommen, was anhand dieser Daten in Kapitel 3.4 auch gezeigt werden konnte.

Da für die Modellerstellung bei der QSAR die biologischen Daten möglichst geringe Fehler enthalten sollten, kam eine gemeinsame Verwendung von Daten des Calcium-Assays und Radioligandbindungsdaten bzw. von Daten, die an CHO- und HEK-Zellen gemessen wurden, nicht in Frage. Die Radioligandbindungsdaten, welche an den HEK-Zellen gemessen wurden, schienen schließlich am geeignetsten für eine QSAR, da sie die geringsten Fehler enthielten und ein guter Kontakt zu den Experimentatoren ständige Verbesserungen in der Genauigkeit ermöglichte.



### 3. Erweiterung der 3D-QSAR-Methode CoMFA mit automatischer PLS

Die CoMFA-Methode ist ein etabliertes Verfahren zur Ableitung von 3D-QSAR-Modellen. Verschiedene Konformationen und Alignments können dabei jedoch nicht ohne weiteres berücksichtigt werden. Außerdem ist die Optimierung von mäßig guten CoMFA-Modellen schwierig und aufwändig. Das Verfahren der automatischen PLS versucht diese Nachteile auszugleichen und ermöglicht die Erstellung besserer CoMFA-Modelle. Durch dieses neue Verfahren werden Alternativen bei Konformationen und Alignment für die Modellerstellung und Evaluierung berücksichtigt.

Eine andere Möglichkeit, das Alignmentproblem zu lösen, ist die Reduktion der 3D-Informationen zu raumunabhängigen Daten, was in Kapitel 4 anhand der COSMO-Sigma-Profiles gezeigt werden wird. Gemeinsam ist beiden Möglichkeiten die Notwendigkeit einer sorgfältigen Vorbereitung in Form einer Konformationsanalyse. Hierfür kann die (selbstentwickelte) Methode des Ähnlichkeitsclusterings zum Einsatz kommen.

Da für die Entwicklung eines neuen Verfahrens ein einfacher Datensatz vorteilhaft ist, wurde dafür die sehr homogene Gruppe der Antagonisten des Dopamin D<sub>2</sub>- und

D<sub>3</sub>-Rezeptors der Literatur entnommen (s. auch Kapitel 2.3, Strukturen in Tabelle 3.3). Dass diese Methode auch für heterogene Datensätze anwendbar ist, zeigt ihre Verwendung zur Ableitung von QSAR-Modellen für Dopamin D<sub>1</sub>-Rezeptorantagonisten.

## 3.1 Möglichkeiten zur Konformationsanalyse

Kovalente Einfachbindungen sind keinesfalls starr aber auch nicht generell uneingeschränkt frei drehbar. Die möglichen Torsionswinkel sind abhängig von den sterischen und elektrostatischen Eigenschaften der benachbarten Atome. Dadurch haben selbst einfache Moleküle keine streng definierte Gestalt, sondern können mehrere Konformationen einnehmen. Um die konformationelle Flexibilität eines Moleküls zu untersuchen, gibt es zwei prinzipiell unterschiedliche Verfahren.

- systematische Suche
- stochastische Suche

### 3.1.1 Systematische Suche

Bei der systematischen Suche wird mittels eines vorgegeben Rasters der gesamte Konformationsraum abgesucht. Die Wahrscheinlichkeit, dass alle lokalen Minima und somit auch das globale Minimum gefunden werden, steigt mit sinkender Rastergröße. Üblicherweise wird bei der systematischen Suche um alle drehbaren Bindungen rotiert, so dass sich das Raster als Intervall des Drehwinkels definieren lässt. Ab einem Winkelintervall  $\leq 10^\circ$  kann davon ausgegangen werden, dass alle Minima gefunden werden.

Die systematische Suche wird hauptsächlich bei Verbindungen ohne geschlossene Ringsysteme angewendet. Bei Ringschlüssen gibt es bei einigen Verfahren die Möglichkeit, den Ring zu öffnen und wie bei einem ringoffenen System zu verfahren. Der Ring wird dann, wenn die gefundene Konformation es zulässt, wieder geschlossen;

ansonsten wird die Konformation verworfen. Hierbei ist zu beachten, dass sich die Konfiguration von stereochemischen Zentren ändern könnte, wenn sich der Ringschluss von der vorherigen Ringöffnung unterscheidet.

Die maximale Anzahl der gefundenen Konformationen verhält sich umgekehrt proportional zum Winkelintervall  $\theta_i$ , wächst exponentiell mit der Anzahl drehbarer Bindungen  $N$  und ist gegeben durch die Formel 3.1.

$$\text{Maximale Zahl der Konformationen} = \prod_{i=1}^N \frac{360}{\theta_i} \quad (3.1)$$

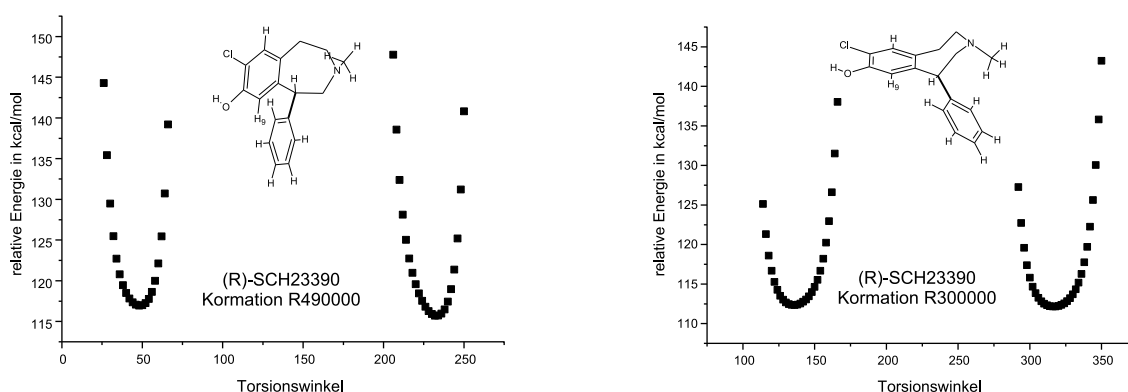


Abbildung 3.1: Torsionswinkeldiagramme abhängig von der Konformation des Siebenringes

Abbildung 3.1 zeigt das Torsionswinkel-Energie-Diagramm des Dopamin  $D_1$ - und  $D_5$ -Antagonisten (R)-SCH23390. Bei dieser Verbindung wird der Phenylring um die Bindung rotiert, mit der dieser mit dem Benzazepin verknüpft ist. Durch die gewählte Konformation des Siebenringes wird der Phenylring in seiner Drehung behindert. Dies findet Ausdruck in den extrem starken Anstiegen der Energiefunktion des Torsionswinkels. Ändert sich jedoch die Konformation des Grundgerüsts, so kann der Phenylring andere Torsionswinkel einnehmen. Wie quantenmechanische Berechnungen vorhersagten und NMR-Untersuchungen bestätigten (s. Anhang B (NMR) auf Seite 117), sind beide Konformationen wahrscheinlich und wandeln sich bei Raumtemperatur ständig ineinander um. Da das Molekül ansonsten aus geschlossenen

Ringen besteht, versagt hier die systematische Suche. Solchen dynamischen Phänomenen wird bei der Konformationsanalyse besser mit Moleküldynamiksimulationen wie dem Simulated Annealing Rechnung getragen.

### 3.1.2 Stochastische Suche

#### Simulated Annealing

Das Simulated Annealing ist im Prinzip eine Moleküldynamikmethode zur Untersuchung des Zustandes von „eingefrorenen“ molekularen Systemen. Das Konzept basiert auf der Art und Weise wie Flüssigkeiten während des Abkühlungsprozesses gefrieren und kristallisieren [57].

Die Schmelze ist anfänglich bei hoher Temperatur in ungeordnetem Zustand. Sie wird nun so langsam abgekühlt, dass sich das System jederzeit ungefähr im thermodynamischen Gleichgewicht befindet. Mit zunehmender Abkühlung nimmt die Ordnung im System bis zur Temperatur  $T = 0\text{ K}$  zu. Es ist also eine asymptotische Annäherung an den Zustand kleinster Energie, wobei praktisch bei genügend kleiner Annäherung an die Zieltemperatur abgebrochen wird. Geschieht die Abkühlung nicht langsam genug, können Zustände höherer Energie „eingefroren“ werden, und das System ist in einem lokalen Minimum gefangen.

Dieses Prinzip wird beim Molecular Modelling für die Konformationsanalyse und auch für Dockingsimulationen verwendet. Bei hoher Temperatur können energetische Barrieren zwischen lokalen Minimumkonformationen leichter überwunden werden. Die Abkühlung erlaubt eine langsame Relaxation, wodurch das Molekül in einen energetisch stabileren Zustand übergeht. Ihm wird dabei nicht nur kinetische Energie, sondern auch potentielle Energie entzogen.

Diesen Vorgang wiederholt man über mehrere Zyklen (s. Abbildung 3.2). Lässt man dem System genügend Zeit zum Equilibrieren während der anfänglichen Plateauphase, erhält man mehrere unterschiedliche Endzustände, wobei unter den dominierenden Konformationen häufig das globale Minimum zu finden ist.



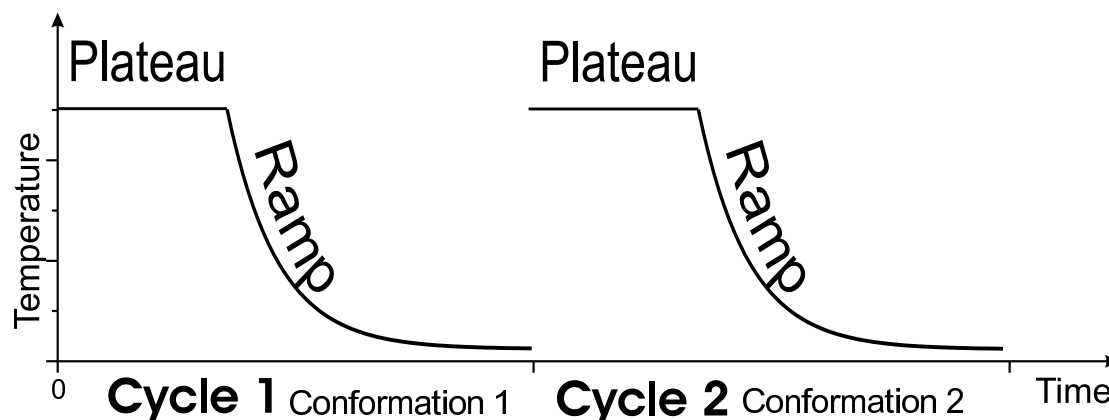


Abbildung 3.2: Prozess des Simulated Annealing

Ein Vorteil des Simulated Annealing gegenüber der systematischen Suche ist die generelle Anwendbarkeit, da es keine speziellen Strukturbeschränkungen gibt. Während bei der systematischen Suche der Aufwand und auch die Zahl der gefundenen Konformationen exponentiell mit der Zahl der Freiheitsgrade steigt, ist das beim Simulated Annealing nicht der Fall. Die Anzahl der Zyklen bleibt dem Anwender überlassen, wobei man allerdings auch hier die Zahl der Zyklen der Flexibilität des Moleküls anpassen muss.

### 3.1.3 Entwicklung eines neuen Verfahrens zum Ähnlichkeits-clustering

Viele der untersuchten Verbindungen besitzen flexible Ringsysteme, weshalb für die Konformationsanalyse in dieser Arbeit dem Simulated Annealing der Vorzug gegeben wurde. Um die durch das Simulated Annealing generierten Konformationen zu analysieren, wurde ein eigenes Verfahren entwickelt, da keine der bestehenden dem Autor bekannten Methoden in einfacher Weise anwendbar war. Bei diesem Verfahren werden die Konformationen miteinander verglichen (jede mit jeder). Mit den resultierenden Vergleichswerten sowie (optional) der Energie der Konformationen werden anhand eines vorgegebenen Schwellenwertes ähnliche Konformationen zu Gruppen (im Folgenden Cluster genannt) zusammengefasst.

Für das Simulated Annealing wurden die folgenden Parameter verwendet: Die Starttemperatur betrug 1000 K, die Plateauphase 2000 fs. Anschließend wurde das Molekül innerhalb 10000 fs mittels einer exponentiellen Annealingfunktion auf 0 K abgekühlt. Als Kraftfeld kam MMFF94 [58–62] zum Einsatz. Die Elektrostatik wurde durch ein entfernungsabhängiges Dielektrikum berücksichtigt, wodurch der elektrostatische Interaktionsbeitrag mit zunehmendem Abstand schneller (proportional zu  $1/r^2$  anstatt zu  $1/r$ ) abnimmt. Die Zyklenzahl wurde der konformationellen Flexibilität des Systems angepasst. In der größten Zahl der Fälle waren 100 oder 200 Zyklen ausreichend.

Die resultierenden Konformationen sollten nach dem Annealing energieminiert werden. Dies ist notwendig, um eventuelle energetisch sehr ungünstige Konformationen, welche z. B. durch zu schnelles Abkühlen erhalten werden können, auszuschließen. Außerdem kann die bei der Minimierung der Strukturen ausgegebene Energie später bei der Bildung der Konformationscluster berücksichtigt werden. Bei Verwendung eines Kraftfeldes empfiehlt sich der Einsatz einer modifizierten Version des für das Simulated Annealing benutzten Kraftfeldes MMFF94 (MMFF94S). MMFF94S ist für statische Minimierungen angepasst und benutzt einen speziellen Term für delokalisierte trigonale Stickstoffzentren (z. B. planare Amid-Bindungen) [63, 64]. Grundsätzlich ist der Einsatz jedes verfügbaren Kraftfeldes möglich, solange es für alle Konformationen dasselbe ist. Zusätzlich kann eine Optimierung mit der semiempirischen quantenmechanischen Methode AM1 durchgeführt werden, was auch mit den in dieser Arbeit untersuchten Strukturen erfolgt ist. Statt der relativen Kraftfeldenergiewerte wurden die AM1-Bildungsenthalpien (Heat of Formation – HoF) aufgezeichnet und später weiterverwendet.

Um die Konformationen sterisch zu vergleichen, wurde der MATCH-Algorithmus von SYBYL<sup>®</sup> verwendet. Dabei werden die relativen Abstände aller Atome verglichen und ein RMS-Wert angegeben. Der RMS-Wert ist, wie Formel 3.2 zeigt, die Quadratwurzel aus dem mittleren Abstandsquadrat  $d^2$  aller einbezogenen Atome.

$$RMS = \sqrt{\frac{\sum d^2}{n}} \quad (3.2)$$

Dabei wird nur das bestmögliche Ergebnis aller möglichen Überlagerungen der angegebenen Atome der beiden Konformationen berücksichtigt. Der RMS beschreibt also im Grunde, wie ähnlich sich zwei Konformationen sind. Da die Einbeziehung der Wasserstoffatome die Zahl der zu überprüfenden Überlagerungen stark ansteigen lässt, werden sie beim Vergleich nicht berücksichtigt. Somit wird nur das Molekülgerüst verglichen. Das für den Vergleich (MATCH) benötigte Skript ist im Anhang D.2 abgedruckt.

Enthält das Molekül eine Symmetrieebene (wie z. B. LE 404, s. Anhang C auf Seite 132), so sind spiegelbildliche Konformere möglich, die eine (fast) identische Energie aufweisen. Sie werden vom MATCH-Algorithmus durch einen großen RMS-Wert als sehr unterschiedlich charakterisiert. Wie in Tabelle 3.2 an der Energie (und zum Teil auch an der Häufigkeit) der Konformationspaare (49 + 93, 47 + 84, 99 + 56) ersichtlich ist, bilden solche Konformationen eigene Cluster. Normalerweise ist das gewollt, denn es ist durchaus denkbar, dass eine Konformation in die Rezeptortasche passt, ihr Spiegelbild jedoch nicht.

Da jede Konformation mit allen anderen verglichen wird, erhält man eine RMS-Matrix (s. Tabelle 3.1), die später von einem MATLAB-Programm ausgewertet werden kann. Ein Nachteil dieser Methode ist die mit  $n^2/2$  steigende Anzahl der notwendigen Vergleiche und die damit ebenso größer werdende RMS-Matrix. Andererseits funktionieren andere Ähnlichkeitsmaße vergleichbar und benötigen oft noch einen größeren Aufwand zur Berechnung. Weiterhin ist der RMS-Wert im Gegensatz zu manchen Fingerprintwerten keine abstrakte Größe, sondern direkt interpretierbar.

Das MATLAB Programm `conf_elecT` bildet simple Konformationscluster aus den RMS-Werten der MATCH-Analyse. Da das Programm auf MATLAB aufbaut, ist es sehr einfach erweiterbar. Der Programmcode von `conf_elecT` ist im Anhang D.3 abgedruckt.

Konf	1	2	3	4
1	0	RMS(1,2)	RMS(1,3)	RMS(1,4)
2		0	RMS(2,3)	RMS(2,4)
3			0	RMS(3,4)
4				0

Tabelle 3.1: Beispielhafte RMS-Matrix

Die von `conf_elecT` ausgewählten Konformationen decken als Repräsentanten den gesamten Konformationsraum ab. Der Mindestunterschied zwischen im Konformationsraum benachbarten Repräsentanten wird durch den RMS-Schwellenwert (Threshold-RMS) beim Programmaufruf bestimmt. Durch die Wahl dieses Wertes kann die Anzahl der selektierten repräsentativen Konformationen beeinflusst werden. Zusätzlich erhält man eine Häufigkeitsstatistik (s. Tabelle 3.2) über die Anzahl der zu ihren Repräsentanten ähnlichen Konformationen (deren RMS zum jeweiligen Repräsentanten kleiner ist als der Threshold-RMS). Diese wird vom Programm unter anderem als Balkendiagramm ausgegeben (s. Abbildung 3.3(b)).

LE 404				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
45	R490000	0.546	36	1.109
93	R920000	0.546	42	1.060
47	R500000	3.265	25	1.205
84	R840000	3.272	29	1.130
99	R980000	8.103	3	1.476
56	R590000	8.149	7	1.471

Tabelle 3.2: Beispiel für repräsentative Konformationen ermittelt durch `conf_elecT` (Molekül: LE 404 ; Schwelle = 1.0 Å)

Einen Überblick über die Häufigkeitsverteilung der durchschnittlichen RMS-Werte gibt ein weiteres Balkendiagramm (s. Abbildung 3.3(a)). Hierfür wurde der Mittelwert der RMS-Werte jeder Konformation (verglichen mit jeder anderen) gebildet. Es ist also der Spalten- oder Zeilenmittelwert der RMS-Matrix. Die Form des

Diagramms entspricht typischerweise einer logarithmischen Normalverteilung, denn um das globale Minimum gibt es üblicherweise sehr viele Konformationen. Dieses Histogramm kann dazu benutzt werden, einen sinnvollen Wert für den Threshold-RMS auszuwählen. Werte, die viel kleiner sind als der kleinste im Histogramm (mit Häufigkeit  $> 0$ ) dargestellte durchschnittliche RMS-Wert, sollten nicht für die Clusterbildung verwendet werden. Es würden sich zu viele Konformationscluster mit zu wenigen Konformationen bilden. Andererseits kann man den Threshold-RMS-Wert erhöhen, wenn die Repräsentanten noch zu ähnlich scheinen oder man die Zahl der Konformationscluster verringern will. Im Folgenden wird der grundlegende Algorithmus am Beispiel der Verbindung LE 404 etwas ausführlicher beschrieben.

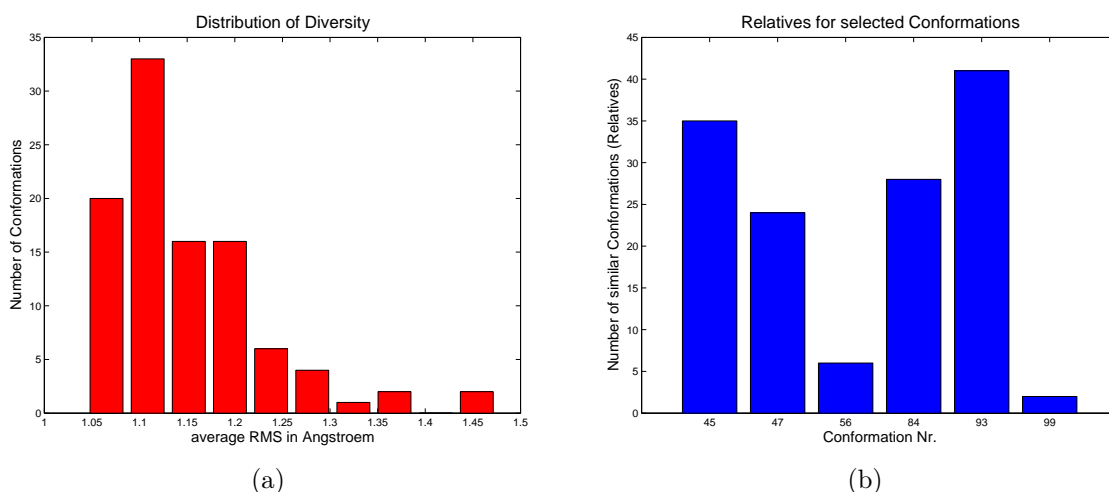


Abbildung 3.3: Diversitätsverteilung und ausgewählte Repräsentanten ermittelt durch `conf_elecT` (Molekül: LE 404; Schwelle = 1.0 Å)

Das Simulated Annealing mit anschließender AM1-Optimierung liefert einhundert Konformationen, welche den Konformationsraum gut abdecken. In der Abbildung 3.4 sind diese Konformationen auf den Phenolring von LE 404 überlagert dargestellt. Das Ziel ist es nun, aus dieser Konformationsvielfalt repräsentative Konformationen mit möglichst niedriger Energie zu extrahieren.

Im Programm `conf_elecT` wird aus der Liste der Vergleichswerte zunächst die symmetrische (an der Diagonalen gespiegelte) RMS-Matrix generiert. Anschließend wird für jede Konformation die Summe aller RMS-Werte berechnet und auf die Gesamt-

zahl der Konformationen bezogen, so dass man den durchschnittlichen RMS-Wert jeder Konformation erhält. Diese Werte sind in der letzten Spalte ( $\emptyset$  RMS) der Tabelle 3.2 zu sehen und werden im Histogramm „Distribution of Diversity“ dargestellt (s. Abbildung 3.3(a)).

Nun wird für jede Konformation die Liste der ähnlichen Konformationen erstellt („Relatives“), indem die Konformationen gesucht werden, deren RMS-Wert kleiner als der Schwellenwert ist. Schließlich wird die RMS-Matrix sukzessive verkleinert, indem in der Reihenfolge steigender Energiewerte (falls vorhanden) mit der globalen Energieminimumkonformation beginnend, die Zeilen und Spalten der zugehörigen ähnlichen Konformationen entfernt werden. So wird sichergestellt, dass die globale Energieminimumkonformation Repräsentant (bzw. RMS-Zentrum) des ersten Clusters wird und die Repräsentanten aller weiteren Cluster höhere Energiewerte aufweisen.

Am Ende bleiben nur die Repräsentanten ihrer jeweiligen Cluster übrig, welche in einem Balkendiagramm (s. Abbildung 3.3(b)) sowie auch in der Textkonsole mit der Anzahl ihrer ähnlichen Konformationen dargestellt werden. Hat eine solche ausgewählte Konformation keine „Verwandten“, so wird kein Balken dargestellt. Ein solcher Fall ist z. B. bei der Verbindung (+)2b SCH 39166 (s. Abbildung C.27(b) rechts im Anhang) zu sehen. Für die Beispielverbindung LE 404 sind die 3D-Strukturen der ausgewählten repräsentativen Konformationen in Abbildung 3.5 dargestellt.

Unter der Annahme, dass das Simulated Annealing die Realität (Molekül in Lösung) relativ gut abbildet, lässt sich aus den Häufigkeitswerten eine Aussage über die wahrscheinlich „realen“ Konformationen treffen. Damit sind die Konformationen gemeint, in denen das Molekül am häufigsten vorliegt und die auch energetisch am stabilsten sind. Zu diesen gehört nicht notwendigerweise die bioaktive Konformation. Bei den gegebenen Temperaturen (meist mindestens 298 K) sind konformationelle Änderungen ohne weiteres möglich. Wie weit sich die bioaktive Konformation von der Energieminimumkonformation unterscheiden kann, hängt von der Gibbsschen

Enthalpie der Bindungsreaktion ab. Diese lässt sich bisher leider nicht genau berechnen.

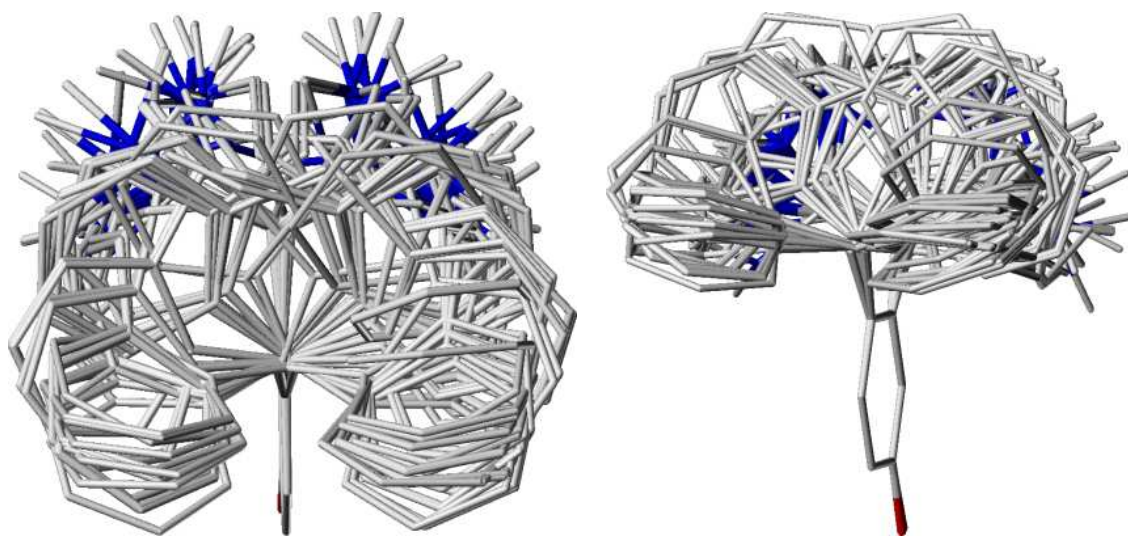


Abbildung 3.4: Konformationen der Verbindung LE 404 nach dem Simulated Annealing.

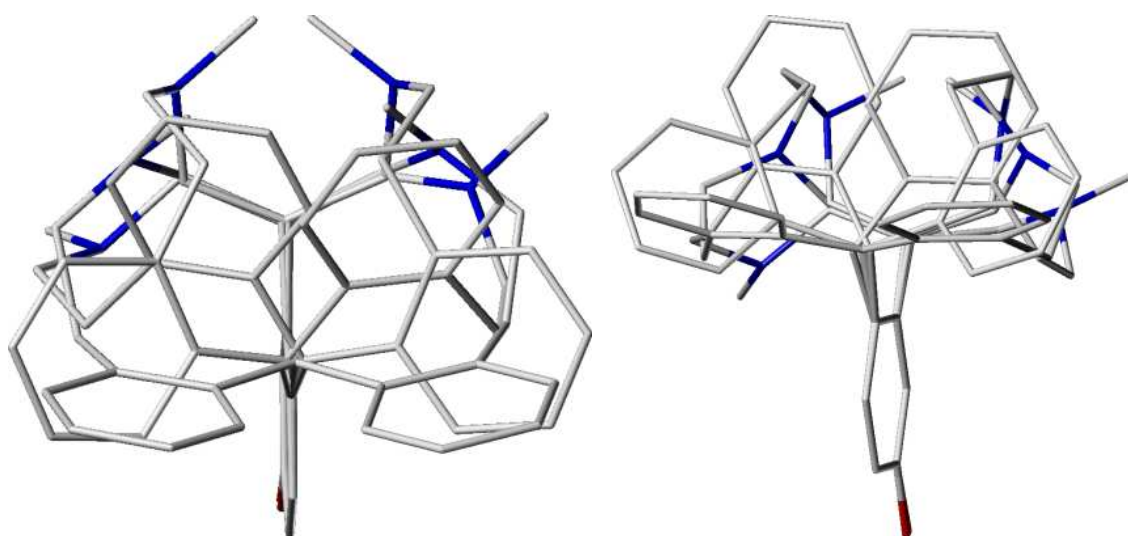


Abbildung 3.5: 3D-Strukturen der durch `conf_elecT` ausgewählten Konformationen der Verbindung LE 404.

## 3.2 CoMFA als Standardmethode der 3D-QSAR

Eine der heute am häufigsten eingesetzten Methoden, um Strukturmerkmale dreidimensional mit der biologischen Aktivität zu korrelieren, ist die vergleichende molekulare Feldanalyse CoMFA®. Die Grundlagen legten Cramer et al. Anfang der Achtziger Jahre, bevor sie 1988 [65, 66] die Methode veröffentlichten.

Die Moleküle werden in einer einheitlichen Orientierung und Platzierung in ein dreidimensionales Gitter (s. Abbildung 3.6) gelegt, wobei jeder der Gitterpunkte ein hypothetisches Rezeptoratom darstellt. Mit diesen werden nun mögliche sterische oder elektrostatische (oder sonstige Wechselwirkungen) berechnet, die in ihrer Gesamtheit als Feld bezeichnet werden. Bei der Korrelation der Felder mit biologischen Aktivitätsdaten geht man davon aus, dass die Gesetze der Gleichgewichtsthermodynamik (s. Gleichung 3.3 u. 3.4) gelten.

$$\Delta G = -RT \ln K \quad (3.3)$$

$$\Delta G = \Delta H - T\Delta S \quad (3.4)$$

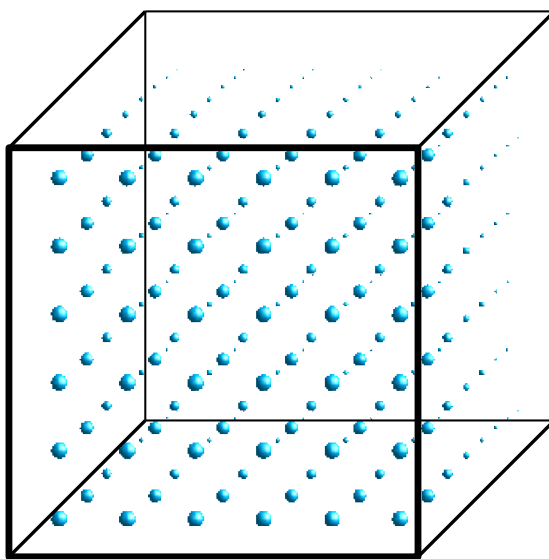


Abbildung 3.6: Das CoMFA-Gitter



Die Voraussetzung hierfür ist, dass die Rezeptorbindung die einzige bzw. die dominierende Basis der gemessenen Wirkung einer Verbindung ist. Diese Annahme wird häufig dann verletzt, wenn in vivo Daten verwendet werden. Aber auch Daten funktioneller Assays erfüllen nicht immer diese Voraussetzung (s. Kapitel 2.2 auf Seite 11).

Der enthalpische Anteil, der energetische Wechselwirkungen beschreibt, besteht vorwiegend aus sterischen und elektrostatischen Interaktionen. Die Entropie  $S$  spielt bei hydrophoben Wechselwirkungen eine wichtige Rolle. Ein Beispiel hierfür ist die Solvation aller hydrophoben Bereiche der beteiligten Strukturen, die in Abhängigkeit von ihrer Größe Einfluss auf die Entropie nehmen. Vergleicht man in der 3D-QSAR homologe bzw. nicht völlig verschiedene Verbindungen, ist der entropische Anteil vernachlässigbar, weil dieser annähernd gleich ist.

### 3.2.1 CoMFA Feldtypen

Das sterische Feld entspricht den van-der-Waals-Wechselwirkungsenergien zwischen einem an den Gitterpunkten platzierten ungeladenen Atom und dem Molekül. Die Berechnungsgrundlage ist das Lennard-Jones-Potenzial, welches wie folgt definiert ist:

$$E_{vdW(j)} = \sum_{i=1}^n (A_{ij}r_{ij}^{-12} - C_{ij}r_{ij}^{-6}) \quad (3.5)$$

Dabei ist  $E_{vdW}$  die Summe der van-der-Waals-Wechselwirkungsenergien,  $r_{ij}$  der Abstand zwischen einem Atom  $i$  des Moleküls und dem Gitterpunkt  $j$ . Die van-der-Waals-Radien der beteiligten Atome sind in den Konstanten  $A_{ij}$  und  $C_{ij}$  berücksichtigt. Als Sondenatom wird meistens ein  $sp^3$ -hybridisiertes Kohlenstoffatom verwendet. Das sterische Feld beschreibt somit mögliche Ligand-Rezeptor-Wechselwirkungen ohne Berücksichtigung der Elektrostatik. Negative Potenzialwerte entsprechen einer Anziehung und positive einer Abstoßung der Teilchen. Das zweite wichtige CoMFA-Feld beschreibt mögliche elektrostatische Interaktionen zwischen Molekül

und Rezeptor, wobei man sich hierfür der Coloumb-Energie  $E_C$  (s. Gleichung 3.6) bedient.

$$E_{C(j)} = \sum_{i=1}^n \frac{q_i q_j}{\varepsilon r_{ij}} \quad (3.6)$$

Hierbei sind  $q_i$  und  $q_j$  die Ladungen der Atome des Moleküls bzw. der Sonde,  $r_{ij}$  der Abstand beider voneinander und  $\varepsilon$  die Dielektrizitätskonstante. Da als Sonde üblicherweise eine positive Elementarladung ( $q_j = +1$ ) zum Einsatz kommt, ist die erhaltene Energie für positiv geladene Atome (gleichsinniger Charakter) positiv und für negativ geladene Atome (gegensinniger Charakter) negativ. Der abstandsabhängige Verlauf der Feldfunktionen ist in Abbildung 3.7 zu sehen.

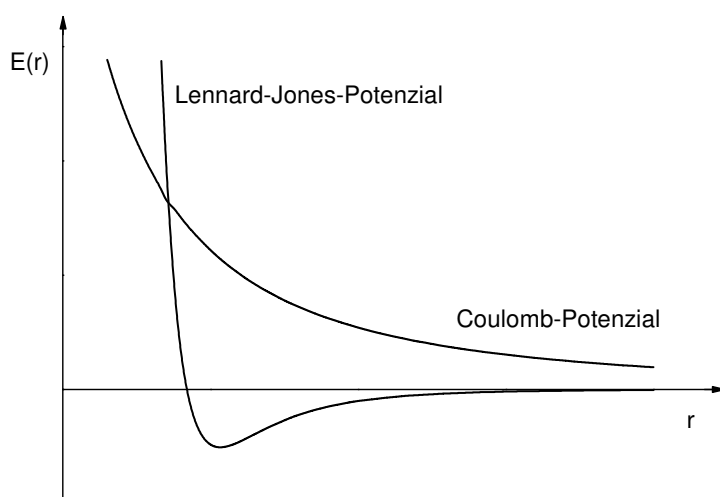


Abbildung 3.7: Sterische und elektrostatische Potenzialfunktion

Da Cramer und Wold dieses Verfahren unter der Nummer 5025388 in den USA patentierten [67], ist es nur im Softwarepaket SYBYL<sup>®</sup> von TRIPOS [68] verfügbar. Solange der Patentschutz andauert (bis 2007), wird sehr wahrscheinlich niemand eine 3D-QSAR Methode, die die Korrelation von Gitterfeldwerten mittels PLS zur Basis hat, in seine Modellingssoftware integrieren.

### 3.2.2 Vorteile und Probleme des CoMFA-Verfahrens

Vorteile der CoMFA sind:

- Es ist heute eine Standardmethode der 3D-QSAR.
- Es liefert robuste Vorhersagen.
- Die Ergebnisse sind einfach zu visualisieren.

Natürlich besitzt diese Methode auch einige Nachteile:

- Abhängigkeit vom Alignment
- Abhängigkeit von der Positionierung im Gitter
- Abhängigkeit von der Gitterdichte (in Zusammenhang mit ihrer Positionierung zu sehen)
- Es kann immer nur eine Konformation berücksichtigt werden.

Die relative Anordnung von Molekülen mit einer festgelegten Konformation zueinander bezeichnet man als Überlagerung oder Alignment. Die Alignmentabhängigkeit stellt das schwierigste Problem der CoMFA-Methode dar. Eine universell anwendbare Lösung kann es dafür auch nicht geben. Für ein realitätsnahes Modell sollten die Verbindungen in ihrer rezeptorgebundenen Form überlagert werden. Das betrifft sowohl die Konformation, die der bioaktiven Konformation entsprechen sollte, als auch die Orientierung zum (fiktiven) Rezeptor.

Damit befindet man sich in einem Dilemma, da in der Mehrzahl der Fälle weder der Rezeptor, noch die bioaktive Konformation, noch der Bindungsmodus der Liganden bekannt sind. Entlastend sei hier erwähnt, dass es sich bei Annahme einer allen Liganden gemeinsamen Konformation nicht um die bioaktive Konformation handeln muss. CoMFA extrahiert und korreliert nur die Unterschiede zwischen den Molekülen. Somit ist das Ergebnis für beliebige Konformationen ähnlich, sofern diese für alle Verbindungen gleich sind.

Es gibt eine Vielzahl von Ansätzen, Moleküle zu überlagern, und es gibt ebenso viele Kriterien für die Güte der Überlagerung. Die angewandten Alignmentmethoden zu erklären, würde ein separates Buch füllen und ist auch nicht Thema dieser Arbeit. Im allgemeinen beruhen die meisten Methoden auf der Minimierung des RMSD (s. Gleichung 3.2) von gemeinsamen Merkmalen der betrachteten Moleküle. Diese Minimierung kann durch verschiedene Verfahren erreicht werden. Häufig werden genetische Algorithmen, Gradientenverfahren und Monte-Carlo-Methoden angewandt. Die Auswahl und Gewichtung der Merkmale sowie die Auswahl der geeigneten Konformation sind für ein erfolgreiches Alignment ausschlaggebend. Hier sollen nur einige Möglichkeiten erwähnt werden:

- Überlagerung des gemeinsamen Grundgerüsts und möglichst gleichartige Orientierung von Substituenten
- Überlagerung anhand von Merkmalen einer vorhandenen Pharmakophorhypothese/gemeinsame Ausrichtung in Bezug auf wahrscheinliche Bindungspartner auf der Rezeptorseite (z. B. mögliche Wasserstoffbrückenbindungspartner)
- Überlagerung der Elektronendichte (siehe z. B. [69])
- Überlagerung des elektrostatischen Potenzials (als Ergebnis der Elektronendichte)
- Verwendung der für die CoMFA berechneten Felder für die Überlagerung
- Anpassung flexibler Verbindungen an eine möglichst sehr aktive aber rigide (bzw. konformativ stark eingeschränkte) Verbindung

Standardmäßig wird bei CoMFA ein Gitterabstand von 2 Å verwendet. Ursprünglich geschah dies, um den rechnerischen Aufwand und den Speicherbedarf handhabbar zu halten, denn die Zahl der Gitterpunkte und damit auch die Zahl der zu verarbeitenden Variablen steigt in der dritten Potenz mit der Verkleinerung (Teilung) des Gitterabstand. Mit der gegenwärtig verfügbaren Rechenleistung einfacher Personalcomputer sind nun auch erheblich kleinere Gitterabstände in vertretbarer Zeit

berechenbar geworden. Mit der Zahl der Gitterpunkte steigt aber auch der Anteil des Rauschens, d. h. es steigt die Zahl der Variablen, die wenig oder gar keine Informationen über Molekülunterschiede tragen, die für die Aktivität relevant sind.

Cramer selbst zeigte zwar, dass die PLS-Methode relativ robust gegenüber nicht-signifikanten Variablen ist [70, 71], jedoch gilt das nur für Datensätze mit einer großen Varianz bei den abhängigen Variablen. Der Wunsch, die unwichtigen Variablen mittels genetischer Algorithmen oder Programmen wie GOLPE [72] zu entfernen, ist deshalb verständlich. In Fällen mit insgesamt genügend großer Varianz wird das Problem der hohen Zahl von Variablen mit geringer Varianz recht effektiv dadurch gelöst, dass alle Variablen eine festgelegte Mindestschwankungsbreite aufweisen müssen. Legt man diesen Wert auf 1 bis 2 kcal fest, eliminiert man somit meist bis zu 90 % der unabhängigen Variablen.

### 3.2.3 Verbesserung des CoMFA-Verfahrens mittels Automation

Bei der CoMF-Analyse gibt es eine Reihe von Parametern, die einen nicht geringen Einfluss auf die Modellgüte haben und deren optimale Einstellung von Hand bestenfalls sehr zeitaufwändig bzw. unmöglich ist. Somit liegt der Einsatz von automatisch ablaufenden Programmen zur Optimierung der CoMFA-Modelle nahe.

#### All-Orientation-/All-Placement-Search

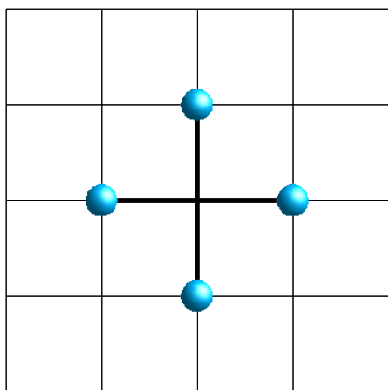
Ein wichtiges Kriterium ist die optimale Positionierung im CoMFA-Gitter, so dass Moleküloberflächenpunkte in für die Aktivität relevanten Bereichen keinen zu geringen Abstand von den Gitterpunkten haben. Durch die Verwendung des Lennard-Jones-Potenzials für die sterische und des Coulomb-Potenzials (s. Abbildung 3.7) für die elektrostatische Wechselwirkung ergibt sich ein Problem bei zu großer Annäherung der Moleküloberflächenpunkte an die Gitterpunkte.

Der Anstieg der beiden Potenzialfunktionen ist unterhalb eines gewissen Abstandes der Wechselwirkungspartner so groß, dass sehr hohe Energiewerte berechnet werden. Diese würden in einer Regression eine viel höhere Bedeutung gegenüber den

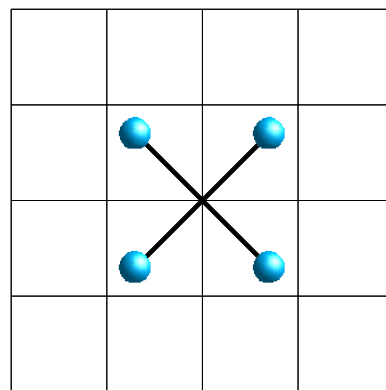
restlichen Feldwerten erhalten und so den wahren Zusammenhang überdecken. Aus diesem Grund ist bei der CoMF-Analyse eine Obergrenze für diese Werte vorgesehen, so dass alle Gitterpunkte, für die ein höherer Wert berechnet wird, auf diesen Maximalwert gesetzt werden.

Dadurch tritt nun ein gegenteiliges Problem auf: Wurden ohne Obergrenze unwichtige Bereiche des Moleküls überbewertet, werden mit dieser Obergrenze nun potenziell wichtige Unterschiede nivelliert. Da die Interaktionsenergien weiter entfernt liegender Gitterpunkte aber meist unter die Schwelle fallen, sind die Auswirkungen nicht so nachteilig, wie ohne die Beschränkung der Energiewerte.

Abbildung 3.8 zeigt schematisch zwei mögliche Orientierungen. In Abbildung 3.8(a) befinden sich die Atome des Moleküls auf den Gitterpunkten, an denen auch die Interaktionspotentiale ermittelt werden, während das Molekül durch Rotation bei Abbildung 3.8(b) so orientiert ist, dass die Atome zwischen den Gitterpunkten liegen. Das führt bei 3.8(a) zu sehr hohen berechneten Feldwerten, welche dann auf einen einheitlichen maximalen Wert gesetzt werden, wodurch die Variable ihre Varianz verliert (wenn die Atome der Verbindungen an dieser Stelle exakt überlagert wurden).



(a) Die Atome liegen genau auf den Gitterpunkten



(b) optimale Positionierung zwischen den Gitterpunkten

Abbildung 3.8: Orientierung im Gitter

Über das Phänomen der Abhängigkeit der Modellgüte von der Orientierung im Gitter wurde schon 1995 berichtet. Cho et al. bemerkten eine Schwankung des  $q^2$ -Wertes

um bis zu 0,5 Einheiten [73]. Die Liganden wurden dabei systematisch um die x-, y- und z-Achse rotiert. Wang et al. untersuchten konsequenterweise auch die Platzierung im Gitter und versuchten dieses Problem mit einer automatischen Routine zu lösen [74]. Ihr Verfahren des All-Orientation Search and All-Placement Search rotiert und verschiebt die überlagerten Moleküle im CoMFA-Gitter systematisch um einen bestimmten Betrag bzw. Winkel und findet so die beste Position und Orientierung. Das Kriterium dafür ist der  $q^2$ -Wert. Sie untersuchten außerdem die Abhängigkeit der  $q^2$ -Werte von der Gitterdichte.

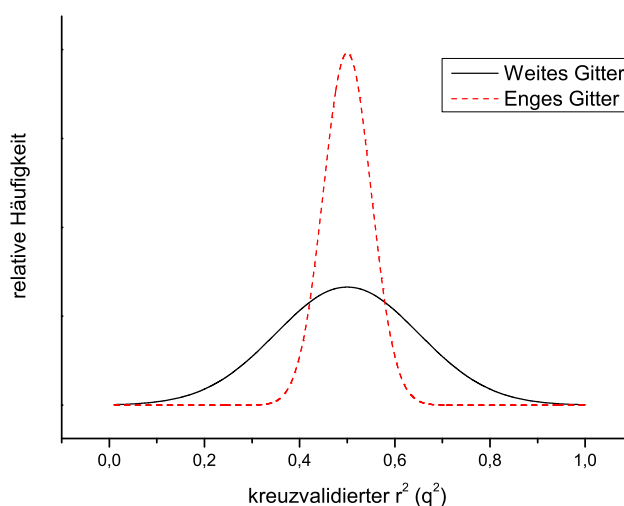


Abbildung 3.9: Die Verteilung der  $q^2$ -Werte in Abhängigkeit von der Gittergröße

Wie schon Cramer et. al. [75] darlegten, hat eine Erhöhung der Gitterpunktdichte keine direkte Erhöhung des  $q^2$ -Wertes zur Folge, da gleichzeitig mit der genaueren Abtastung der Moleküle (grössere Zahl an Information tragenden Variablen) die Anzahl der nichtinformativen Variablen steigt. Wang et al. verdeutlichten diese Aussage, indem sie zeigten, dass bei zunehmender Gitterpunktdichte nur die Variationsbreite um das Häufigkeitsmaximum der annähernd normalverteilten  $q^2$ -Werte geringer wird. Der am häufigsten vorkommende Wert bleibt gleich (s. Abbildung 3.9).

### Automatische PLS zur Selektion von Konformationen

Wie bereits erwähnt, stellt das Alignment und die Auswahl der richtigen Konformation eines der grundlegenden Probleme der CoMF-Analyse dar. Hier soll eine Methode vorgestellt werden, wie mit systematischen PLS-Analysen „falsche“ Konformationen identifiziert und ausgeschlossen werden können. Die generelle Vorgehensweise beruht auf der Analyse der Residuen nach einer durchgeführten PLS-Analyse mit Kreuzvalidierung. Bei der Kreuzvalidierung werden während der Modellerstellung ein (systematisch) oder mehrere (zufällig) Objekte ausgelassen, um deren Aktivität mit genau diesem Modell vorherzusagen. Dies wird solange wiederholt bis jedes Objekt genau einmal ausgelassen und seine Aktivität vorhergesagt wurde. Aus den Residuen — der Differenz aus gemessener und vorhergesagter Aktivität — lässt sich nach Formel 3.7 die Vorhersagekraft ( $q^2$ -Wert) ermitteln.

$$q^2 = 1 - \frac{\sum (y_{\text{vorhergesagt}} - y_{\text{gemessen}})^2}{\sum (y_{\text{gemessen}} - y_{\text{gemittelt}})^2} \quad (3.7)$$

Es soll angemerkt werden, dass der Nutzen des  $q^2$ -Werts und seine Interpretation kontrovers diskutiert werden [76–78]. Abhängig von den verwendeten Daten bedeutet ein hoher  $q^2$  nicht immer auch eine hohe externe Vorhersagekraft des Modells. Welcher  $q^2$ -Wert „sinnvoll“ ist, hängt von der jeweiligen Situation ab. Meist wird ein QSAR-Projekt mit kleinen fehlerbehafteten Datensätzen begonnen, wobei die Intention zunächst die Auswahl von aussagekräftigen Deskriptoren ist. Hier kann schon ein  $r^2$ -Wert von 0,7 und ein  $q^2$ -Wert ab 0,3 sinnvoll sein. Bei Anwachsen des Datensatzes und zunehmender Homogenität des Deskriptorraumes sollten  $r^2$ -Wert und  $q^2$ -Wert ansteigen und gegen den gleichen Wert konvergieren, der natürlich (ebenso wie der Standardfehler der Vorhersage) durch die praktischen Fehlergrenzen des Experiments beschränkt wird.

Mit der hier beschriebenen Methode wurde versucht, die besser vorhergesagten Konformationen im Modell zu konservieren und schlechter vorhergesagte Konformatio-



nen (mit hohen Residuen) zu eliminieren. Dabei sind zwei Strategien möglich: Beim Backward-Verfahren beginnt man mit Modellen, die mehrere Konformationen pro Verbindung enthalten und eliminiert die jeweils schlechter vorhergesagten. Vorteilhaft ist hier der geringere Aufwand, jedoch führt die Redundanz in den Daten zu Artefakten bei der Vorhersage.

Da bei Leave-One-Out-Kreuzvalidierung Verbindungen nicht komplett ausgeschlossen werden und Teile der Information durch die verbliebenen Konformationen bei der Modellerstellung berücksichtigt werden, ist die Vorhersage immer etwas zu „gut“. Der  $q^2$ -Wert wird demzufolge beim Fortschreiten des Verfahrens niedriger, was ihn als mögliches Optimierungskriterium ausschließt. Die Forward-Strategie beinhaltet nur eine Konformation pro Modell und tauscht Konformationen mit hohen Residualwerten aus. Das Blockschema dieser Optimierung ist in Abbildung 3.10 dargestellt.

Vorteilhaft ist hier, dass sich der  $q^2$ -Wert erhöhen sollte, sobald man erfolgreich eine Konformation durch eine „passendere“ ausgetauscht hat. Der größte Nachteil besteht im exponentiell mit der Zahl der möglichen Konformationen pro Verbindung steigenden Aufwand, so dass eine direkte Anwendung mit allen Konformationen unmöglich sein kann.

Das Verfahren ist sowohl auf mehrere Konformationen als auch auf verschiedene Alignments anwendbar. Es ist lediglich durch die nötige Rechenzeit und kombinatorische Explosion der nötigen Rechnungen limitiert. Hat man beispielsweise für 20 Verbindungen lediglich 2 Konformationen zur Auswahl, ergeben sich  $2^{20} = 1048576$  durchzuführenden Analysen. Bei einer angenommenen Geschwindigkeit von einer QSAR pro Sekunde würde das systematische Durchrechnen mehr als 10 Tage dauern. Glücklicherweise kann man diese Zahl durch Gruppenbildung reduzieren, da es relativ unwahrscheinlich ist, dass keine Verbindung eine ähnliche Konformation zu einer anderen besitzt.

Im hier entwickelten Verfahren wurde die Geschwindigkeit durch zwei Maßnahmen gegenüber der konventionellen PLS gesteigert und somit ein höherer Durchsatz (ca. 20000 PLS/h) erreicht. Zunächst wird mit der vorgegebenen Konformationsauswahl

eine vereinfachte PLS durchgeführt. Es handelt sich dabei um das in SYBYL implementierte SAMPLS-Verfahren (SAMple-distance PLS) [79], welches von Bruce Bush bei Merck & Co., Ltd. entwickelt wurde. Diese SAMPLS-Analyse liefert nicht immer exakt dieselben Ergebnisse, wie eine vollständige PLS (welche außerdem eine ausführlichere Auswertung ermöglicht), jedoch ist die Abweichung gering.

Die Ausgabe der SAMPLS-Analyse wird abweichend vom SYBYL-Standardalgorithmus nach Maßgabe des kleinsten Standardfehlers analysiert. Es wird für die Zahl an Komponenten, bei der der Fehler am geringsten ist, der  $q^2$ -Wert bestimmt. Das SYBYL-Programm bestimmt die optimale Zahl der Komponenten einer PLS-Analyse nach dem maximalen  $q^2$ -Wert. Das führt häufig (s. Abschnitt 3.3.3) zu einer Erhöhung der Komponentenzahl bei gleichzeitig steigendem Fehler und nur geringfügig steigendem  $q^2$ -Wert.

Liegt der durch SAMPLS ermittelte  $q^2$ -Wert über einem vorgegebenen Schwellenwert, wird eine vollständige PLS durchgeführt. Dabei wird genau die Zahl der Komponenten benutzt, die nach Auswertung der SAMPLS-Analyse als optimal befunden wurde. Für die PLS wird eine Report-Datei generiert. Diese kann einzeln begutachtet werden, oder man wertet nach Beendigung des Verfahrens mehrere Reportdateien zusammen mit dem MATLAB-Programm `plsreport` aus.

Das MATLAB-Programm `plsreport` wertet die Residuen einer Liste von Report-Dateien aus und stellt deren Mittelwerte in einem Balkendiagramm dar. Ebenfalls erfasst wird die jeweilige Häufigkeit einer Konformation. Daraus lässt sich meistens erkennen, wenn eine Konformation häufiger als die anderen zur Überschreitung des Schwellen- $q^2$ -Wertes geführt hat, was wiederum für eine Konservierung dieser Konformation sprechen würde. In den meisten Fällen sind die häufigsten Konformationen auch mit den geringsten durchschnittlichen Residuen behaftet. Ist die QSAR-Beziehung hinreichend stabil, gewinnt man durch Auswahl der konservierten Konformationen ein Modell mit hoher Vorhersagekraft, welches ebenfalls sehr stabil ist.

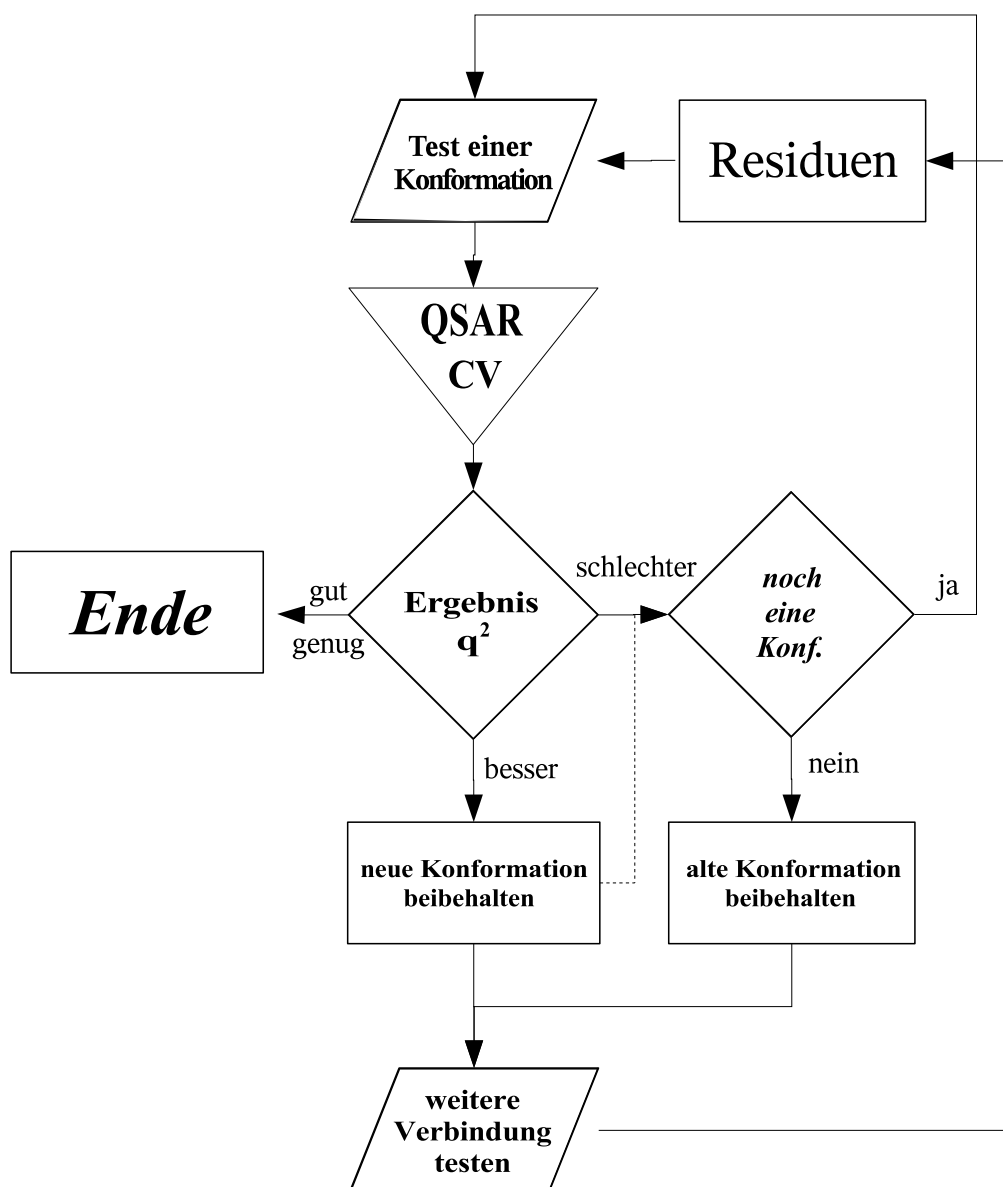


Abbildung 3.10: Blockschema der Forward-Optimierung

### Stabilität der von der automatischen PLS abgeleiteten Modelle

Ob die QSAR-Beziehung stabil genug ist, oder ob die hohen  $q^2$ -Werte durch Übertraining erzielt werden, kann man leicht anhand der Histogramme der  $q^2$ -Werte und SDEP-Werte (wie in den Abb. 3.11 und 3.12 zu sehen) abschätzen. Dort ist die Häufigkeit der  $q^2$ -Werte über die gesamte Optimierung abgebildet. Da nicht alle möglichen Konformationen permutiert werden, ist die Verteilung nicht symmetrisch, sondern auf der linken Seite abgeschnitten. Modelle mit um 0,2 Einheiten höheren  $q^2$ -Werten als das Maximum der Verteilung sind als bedenklich einzustufen. Bei diesen „extremen“  $q^2$ -Werten sind Modelle, die nach den Kriterien Häufigkeit und kleinste Residuen der Konformationen gewonnen werden, nicht mehr stabil. Die Konformationen, die man so auswählen würde, sind sehr heterogen. Nimmt man dagegen  $q^2$ -Werte um das Maximum der Verteilung, werden homogene Konformationen ausgewählt.

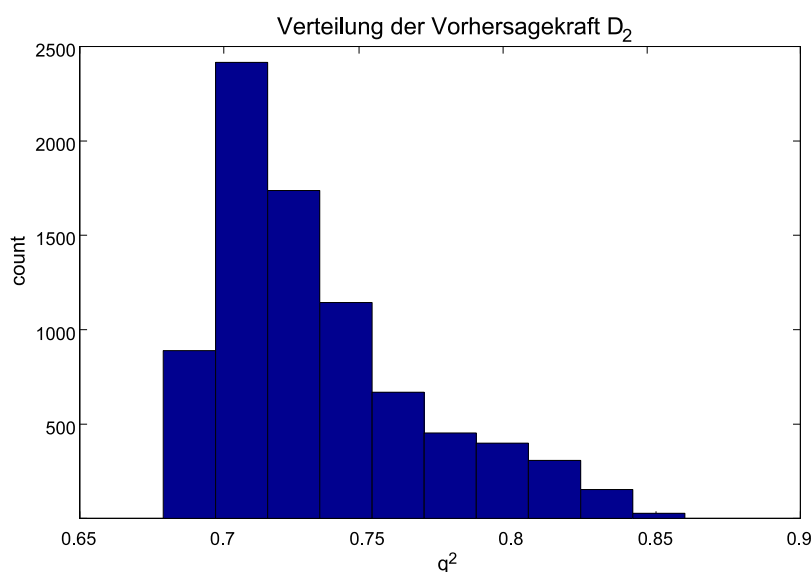


Abbildung 3.11: Verteilung der  $q^2$ -Werte aller durchgeführten SAMPLS-Analysen

Hat man ein Modell mit den als konserviert ermittelten Konformationen erstellt, kann man es auch leicht mit einer wiederholten „Random-Groups“-Kreuzvalidierung auf Stabilität testen. Dabei wird jeweils nicht nur eine Verbindung bei der Kreuzvalidierung ausgelassen, sondern, je nach Anzahl der Gruppen, bis zur Hälfte der

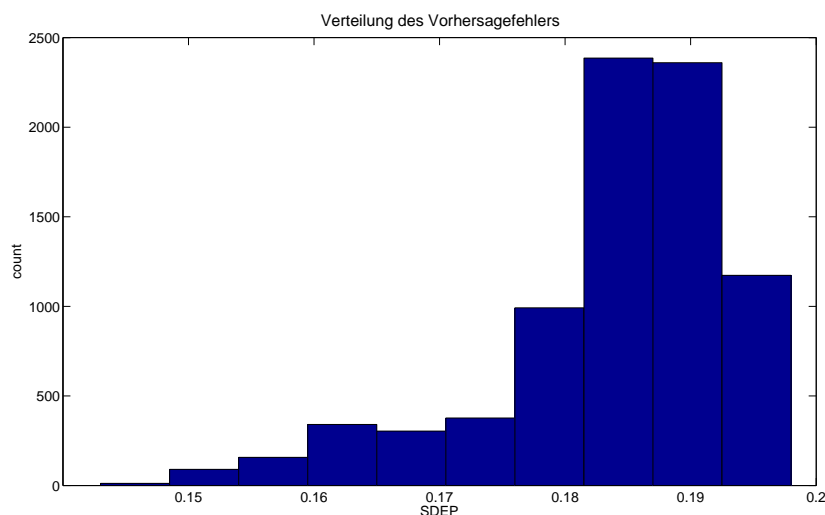


Abbildung 3.12: Verteilung der Vorhersagefehler aller durchgeführten SAMPLS-Analysen

Verbindungen. Üblicherweise werden minimal fünf Gruppen verwendet, so dass etwa ein Fünftel der Verbindungen als Testdatensatz dient. Je nach Menge der bei der Modellbildung vorenthaltenen Information sinkt die Vorhersagekraft etwas. Sie sollte jedoch bei einem stabilen Modell nicht drastisch einbrechen.

Da die Gruppenzusammenstellung zufällig erfolgt, ist es notwendig, die kreuzvalidierte PLS mehrfach durchzuführen und den Mittelwert und die Standardabweichung der Ergebnisse zu betrachten. Das SPL-Skript `random_groups_pls` tut genau das (s. Anhang D.6). Damit ist es möglich, in sehr kurzer Zeit 100 oder mehr Validierungen durchzuführen. Man erhält sowohl Mittelwert und Standardabweichung aller Berechnungen als auch die Einzelergebnisse in einer Ausgabedatei. In Abbildung 3.13 ist das Ergebnis einer solchen Validierung eines stabilen Modells zu sehen. Das zugehörige Einzelmodell ergab nach LOO-Kreuzvalidierung einen  $q^2$ -Wert von 0,84 (SDEP = 0,15).

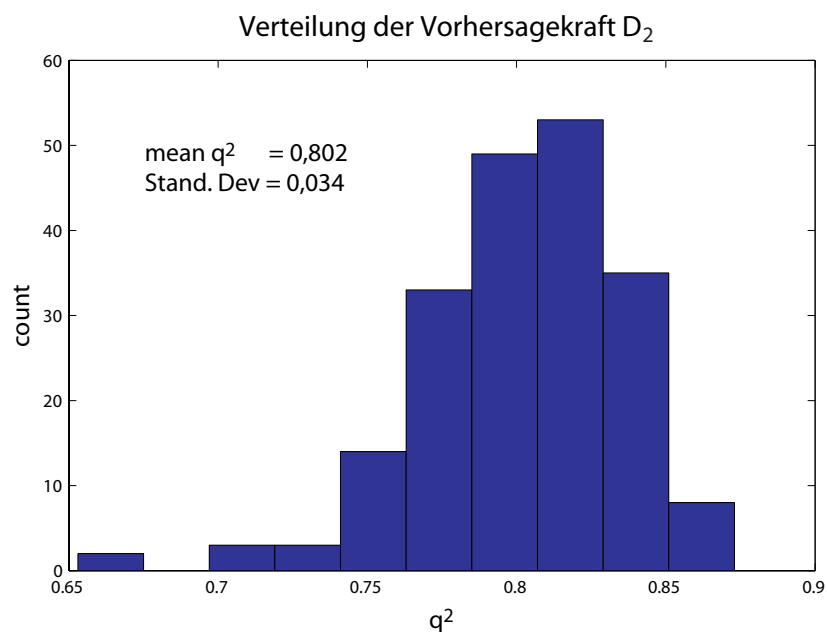
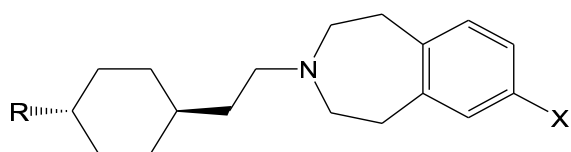


Abbildung 3.13: Verteilung der  $q^2$ -Werte von 200 „Random-Groups“-PLS mit 3 Kreuzvalidierungsgruppen

### 3.3 Automatische PLS am Beispiel der $D_2$ - und $D_3$ -Rezeptorantagonisten

Im Folgenden wird die Anwendung des Auto-PLS-Verfahrens auf den relativ homogenen Datensatz von ausgewählten Antagonisten des Dopamin  $D_2$ - und  $D_3$ -Rezeptors erläutert. Zunächst wurde versucht, die CoMFA-Modelle auf konventionelle Weise zu erstellen. Im Anschluß wurden diese durch das Auto-PLS-Verfahren optimiert.

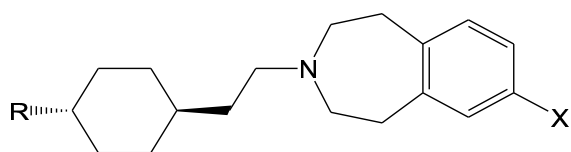
Bei den verwendeten Strukturen handelt es sich um 7-Methylsulfonyl-2,3,4,5-Tetrahydro-1H-3-Benzazepine und 7-Methylsulfonyloxy-2,3,4,5-Tetrahydro-1H-3-Benzazepine. Die Strukturen und Hemmkonstanten wurden der Veröffentlichung [24] entnommen. Tabelle 3.3 zeigt in der Übersicht die Strukturen der Verbindungen und ihre  $pK_i$ -Werte für die Dopaminrezeptoren  $D_2$  und  $D_3$ . Außerdem ist die Anzahl der unterschiedlichen Konformationen angegeben. Die möglichen Konformationen sind in Tabelle A.1 im Anhang A dargestellt. Sie ergibt sich aus den (durch eine dickere farbige Bindung angedeuteten) verschiedenen Orientierungen der Endgruppen im Rest R sowie aus einer möglichen Drehung des mittleren Aromaten.



MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)  
 MS = Methylsulfonyl (-SO<sub>2</sub>Me)

Nr.	<i>R</i>	<i>X</i>	<i>pK<sub>i</sub></i> (D <sub>2</sub> )	<i>pK<sub>i</sub></i> (D <sub>3</sub> )	<i>D<sub>3</sub>/D<sub>2</sub> Sel.</i>	<i>Konf.</i>
18		MSO	7,2	9,1	80	1
19		MSO	7,1	8,9	65	2
20		MSO	7,0	8,9	80	2
21		MSO	7,5	9,0	35	2
22		MSO	7,2	8,7	35	4
23		MSO	7,0	8,9	80	2
24		MSO	6,5	8,6	125	2
25		MSO	6,7	8,7	100	2
26		MSO	6,7	8,9	160	2
27		MSO	6,2	7,7	35	2
28		MSO	6,7	8,6	80	4

Tabelle 3.3a: Übersicht über die verwendeten Strukturen der Benzazepine und ihre Aktivitäten

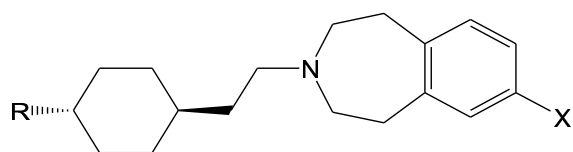


MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)  
 MS = Methylsulfonyl (-SO<sub>2</sub>Me)

Nr.	R	X	$pK_i(D_2)$	$pK_i(D_3)$	$D_3/D_2$ Sel.	Konf.
29		MSO	6,7	8,9	160	4
30		MSO	6,6	8,5	80	2
31		MSO	6,7	8,5	65	4
32		MSO	6,8	8,7	80	4
33		MSO	6,5	8,5	100	4
34		MSO	6,5	8,7	160	2
43		MS	7,1	9,0	80	1
44		MS	6,9	8,8	80	2
45		MS	6,9	8,9	100	2
46		MS	6,6	8,7	125	1
47		MS	6,4	8,6	160	1

Tabelle 3.3b: Übersicht über die verwendeten Strukturen der Benzazepine und ihre Aktivitäten





MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)  
MS = Methylsulfonyl (-SO<sub>2</sub>Me)

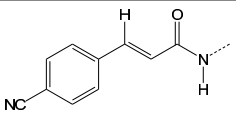
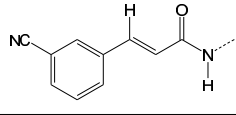
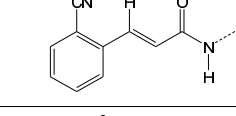
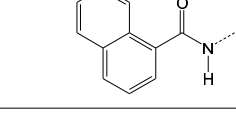
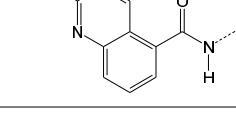
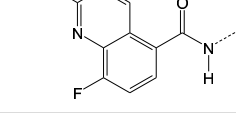
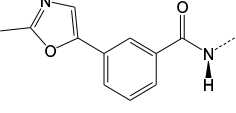
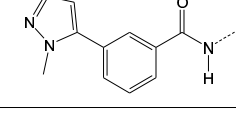
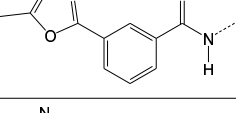
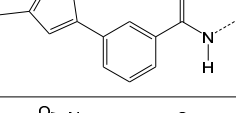
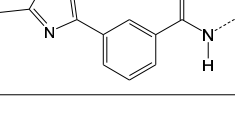
<i>Nr.</i>	<i>R</i>	<i>X</i>	<i>pK<sub>i</sub>(D<sub>2</sub>)</i>	<i>pK<sub>i</sub>(D<sub>3</sub>)</i>	<i>D<sub>3</sub>/D<sub>2</sub> Sel.</i>	<i>Konf.</i>
48		MS	6,9	8,7	65	1
49		MS	6,8	8,7	80	2
50		MS	6,8	8,8	100	2
51		MS	6,2	8,2	100	2
52		MS	6,5	8,4	80	2
53		MS	6,1	8,5	250	2
54		MS	6,2	8,2	100	4
55		MS	6,3	8,5	160	4
56		MS	6,2	8,1	80	4
57		MS	6,3	8,4	125	4
58		MS	6,4	8,4	100	4

Tabelle 3.3c: Übersicht über die verwendeten Strukturen der Benzazepine und ihre Aktivitäten

Zunächst wurde eine Konformationsanalyse der Grundstruktur, wie in Kapitel 3.1.3 auf Seite 27 erläutert, durchgeführt. Daraufhin erfolgte eine Geometrieoptimierung der möglichen Konformationen der Reste R mit quantenmechanischen Methoden. Anschließend wurden verschiedene Gruppen (Sets) bezüglich der Konformationen ausgewählt und mit diesen Sets CoMF-Analysen durchgeführt.

### 3.3.1 Konformationsanalyse der Grundstruktur

Das Simulated Annealing wurde über 200 Zyklen durchgeführt. Hierfür wurde Verbindung 18 benutzt, da sie eine sehr hohe Affinität zum D<sub>2</sub> und D<sub>3</sub>-Rezeptor besitzt und für den Rest R nur eine mögliche Konformation aufweist. Als dielektrische Funktion wurde „DISTANCE“ gewählt, d. h. die Größe der Dielektrizitätskonstante ist entfernungsabhängig. Die erhaltenen Strukturen wurden anschließend mit dem Kraftfeld MMFF94S [63, 64] minimiert. Die minimierten Strukturgerüste wurden mit dem MATCH-Programm verglichen und in MATLAB mit dem Programm `conf_elecT` analysiert.

Abbildung 3.14 und Tabelle 3.4 zeigen die Ergebnisse der Analyse. Die Verteilung der RMS-Werte zeigt, dass die Mehrzahl der Konformationen um 1.2–1.4 Å RMS verschieden sind. Daraufhin wurde die Schwelle für die Selektion der einander ähnlichen Konformationen auf 1.2 Å gesetzt und die vier häufigsten Konformationen in die engere Wahl gezogen. Wie aus Tabelle 3.4 ersichtlich ist, weist die Konformation R2256000 einen wesentlich höheren Energieinhalt auf und wurde deshalb nicht weiter berücksichtigt. Abbildung 3.15 zeigt die drei übrigen Konformationen in der Überlagerung, wobei die unterschiedlichen Gerüste verschieden eingefärbt wurden. R1104000 und R972000 verhalten sich fast spiegelbildlich zueinander. Weiterhin unterscheiden sich R1104000 und R1716000 hinsichtlich der Konformation des Cyclohexanrings. R972000 wurde aufgrund der niedrigsten Energie als Grundstruktur für alle Verbindungen ausgewählt. An diese Grundstruktur wurden die verschiedenen in Tabelle 3.3 aufgeführten Reste in einheitlicher Weise mit dem Programm SYBYL

angeknüpft, und anschließend wurden die Strukturen mit dem Kraftfeld MMFF94S minimiert.

Ergebnisse der MATLAB-Analyse (Schwelle = 1.2 Å)				
Nr.	Konformation	MMFF94s Energie (kcal)	Häufigkeit	Ø RMS
198	R972000	47.073	89	1.296
67	R1716000	48.694	86	1.287
10	R1104000	48.076	81	1.340
116	R2256000	53.470	101	1.252

Tabelle 3.4: Repräsentative Konformationen des Grundgerüsts der Benzazepine

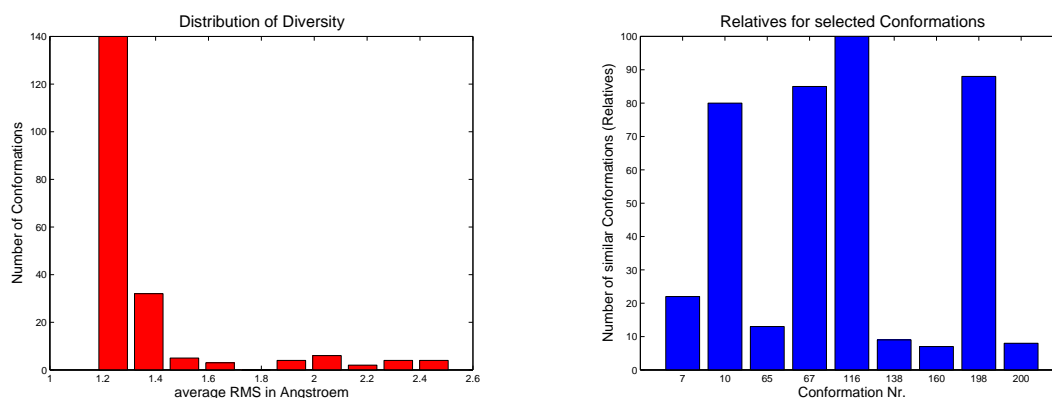


Abbildung 3.14: Diversitätsverteilung und Repräsentanten der Benzazepine

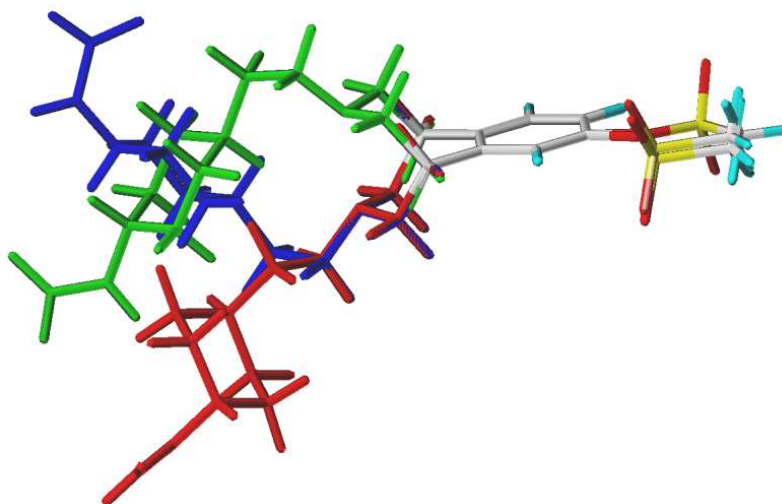


Abbildung 3.15: Ausgewählte Konformationen des Grundgerüsts der Benzazepine: blau – R1104000, rot – R1716000, grün – R972000

### 3.3.2 Überlagerung und Optimierung der Einzelkonformationen

Da die Verbindungen über ein einheitliches Grundgerüst verfügen, wurde dieses als Schablone für die Überlagerung benutzt. Die verschiedenen Reste R am Cyclohexanring hatten jedoch durch die Kraftfeldoptimierung jeweils leicht unterschiedliche Torsionswinkel erhalten. Entsprechend uneinheitlich war die Überlagerung der durch Optimierung mit dem Kraftfeld MMFF94S erhaltenen Konformationen (s. Abbildung 3.17).

#### Optimierung der Torsionswinkel

Um eine Vereinheitlichung zu erreichen, wurden die optimalen Torsionswinkel ab dem Cyclohexanring mit quantenmechanischen Methoden berechnet. Zunächst kam die AM1-Methode des Programms MOPAC 7.0 [80] zum Einsatz. Speziell die Torsionswinkel für an Aromaten gebundene Substituenten wichen jedoch von der Lehrbuchmeinung ab. Es zeigte sich, dass die semiempirische AM1-Methode hierfür nicht genau genug arbeitet bzw. nicht entsprechend parametrisiert ist. Daher wurde zur Geometrieoptimierung ein ab initio-Verfahren eingesetzt. Alle Berechnungen wurden mit der DFT/HF-Hybridmethode B3LYP und dem Basissatz 6-31G mit zusätzlichen p- und d-Orbitalen (6-31G\*\*) mit Hilfe des Programmes GAUSSIAN03 [81] durchgeführt. Diese Berechnungsmethode ist allgemein anerkannt als guter Kompromiss zwischen Genauigkeit der Ergebnisse und rechnerischem Aufwand. Die Ergebnisse entsprachen den vorherrschenden theoretischen Vorstellungen.

Allerdings wurde bei diesen Berechnungen deutlich, dass die Reste R am Cyclohexanring bis zu vier unterschiedliche Konformationen bzw. Orientierungen besitzen können. Jede mögliche Konformation wurde deshalb mit den folgenden Vorgaben berechnet. Der Rest der Verbindung 25 wurde stellvertretend für die Verbindungen mit naphthylartigen aromatischen Substituenten gewählt. Die Reste 28, 29, 30, 32, 34 und 55 wurden für Verbindungen mit phenylartigen aromatischen Substituenten und die Reste 20, 23, 47, 50 für Verbindungen mit cinnamoylartigen Substituenten

gewählt. Für die anderen Verbindungen wurden die Winkel entsprechend eingestellt. Die berechneten bzw. eingestellten Winkel und die für die Berechnung verwendeten Grundkörper sind in Abbildung 3.16 dargestellt. Die Winkel  $\alpha$  und  $\beta$  betragen in allen Fällen  $-143,3^\circ$  bzw.  $172^\circ$ . Tabelle 3.5 zeigt die letztendlich eingestellten Torsionswinkel aller berücksichtigten Konformationen der Verbindungen. Im Anhang A (s. Tabelle A.1) sind die möglichen Konformationen schematisch dargestellt. Das Ergebnis der Anpassung der Torsionswinkel ist in Abbildung 3.17 für die Grundkonformationen zu sehen.

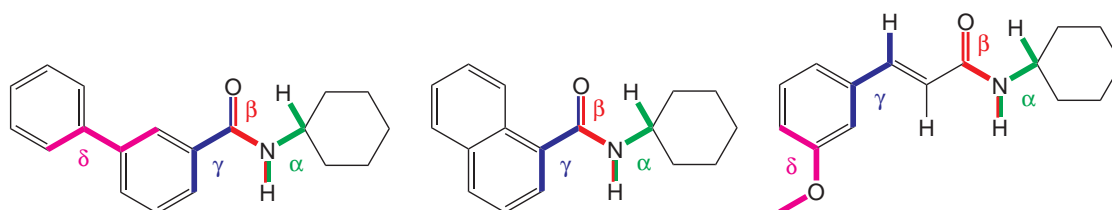


Abbildung 3.16: Bei den Benzazepinen veränderte Torsionswinkel

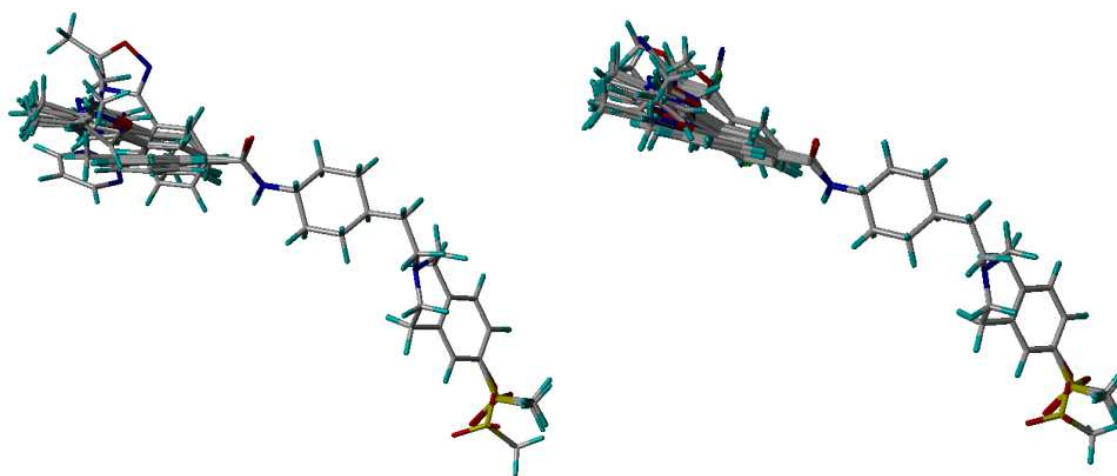


Abbildung 3.17: Überlagerung der Benzazepine vor und nach der Winkelanpassung

Konf.	$\gamma$	$\delta$
18	180	–
19	180	–
19a	0	–
20	0	–
20a	180	–
21	180	180
21a	180	0
22	180	0
22a	0	180
22b	180	180
22c	0	0
23	166	180.9
23a	180	0
24	140	–
24a	-40	–
25	140	–
25a	-40	–
26	140	–
26a	-40	–
27	140	–
27a	-40	–
28	157.1	180
28a	-22.9	0
28b	157.1	0
28c	-22.9	180
29	157.1	180
29a	-22.9	0
29b	157.1	0

Konf.	$\gamma$	$\delta$
29c	-22.9	180
30	157.1	-143.9
30a	-22.9	-143.9
31	157.1	180
31a	-22.9	0
31b	157.1	0
31c	-22.9	180
32	157.1	0
32a	-22.9	180
32b	157.1	180
32c	-22.9	0
33	157.1	180
33a	-22.9	0
33b	157.1	0
33c	-22.9	180
34	157.1	180
34a	-22.9	180
43	180	–
44	180	–
44a	0	–
45	180	–
45a	0	–
46	180	–
47	180	–
48	180	–
49	180	–
49a	0	–
50	173.8	–

Konf.	$\gamma$	$\delta$
50a	0	–
51	140	–
51a	-40	–
52	140	–
52a	-40	–
53	140	–
53a	-40	–
54	157.1	180
54a	-22.9	0
54b	157.1	0
54c	-22.9	180
55	157.1	138.7
55a	-22.9	138.7
55b	157.1	-41.3
55c	-22.9	-41.3
56	157.1	180
56a	-22.9	0
56b	157.1	0
56c	-22.9	180
57	157.1	0
57a	-22.9	180
57b	157.1	180
57c	-22.9	0
58	157.1	180
58a	-22.9	0
58b	157.1	0
58c	-22.9	180

Tabelle 3.5: Torsionswinkel, die bei den Substituenten eingestellt wurden

### 3.3.3 CoMF-Analysen und Ergebnisse

#### CoMFA-Ergebnisse der Grundkonformation

CoMFA-Ergebnisse für D <sub>2</sub> -Antagonismus					
Feld	SDEP	q <sup>2</sup>	NoC	Filter	Ausreißer
ster+ele	0,193	0,707	2	2,0	53(-0,48); 27(-0,39)
hydrophob	0,193	0,737	5	1,0	28(-0,41)
sterisch	0,192	0,749	6	2,0	47(-0,47)
elektrostatisch	0,172	0,775	3	2,0	21(0,34); 50(0,33); 55(0,33)

Tabelle 3.6: Ergebnisse der CoMF-Analyse für den Antagonismus an D<sub>2</sub>-Rezeptoren

CoMFA-Ergebnisse für D <sub>3</sub> -Antagonismus					
Feld	SDEP	q <sup>2</sup>	NoC	Filter	Ausreißer
ster+ele	0,284	0,286	6	2,0	27(-0,97)
hydrophob	0,264	0,304	3	1,0	27(-1,19); 53(1,32); 45(0,61)
sterisch	0,259	0,404	6	2,0	27(-0,8)
elektrostatisch	0,250	0,350	3	2,0	27(-0,88)

Tabelle 3.7: Ergebnisse der CoMF-Analyse für den Antagonismus an D<sub>3</sub>-Rezeptoren

Mit den so ausgerichteten Strukturen wurden erste CoMF-Analysen durchgeführt. Der Gitterabstand betrug 2 Å. Es wurden verschiedene Werte für das Column-Filtering eingesetzt, wobei sich meist 2,0 als optimal erwies. Column-Filtering bedeutet, dass nur Variablen mit der angegebenen Mindestvarianz für die Regression verwendet werden. Tabelle 3.6 zeigt, dass die Bindung an den D<sub>2</sub>-Rezeptor mit q<sup>2</sup>-Werten von 0,7 sehr gut vorhergesagt wurde, während die Vorhersagen der Bindung an den D<sub>3</sub>-Rezeptor sehr schlecht waren. Es kristallisierte sich dabei für die D<sub>3</sub>-Rezeptordaten die Verbindung 27 als Ausreißer heraus. Bei Auslassung dieser Verbindung (s. Tabelle 3.8) verbesserten sich die Vorhersageergebnisse sofort. Da die Vorhersagefehler der D<sub>3</sub>-Ergebnisse kleiner sind als die der D<sub>2</sub>-Ergebnisse, sind die trotzdem kleineren q<sup>2</sup>-Werte zum Teil auf die kleinere Standardabweichung der D<sub>3</sub>-Rezeptordaten (0,296 vs. 0,345) zurückzuführen.

In der Tabelle sind Verbindungen mit einer Vorhersageabweichung von mehr als dem Doppelten des Standardfehlers der Vorhersage (SDEP) für alle Verbindungen als Ausreißer aufgeführt. Die CoMSI-Analysen führten zu ähnlichen Ergebnissen. Die D<sub>3</sub>-Hemmkonstante der Verbindung 27 wurde konsequent zu hoch vorhergesagt. Daran änderte auch der Austausch der Konformation gegen 27a nichts. Während die Aktivität dieser Verbindung für den D<sub>2</sub>-Rezeptor in etwa den Aktivitäten ähnlicher Verbindungen entspricht, weicht sie für den D<sub>3</sub>-Rezeptor deutlich ab. Dafür sind zwei Ursachen denkbar:

- Die Methylgruppe am Chinolinring besetzt im Rezeptor einen „verbotenen“ Raum.
- Der Wert ist falsch.

Beide Varianten sind mit der gegebenen Datenlage nicht nachzuprüfen. Im Folgenden wurde versucht, die Vorhersagekraft der Modelle für den D<sub>3</sub>-Rezeptor zu verbessern.

<b>D<sub>3</sub>-Ergebnisse ohne Verbindung 27</b>				
<b>Feld</b>	<b>SDEP</b>	<b>q<sup>2</sup></b>	<b>NoC</b>	<b>Filter</b>
ster+ele	0,167	0,572	2	4,0
hydrophob	0,174	0,553	3	1,0
sterisch	0,173	0,544	2	2,0
elektrostatisch	0,167	0,591	3	2,0

Tabelle 3.8: Ergebnisse der CoMF-Analyse für den Antagonismus an D<sub>3</sub>-Rezeptoren ohne Verbindung 27

### Rationale Selektion der Konformationen

Angeichts der Vielzahl der möglichen Kombinationen der vorhandenen Konformationen war eine sinnvolle Startauswahl der Konformationen nötig. Es wurden 4 verschiedene Gruppen gebildet, die als Sets der SYBYL-Tabelle definiert wurden. Dabei wurde für die Substituenten eine möglichst gleiche Ausrichtung gewählt. Tabelle 3.9 zeigt die verwendeten Konformationen. Abbildung 3.18 soll die jeweilige Ausrichtung verdeutlichen.



Set	Auswahl
1	18 19 20 21 22 23 24 25 26 27 28A 29A 30A 31A 32A 33A 34A 43 44 45 46 47 48 49 50 51 52 53 54A 55A 56A 57A 58A
2	18 19 20 21 22 23 24 25 26 27 28B 29B 30A 31B 32B 33B 34A 43 44 45 46 47 48 49 50 51 52 53 54B 55B 56B 57B 58B
3	18 19 20 21 22 23 24 25 26 27 28C 29C 30A 31C 32C 33C 34A 43 44 45 46 47 48 49 50 51 52 53 54C 55C 56C 57C 58C
4	18 19 20 21 22 23 24A 25A 26A 27A 28C 29C 30A 31C 32C 33C 34A 43 44 45 46 47 48 49 50 51A 52A 53A 54C 55C 56C 57C 58C

Tabelle 3.9: Die Konformationen der vier Gruppen (Sets)

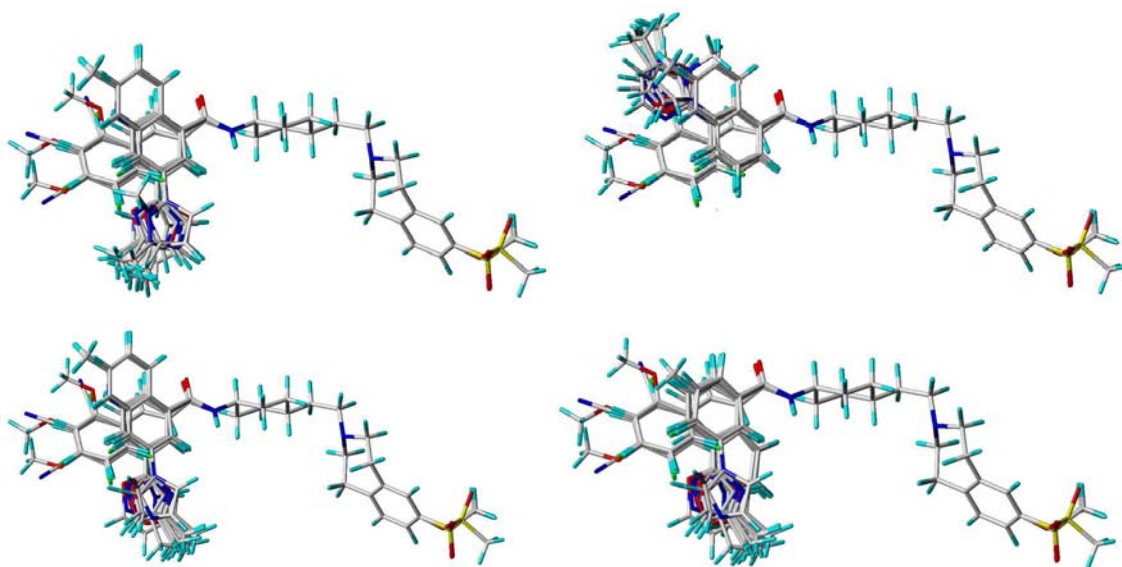


Abbildung 3.18: Vier verschiedene initiale Konformationsauswahlen

Mit diesen Selektionen wurde wiederum eine Reihe von CoMF-Analysen durchgeführt. Die Tabellen A.2, A.3 und A.4 im Anhang A zeigen die detaillierten Ergebnisse. Für den D<sub>2</sub>-Rezeptor blieben die  $q^2$ -Werte für alle berechneten Felder (außer dem elektrostatischen Feld) auf hohem Niveau (0,7–0,8). Für das elektrostatische Feld konnte die Vorhersagekraft außer bei Selektion 4 auf 0,85 erhöht werden. Dies geschah jedoch unter Einbeziehung von 8 bzw. 9 Komponenten, was eine mögliche Interpretation erheblich erschweren würde. Das Programm SYBYL ermittelt als optimale Komponentenzahl ausschließlich die mit dem maximalen  $q^2$ -Wert. Der Vorhersagefehler SDEP wird dabei nicht berücksichtigt. Es ist jedoch zu überlegen, ob ein geringerer Vorhersagefehler in einigen Fällen nicht sinnvoller wäre als ein um wenige Hundertstel höherer  $q^2$ -Wert. Erfahrungen zeigen, dass QSAR-Modelle mit ähnlich hohem  $q^2$ -Wert mit zunehmender Zahl einbezogener Komponenten in der Vorhersage externer Testdaten schlechter werden. Das deutet auf eine Übertrainierung des Modells hin. Die Aktivitäten der bei der Kreuzvalidierung jeweils ausgelassenen Verbindungen können zwar etwas besser vorhergesagt werden, aber die Fähigkeit zur Generalisierung geht verloren. Zur Demonstration wurde für die Selektion 3 mit dem elektrostatischen Feld eine weitere Berechnung für 2 Komponenten (statt 9) durchgeführt. Hier ist der Fehler minimal und der  $q^2$ -Wert mit 0,797 (statt 0,82) etwas niedriger. Der kleinste Standardfehler aller CoMFA-Berechnungen lag bei 0,154.

Die Vorhersagekraft für den D<sub>3</sub>-Antagonismus verbesserte sich ebenfalls durch Verwendung der Selektionen. Der kleinste Standardfehler lag hier bei 0,215. Verbindung 27 (bzw. 27a) stellte sich auch hier wieder als Ausreißer heraus, weshalb die Rechnungen ohne diese Verbindung wiederholt wurden. Daraufhin erhöhte sich der  $q^2$ -Wert auf über 0,6. Der Standardfehler verringerte sich auf minimal 0,152.

Leider konnte mit den CoMFA-Feldern keine befriedigende Korrelation für die Selektivität (dem Verhältnis von D<sub>3</sub>- zu D<sub>2</sub>-Antagonismus) gefunden werden. Korreliert wurde der Logarithmus der in Tabelle 3.3 angegebenen Werte. Die Berechnungen für die Selektivität wurden ebenfalls mit Auslassung der Verbindung 27 durchgeführt, da

deren Aktivität für den D<sub>3</sub>-Rezeptorantagonismus sehr schlecht vorhergesagt wurde. In jeder verwendeten Selektion traten dieselben Ausreißer auf.

Eine so niedrige Vorhersagekraft (maximal  $q^2 = 0,255$ ) bei wenigen angegebenen Ausreißern erscheint zunächst paradox. Jedoch ist die Spanne (0,7) und die Standardabweichung (0,19) der Selektivität nicht sehr hoch, wodurch die Fehler der Vorhersage nur sehr klein sein dürfen, damit ein hoher  $q^2$ -Wert erreicht wird. Der Standardvorhersagefehler (minimal 0,17) bei dieser QSAR ist aber bereits sehr hoch. Die Selektivität wird bei keiner Verbindung ähnlich genau wie ihre Aktivität an den einzelnen Rezeptoren vorhergesagt. Offensichtlich konnte man mit diesem Alignment bzw. mit dieser Auswahl der Konformationen die Selektivität nicht erklären.

### **Optimierung der Konformationsauswahl mit automatischer PLS**

Um die Möglichkeit unterschiedlicher Alignments und mehrerer Konformationen für die vergleichende molekulare Feldanalyse zu berücksichtigen, kam die automatische PLS (s. 3.2.3) zum Einsatz. Es wurde die Arbeitsweise der Forwardoptimierung angewandt, wobei das Verfahren parallelisiert wurde. Da die Vorhersagekraft für den D<sub>2</sub>- und D<sub>3</sub>-Antagonismus bereits recht gut war, wurde zunächst die Selektivität untersucht. Anschließend wurden zum Vergleich der D<sub>2</sub>- und D<sub>3</sub>-Antagonismus behandelt.

Im ersten Schritt wurde eine Vorooptimierung für die weitere Analyse vorgenommen. Die Abweichungen der manuellen PLS der Selektionen 1–4 wurden untersucht und die Verbindungen 22, 26, 29, 31, 52, 53 und 54 für die Optimierung ausgewählt. Die restlichen Verbindungen behielten die Konformation der Selektion 1. Daraus ergab sich eine Gesamtzahl von 2048 Möglichkeiten. Diese wurden für jedes Feld mittels PLS und Leave-One-Out-Kreuzvalidierung mit dem SAMPLS-Verfahren [79] berechnet. Es wurde ein Schwellenwert von 0,3 verwendet, d. h. ab einem mit SAMPLS ermittelten  $q^2$ -Wert von 0,3 wurde eine volle PLS-Analyse mit der durch Auswertung der SAMPLS-Ergebnisse ermittelten optimalen Anzahl an Komponenten durchgeführt und protokolliert. Die Ergebnisse dieser Vorooptimierung sind in Tabelle 3.10 zu sehen.

Feld	Anzahl mit $q^2 >$ Schwellenwert	SDEP	max. $q^2$	NoC
ster+ele	239	0,121	0,663	6
H-Brücken	6	0,167	0,375	7
sterisch	380	0,110	0,698	4
elektrostatisch	213	0,137	0,571	6

Tabelle 3.10: Ergebnis nach 2048 Permutationen der Konformationen der Verbindungen 22, 26, 29, 31, 52, 53 und 54. Der  $q^2$ -Schwellenwert betrug 0,3; Standard Error of Prediction (SDEP) und Number of Components (NoC) für Modell mit max.  $q^2$ .

Schon die Vorooptimierung konnte die Vorhersagekraft der Modelle erheblich steigern. Mit dem MATLAB-Programm `plsreport` wurden die durchschnittlichen Residuen der Konformationen sowie die Häufigkeit ihres Auftretens in einem Modell mit  $q^2$  oberhalb des Schwellenwertes berechnet. Im Anhang A sind die Ergebnisse der Analyse der Report-Dateien in den Tabellen A.5, A.6, A.7 und A.8 ausführlich dargestellt. Für das weitere Vorgehen wurden für jedes CoMFA-Feld einzeln alle Verbindungen mit Residualwerten  $> 0,1$  erneut optimiert. Die Ergebnisse dieser Optimierung sind in Tabelle 3.11 zusammengefasst. Der maximale  $q^2$ -Wert konnte nochmals um 0,1 Einheiten gesteigert werden.

Feld	Schwellenw.	Anzahl	$>$ Schwellenw.	SDEP	max. $q^2$	NoC
ster+ele	0,6	131072	602	0,104	0,761	7
H-Brücken	0,4	65536	410	0,151	0,518	8
sterisch	0,6	131072	1393	0,091	0,803	5
elektrostatisch	0,5	65536	1680	0,119	0,765	13

Tabelle 3.11: Ergebnis nach weiterer Optimierung durch die automatische PLS; Standard Error of Prediction (SDEP) und Number of Components (NoC) für Modell mit max.  $q^2$ .

Die Aufgabe, aus der Vielzahl guter Modelle eines auszuwählen, ist nicht einfach. Hinzu kommt die Frage, wie eine Überoptimierung verhindert bzw. erkannt werden kann. Hierzu muss der gemeinsame Nenner der guten Modelle gefunden werden. Dieser ergibt sich aus der Häufigkeit und den durchschnittlichen Residuen der (guten) Modelle. Ist die Vorhersagekraft dieses „gemeinsamen Nenners“ (Konsensusaus-

wahl) immer noch gut und übersteht es den „Random-Groups-PLS“-Stabilitätstest (s. 3.2.3), erfüllt dieses Modell die angenommenen Kriterien für Stabilität.

Ausgehend von den Modellen mit hohem  $q^2$ -Wert wurde jeweils eine solche Konformationsauswahl getroffen. Ausgewählt wurde die Konformation einer Verbindung mit der geringsten durchschnittlichen Abweichung, wenn keine andere Konformation wenigstens 50 % häufiger in guten Modellen vorkam. In Tabelle 3.12 ist diese Auswahl für das sterische Feld zu sehen. In den meisten Fällen entspricht die Konformation mit der geringsten mittleren Abweichung auch der häufigsten Konformation. War dies nicht der Fall — also gibt es Konformationen einer Verbindung mit geringeren durchschnittlichen Residuen — sind die Werte in der Tabelle kursiv gedruckt (was in Tabelle 3.12 bei den Konformationen 28, 29C und 31A der Fall ist).

Mit dieser Konsensauswahl (s. Tabelle 3.14) wurde eine konventionelle PLS durchgeführt. Da sie nur einen  $q^2$ -Wert von 0,31 erbrachte, wurde auf einen Stabilitätstest verzichtet. Um zu klären, ob Überoptimierung der Grund für das Versagen ist, wurden alle  $q^2$ -Werte der Optimierung in einem Histogramm dargestellt. Dadurch wurde ersichtlich, dass die durch die Optimierung erhaltenen hohen  $q^2$ -Werte weit über dem Verteilungsmaximum liegen. Der geringe  $q^2$ -Wert der Konsensauswahl liegt somit im Rahmen des Erwartbaren. Modelle mit einem  $q^2$ -Wert  $> 0,4$  sind Produkte des Übertrainings.

Für den  $D_2$ -  $D_3$ -Antagonismus wurde die Prozedur ebenfalls durchgeführt. Es wurde das kombinierte CoMFA-Feld „ster+ele“ ausgewählt, um den Aufwand etwas zu verringern.

Die Zusammenfassung der Daten dieser Optimierung ist in Tabelle 3.13 zu sehen. In Abbildung 3.20 ist die Verteilung der  $q^2$ -Werte über die gesamte Optimierung dargestellt. Dementsprechend wäre für die Konsensauswahl aus Tabelle 3.14 ein  $q^2$ -Wert  $> 0,7$  für  $D_2$  und  $> 0,55$  für  $D_3$  zu erwarten.

Mit Werten von  $q^2 = 0,84$  (SDEP = 0,150) für  $D_2$  und  $q^2 = 0,72$  (SDEP = 0,135) für  $D_3$  wurden die Erwartungen übertroffen. Der „Random-Groups-PLS“-Stabilitätstest

Konformation	Anzahl	Ø Residuen	Konformation	Anzahl	Ø Residuen
18	380	0,032	18	1393	0,053
19	380	0,110	19	796	0,091
20	380	0,042	20	1393	0,026
21	380	0,127	21	1123	0,097
22	292	0,061	22	1393	0,060
23	380	0,060	23	1393	0,031
24	380	0,133	24	870	0,114
25	380	0,028	25	1393	0,105
26A	196	0,139	26	642	0,024
28A	380	0,120	28	457	<i>0,076</i>
29C	134	0,187	29C	696	<i>0,165</i>
30A	380	0,077	30A	1393	0,061
31A	113	0,121	31A	587	<i>0,096</i>
32A	380	0,085	32A	1393	0,060
33A	380	0,102	33A	1393	0,092
34A	380	0,167	34A	797	0,140
43	380	0,083	43	1393	0,092
44	380	0,086	44	1393	0,101
45	380	0,014	45	1393	0,014
46	380	0,127	46	1393	0,124
47	380	0,230	47	1393	0,221
48	380	0,032	48	1393	0,051
49	380	0,016	49	1393	0,018
50	380	0,022	50	1393	0,033
51	380	0,131	51	773	0,096
52	241	0,257	52A	634	0,156
53A	227	0,176	53A	759	0,111
54A	108	0,043	54A	1393	0,041
55A	380	0,051	55A	1393	0,070
56A	380	0,145	56	611	0,111
57A	380	0,106	57C	459	0,059
58A	380	0,030	58A	1393	0,031

Tabelle 3.12: Durchschnittliche Residuen des sterischen Feldes vor (links) und nach (rechts) Optimierung für die Selektivität. Die Gesamtzahl der ausgewerteten CoMF-Analysen betrug 380 (links) bzw. 1393 (rechts). Kursiv gedruckte Werte stellen nicht die niedrigsten Residuen dar (andere Konformationen wurden mehr als 50 % häufiger selektiert).

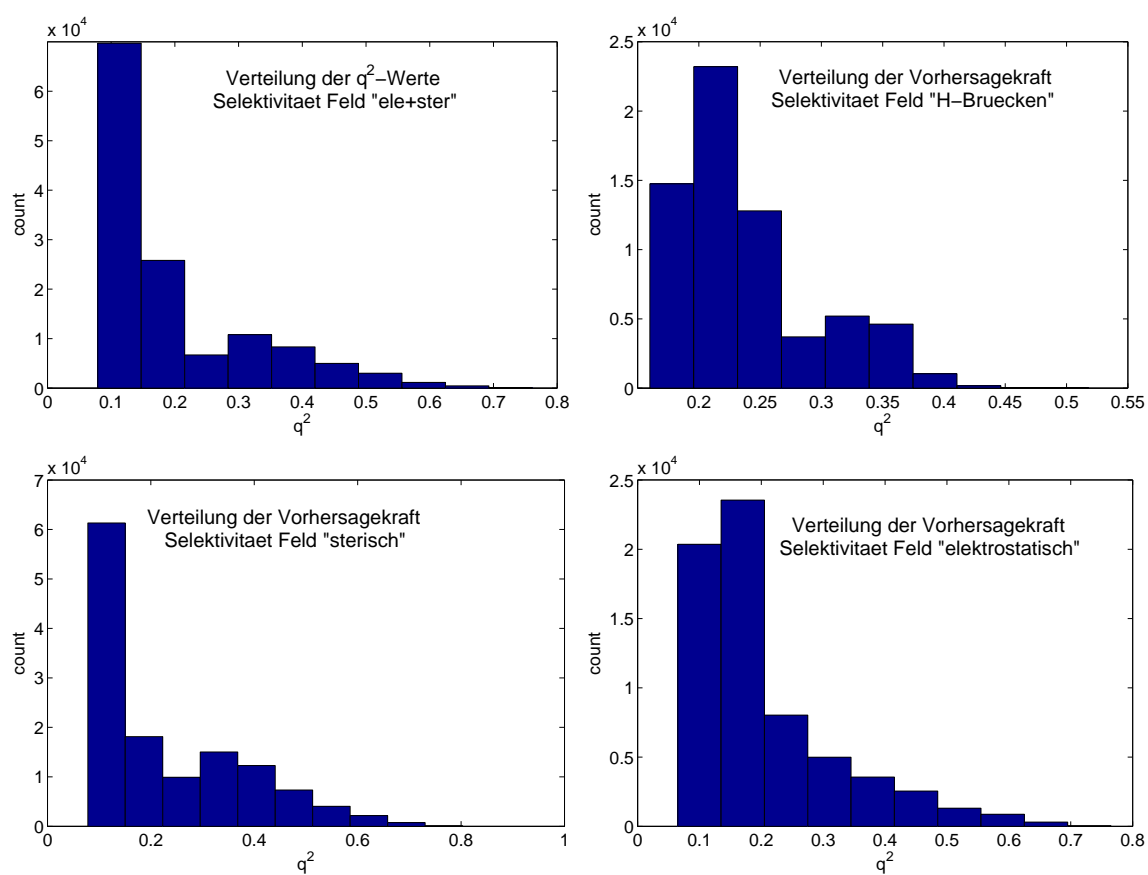


Abbildung 3.19: Histogramme der  $q^2$ -Werte über die gesamte Optimierung der Modelle für die  $D_2/D_3$ -Selektivität

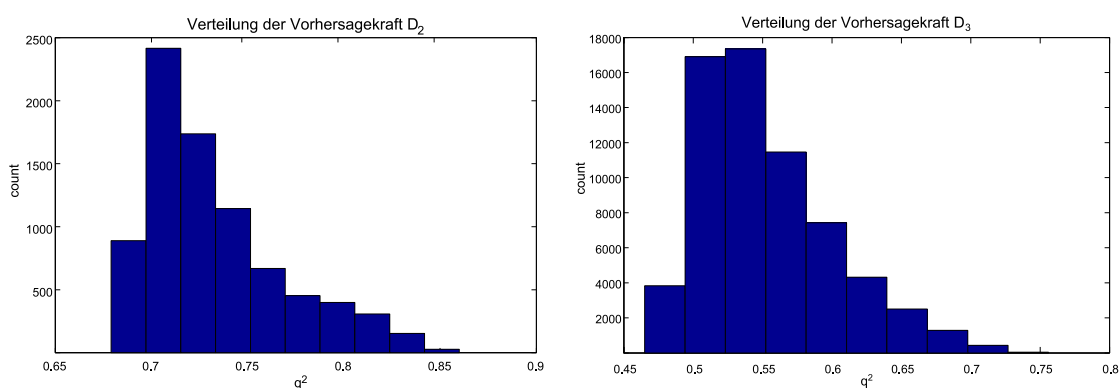


Abbildung 3.20: Histogramme der  $q^2$ -Werte (Feld „ster+ele“) über die gesamte Optimierung der Modelle für die  $D_2$ - und  $D_3$ -Antagonisten

Rezeptor	Schwellenwert	Anzahl	> Schwellenwert	SDEP	max. $q^2$	NoC
D <sub>2</sub>	0,6	8192	3185	0,143	0,86	6
D <sub>3</sub>	0,4	65536	6414	0,139	0,756	7

Tabelle 3.13: Ergebnis der Optimierung für den Antagonismus am D<sub>2</sub>- und D<sub>3</sub>-Rezeptor nach automatischer PLS; Standard Error of Prediction (SDEP) und Number of Components (NoC) für Modell mit max.  $q^2$ .

bestätigte die Stabilität der Modelle. Es wurden jeweils 200 PLS-Analysen mit einer Kreuzvalidierung von 3 Gruppen durchgeführt. Bei den Modellen für den D<sub>2</sub>-Antagonismus lag der durchschnittliche  $q^2$ -Wert bei 0,802 mit einer Standardabweichung von 0,034. Für die D<sub>3</sub>-Modelle ergab sich ein durchschnittlicher  $q^2$ -Wert von 0,701 mit einer Standardabweichung von 0,038 (s. Abbildung 3.21).

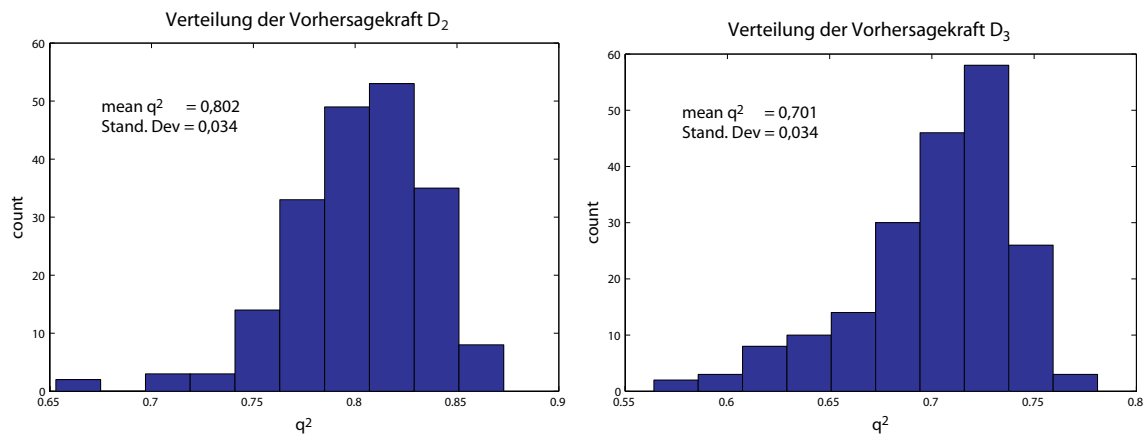


Abbildung 3.21: Verteilung der  $q^2$ -Werte von 200 „Random-Groups“-PLS-Analysen mit 3 Kreuzvalidierungsgruppen für die D<sub>2</sub>- (links) und D<sub>3</sub>-Modelle (rechts)



Konform.	Anzahl	Ø Residuen D <sub>2</sub>	Konform.	Anzahl	Ø Residuen D <sub>3</sub>
18	3185	0,041	18	6414	0,119
19	3185	0,053	19	6414	0,087
20	3185	0,115	20	6414	0,065
21A	1556	0,340	21	6414	0,034
22	1099	0,028	22C	2247	0,252
23	3185	0,083	23	6414	0,052
24	3185	0,122	24	6414	0,095
25	1595	0,105	25	6414	0,048
26	3185	0,117	26	3650	0,237
27	1715	0,375	27	—	—
28A	3185	0,056	28A	6414	0,049
29A	3185	0,102	29	2433	0,142
30A	3185	0,060	30A	3760	0,131
31A	3185	0,076	31A	6414	0,130
32	1076	<i>0,045</i>	32C	2176	0,061
33A	3185	0,147	33A	6414	0,132
34A	3185	0,152	34A	6414	0,072
43	3185	0,310	43	6414	0,308
44	1607	0,136	44	6414	0,113
45	1623	0,154	45	3434	0,220
46	3185	0,205	46	6414	0,041
47	3185	0,397	47	6414	0,108
48	3185	0,083	48	6414	0,063
49	3185	0,038	49	6414	0,032
50	3185	0,036	50	6414	0,100
51	3185	0,080	51A	2838	0,212
52	1723	0,270	52	6414	0,059
53A	1639	0,128	53	6414	0,072
54A	3185	0,128	54C	2130	<i>0,155</i>
55A	3185	0,032	55B	1604	0,042
56A	3185	0,064	56A	2557	<i>0,244</i>
57A	3185	0,038	57A	6414	0,069
58	1082	0,070	58A	6414	0,079

Tabelle 3.14: Residuen ausgewählter Konformationen des Feldes „ster+ele“ für die Rezeptoren D<sub>2</sub> und D<sub>3</sub>. Kursiv gedruckte Werte stellen nicht die niedrigsten Residuen dar (andere Konformationen wurden mehr als 50 % häufiger selektiert).

## 3.4 Auto-PLS zur Erstellung von CoMFA-Modellen für D<sub>1</sub>-Antagonisten

Durch die Forschung an Dopaminrezeptoren im Institut für pharmazeutische Chemie der Universität Bonn standen eine Reihe Rezeptorbindungsdaten zu Antagonisten des Dopamin D<sub>1</sub>-Rezeptors zur Verfügung. Aufgrund der Überlegungen aus Kapitel 2 wurden die Strukturen, für die Hemmkonstanten zum Dopamin D<sub>1</sub>-Rezeptor aus den Messungen von Frau A. Hamacher zur Verfügung standen, für die QSAR ausgewählt.

Für die durchgeführten CoMFA-Untersuchungen ist das Alignment ein wesentlicher Faktor. Das initiale Alignment basiert auf einem Pharmakophormodell, welches von dem Standardantagonisten ((R)-(+)-SCH 23390) abgeleitet wurde. Hierfür wurden die Konformationen aller Verbindungen mit der in Kapitel 3.1.3 vorgestellten Methode des Konformationsclustering untersucht. Bei der Verbesserung des Alignments kam die automatische PLS zum Einsatz. Die finalen CoMFA-Modelle wurden schließlich mit der Random-Groups-PLS-Methode auf ihre Stabilität überprüft.

### 3.4.1 Strukturen und biologische Daten

Die Strukturen der für die CoMFA herangezogenen Verbindungen sind in Abbildung 3.22 dargestellt und ihre am D<sub>1</sub>-Rezeptor gemessenen Hemmkonstanten in Tabelle 3.15 wiedergegeben. Aus den ebenfalls angegebenen Fehlerwerten lässt sich bei angenommener Leave-One-Out-Kreuzvalidierung ein maximal sinnvoller  $q^2$ -Wert (s. Gleichung 3.7) berechnen. Wird dieser von einem QSAR-Modell überschritten, liegt mit hoher Wahrscheinlichkeit ein Overfitting vor. Dieses QSAR-Modell würde die Hemmkonstanten mit einem Fehler vorhersagen, der kleiner ist als der der Experimente. Für diese Daten beträgt der maximal sinnvolle  $q^2$ -Wert 0,99, da die Abweichungen bei Radioligandbindungsuntersuchungen recht klein sind.

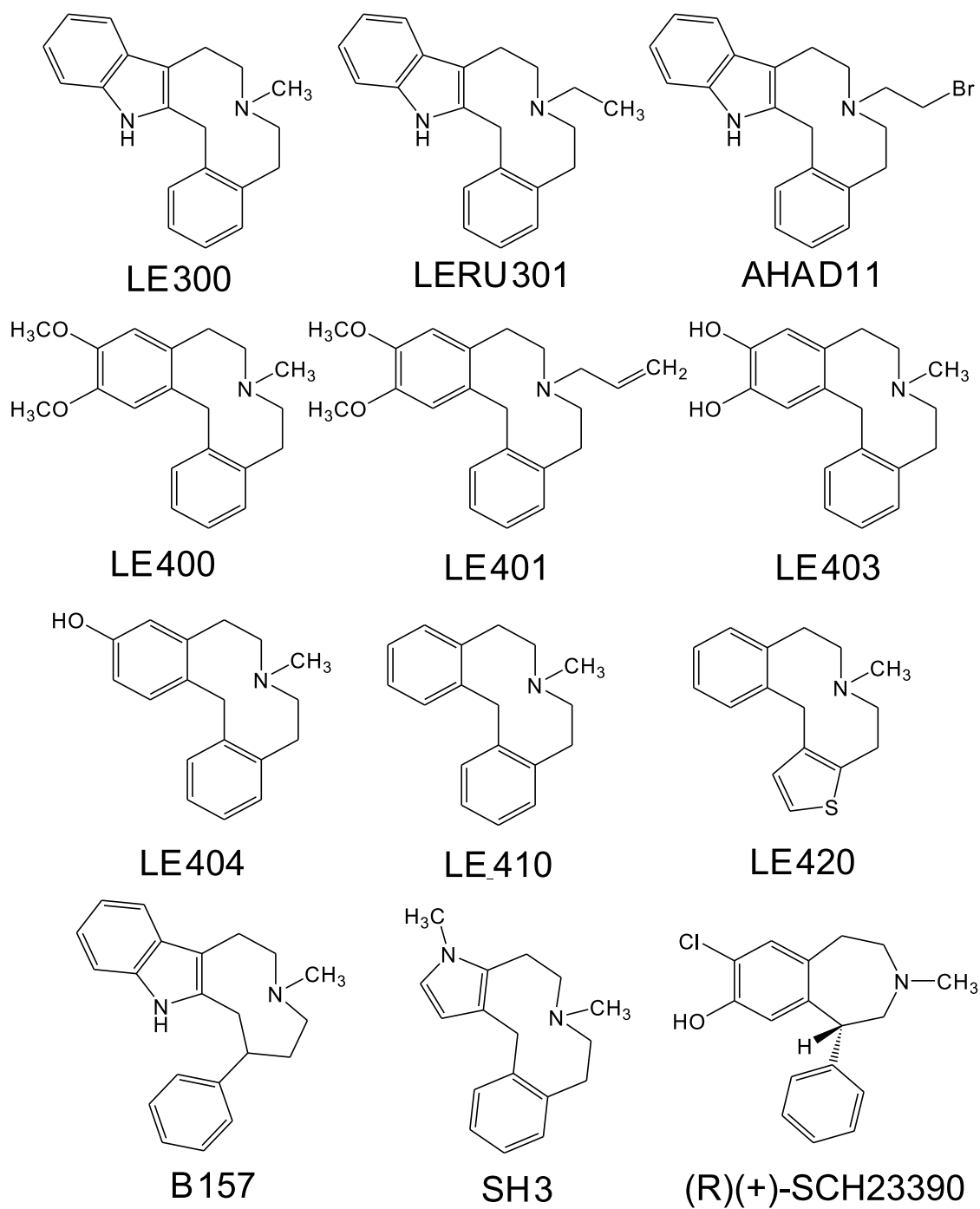


Abbildung 3.22: Übersicht über die 2D-Strukturen der ausgewählten Dopamin  $D_1$ -Antagonisten

Verbindung	$K_i$ in $\text{nmol}\cdot\text{l}^{-1}$	$\text{p}K_i$	$\log(\text{Fehler})$
LE 300	10,4	7,98	0,05
LE 400	2661	5,57	0,16
LE 401	16872	4,77	0,25
LE 403	11,5	7,94	0,06
LE 404	3,4	8,47	0,10
LE 410	17,5	7,76	0,04
LE 420	128	6,89	0,07
(R)-(+)-SCH 23390	2,91	8,54	0,11
AHA D11	1501	5,82	0,07
B 157	104,8	6,98	0,05
LERU 301	55,5	7,26	0,03
SH 3	682,8	6,17	0,04

Tabelle 3.15: Rezeptorbindungsdaten am  $D_1$ -Rezeptor für die ausgewählten Antagonisten

### 3.4.2 Pharmakophorbasiertes Alignment als Ausgangspunkt für die Auto-PLS

Für die CoMFA-Modelle wurde ein pharmakophorbasiertes Alignment benutzt. Als Schablone (Template) sollte (R)-(+)-SCH 23390 zum Einsatz kommen. Um ein geeignetes Konformer auszuwählen, wurde eine Konformationsanalyse (wie in Kapitel 3.1.3 beschrieben) durchgeführt. Bei einem zugrunde gelegten durchschnittlichen RMS von  $0,6 \text{ \AA}$  wurden vier Konformationscluster ermittelt. Die globale Minimumkonformation ist R490000. Sie besitzt einen mittleren RMS von  $0,67 \text{ \AA}$  zu allen anderen 99 durch das Simulated Annealing erhaltenen Konformeren. Ebenfalls durchgeführte NMR-Untersuchungen bestätigten diese Ergebnisse. Sie belegen auch, dass bei Raumtemperatur keines der häufigsten Konformere bevorzugt vorliegt. Die detaillierten Ergebnisse der Konformationsanalyse sind im Anhang C auf Seite 130 und im Anhang B ab Seite 117 dargestellt.

Diese konformationelle Flexibilität veranlasste Chipkin et al. [82, 83], rigidisierte Analoga von SCH23390 zu synthetisieren, um weitere Hinweise auf die bioaktive

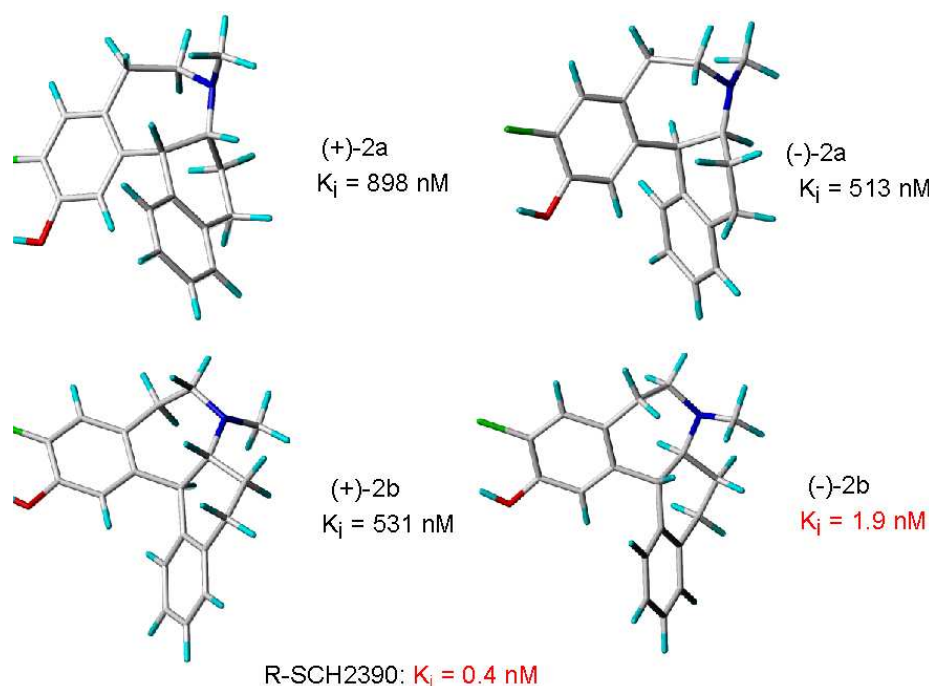


Abbildung 3.23: Die globalen Minimum-Energie-Konformationen der vier Stereoisomere von SCH 39166, angegebene  $K_i$  wurden am Rattenstriatum gemessen.

Konformation zu erhalten. Wie Abbildung 3.23 zeigt, sind die vier Stereoisomere von SCH 39166 durch eine Ringverknüpfung vom Azepinring zum Phenylring von SCH23390 abgeleitet.

Die angegebenen Hemmkonstanten wurden an  $D_1$ -Rezeptoren aus dem Rattenstriatum gemessen und sind somit nicht vollständig mit Daten von humanen Dopaminrezeptoren vergleichbar (s. auch Kapitel 2.2.1). Da insgesamt eine befriedigende Korrelation zwischen Bindungsdaten an menschlichen Dopaminrezeptoren und denen der aus der Ratte gewonnenen besteht ( $r^2 = 0,94$ , s. Abbildung 2.2), sollten zumindest die relativen Aktivitäten von SCH 39166 vergleichbar sein (was von M. A. Tice [84] bestätigt wurde). Für das Analogon (R)-(+)-SCH 23390 wurde am Rattenstriatum ein  $K_i$ -Wert von  $0,4 \text{ nmol} \cdot \text{l}^{-1}$  angegeben, was in der selben Größenordnung liegt, wie beim Humanrezeptor. Leider konnte die Firma Essex Pharma - GmbH (bzw. Schering - Plough) auf Nachfrage bis zum Zeitpunkt dieser Arbeit keine der SCH 39166 Verbindungen für eine Testung an humanen Dopaminrezeptoren zur Verfügung stellen.

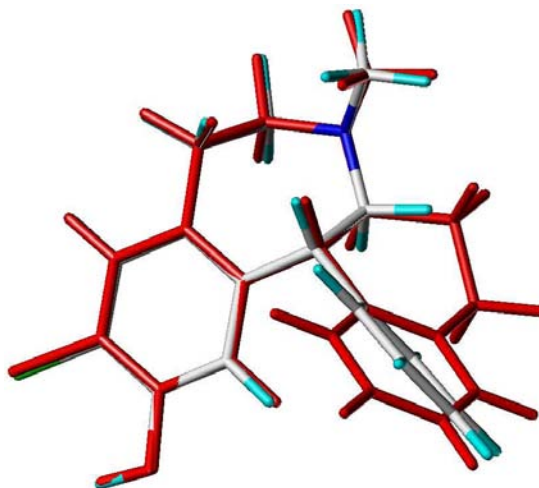


Abbildung 3.24: Überlagerung der Minimum-Energie-Konformationen von (-)-2b-SCH 39166 (rot) und (R)-(+)-SCH 23390

Das aktivste Isomer ist (-)-2b-SCH 39166 mit einem  $K_i$ -Wert von  $1,9 \text{ nmol} \cdot \text{l}^{-1}$ . Es ist an der direkten Verknüpfung des Phenylrings mit dem Azepinring wie (R)-(+)-SCH 23390 R-konfiguriert. Eine Konformationsanalyse (s. Anhang C, Seite 140) ergab als globales Minimum Konformer R1670000 bei lediglich drei Konformationsclustern.

Ein Vergleich von Konformer R1670000 (von (-)-2b-SCH 39166) mit dem globalen Minimumkonformer von (R)-(+)-SCH 23390 zeigt eine gute Übereinstimmung (s. Abbildung 3.24). Mit dem globalen Minimumkonformer von (+)-2a-SCH 39166 ergibt sich ebenfalls eine gute Überlagerung. Dem aktivsten Isomer (-)-2b-SCH 39166 wurde schließlich der Vorzug gegeben. Somit erschienen diese Konformationen der beiden hochaktiven Verbindungen als geeignete Vorlage für die Pharmakophorüberlagerung. Abbildung 3.25 zeigt die Positionen der allen Dopaminantagonisten gemeinsamen essentiellen Pharmakophormerkmale (Features): die beiden aromatischen Reste (grün) und das protonierbare Stickstoffatom (blau). Als zusätzliches optionales Merkmal diente die Donor/Akzeptor-Funktion (rot) der OH-Gruppe von (-)-2b-SCH 39166 bzw. (R)-(+)-SCH 23390.

Hierauf wurden die Repräsentanten der Konformationscluster aus den Konformationsanalysen der übrigen Dopamin  $D_1$ -Antagonisten (s. Anhang C) überlagert. Die

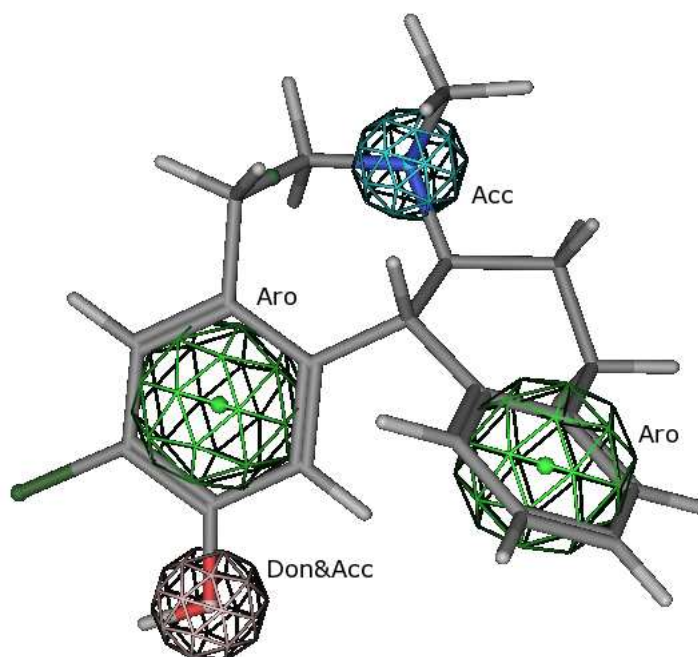


Abbildung 3.25: Die für die Überlagerung verwendeten Pharmakophormerkmale: zwei aromatische Reste (grün), protonierbares Stickstoffatom (blau), Donor/Akzeptor-Funktion (rot, optional)

Pharmakophorüberlagerung wurde mit dem Programm MOE [85] durchgeführt, welches als Ergebnis mehr als 400 Überlagerungen präsentierte, da mehrere Überlagerungen pro Konformation möglich sind. Leider wird bei der Berechnung des Überlagerungs-RMS-Wertes nur der Mittelpunkt der Pharmakophormerkmale herangezogen. Da die aromatischen Reste ebenfalls als Sphären ausgeführt sind, kommt es bei senkrechter Überlagerung der aromatischen Ringe zu kleinen RMS-Werten. Damit wird der berechnete RMS-Wert als Gütekriterium für die Überlagerung unbrauchbar, weshalb eine manuelle Auswahl durchgeführt wurde.

Es wurden die subjektiv optisch besten Überlagerungen (bis zu 8 pro Verbindung) ausgesucht. Unter den Konformeren musste das globale Minimum der jeweiligen Verbindung vertreten sein. Für die Verbindung LE 401 waren zu diesem Zeitpunkt noch keine biologischen Daten verfügbar, und sie war somit im initialen Alignment (s. Abbildung 3.26) nicht vertreten. Für die restlichen 11 Antagonisten wurden 44 Konformationen und Überlagerungen ausgesucht, die sich wie folgt auf die Verbindungen verteilten: AHAD11: 5, B157 (R/S): 8, LE 400: 3, LE 403: 3, LE 404: 3,

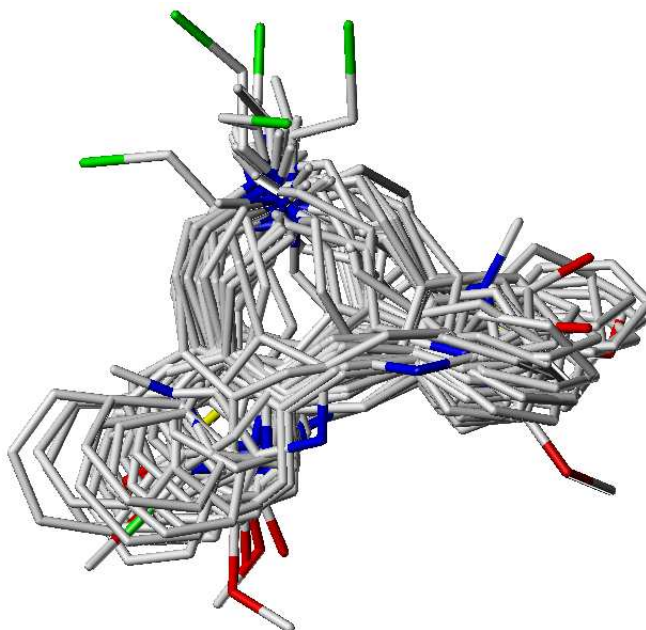


Abbildung 3.26: Überlagerung aller 44 anfänglichen Konformationen

LE 410: 3, LE 420: 5, LE 300: 5, LERU 301: 4, (R)-(+)-SCH 23390: 1, SH 3: 3. Die Kombination aller möglichen Alignments (bzw. Konformationen) ergab 9720000 Permutationen. Diese Alignmentkombinationen wurden der CoMFA mit automatischer PLS zugeführt.

### 3.4.3 Initiale CoMFA-Modelle

Die anfänglich ausgesuchten Konformationen und Überlagerungen wurden mit der Auto-PLS-Methode untersucht. Es kam das kombinierte sterische und elektrostatische Feld zum Einsatz. Die Berechnung des  $q^2$ -Wertes erfolgte, wie im weiteren Verlauf immer, mittels Leave-One-Out-Kreuzvalidierung (LOO-CV). Das beste CoMFA-Modell besaß einen  $q^2$ -Wert von 0,65 bei einer Komponentenzahl von  $NoC = 3$ . Die MATLAB-Auswertung ergab jedoch eine sehr schlechte Verteilung der  $q^2$ -Werte, welche im Mittel bei -0,5 lag (s. Abbildung 3.27). Betrachtet man das initiale Alignment (s. Abbildung 3.26), so fällt auf, dass die Ausrichtung der Substituenten nicht einheitlich ist.



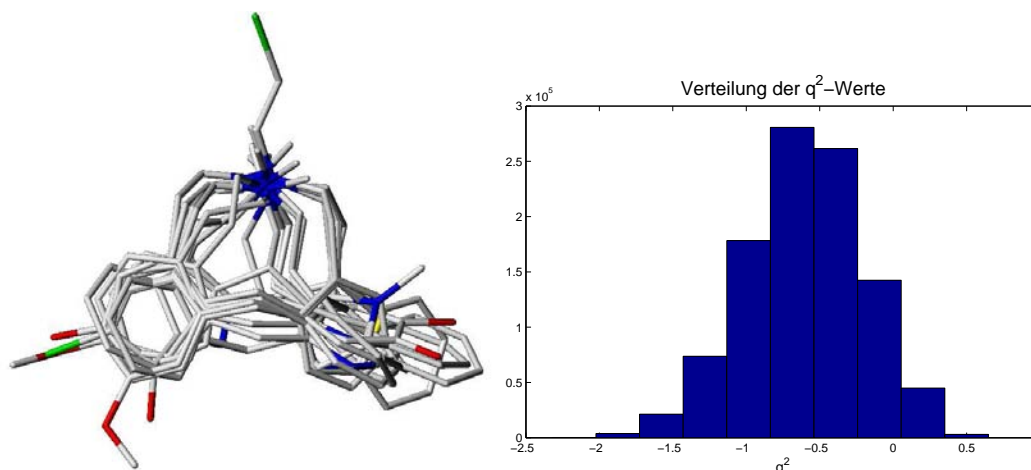


Abbildung 3.27: Bestes CoMFA-Modell des anfänglichen pharmakophorbasierten Alignments (links) und die Verteilung der  $q^2$ -Werte aller CoMFA-Modelle des anfänglichen Alignments (rechts)

### 3.4.4 Optimierung der CoMFA-Modelle

Ziel der automatischen PLS ist es nicht, das einzige zufällig gute Modell aus tausenden sehr schlechten herauszupicken. Vielmehr geht es darum, eine Schar guter Modelle zu finden und das Verteilungsmaximum zu höheren  $q^2$ -Werten zu verschieben. Durch Analyse der Auswahlhäufigkeit und der Residuen von Konformationen einer Verbindung können die Vertreter, die vermehrt fehlerhaft vorhergesagt werden, eliminiert werden. Im vorliegenden Fall war jedoch das Grundalignment ungeeignet, weshalb dieses zunächst überarbeitet werden musste.

#### Sterische Analyse und Verbesserungsmöglichkeiten

Als Ausgangspunkt für weitere Verbesserungen wurde das Alignment des ersten Modells gewählt. Bei genauer Betrachtung der Geometrie fiel auf, dass innerhalb dieses Modells eine Gruppierung der Konformationen vorliegt. In jeder Gruppe sind die Verbindungen untereinander einheitlich überlagert (für die Gruppeneinteilung s. Tabelle 3.16). Diese Gruppierung wurde beibehalten. Innerhalb der zweiten Gruppe wurden Konformations- und Alignmentanpassungen vorgenommen. Die Verbindungen dieser Gruppe wurden mehr an die Konformation von LE 404 angepasst. War einer der aromatischen Ringe substituiert (wie bei LE 404), so wurde ebenfalls die seitenverkehrte Konformation (Substitution am anderen Ring) mit aufgenommen. Gruppe 1

wurde durch die Verbindung LE 401 ergänzt, nachdem für LE 401 ebenfalls eine Konformationsanalyse und Pharmakophorüberlagerung durchgeführt worden war.

Gruppe	Verbindungen				
1	AHAD11	LE 400	LE 420	<i>LE 401</i>	
2	LE 404	LE 403	LE 410	LE 300	LERU 301
3	B 157				
4	(R)-(+)-SCH 23390				
5	SH 3				

Tabelle 3.16: Gruppeneinteilung des ersten Modells. Verbindung LE 401 wurde im Zuge der Verbesserung dieses Modells aufgenommen.

### Sukzessive Optimierung der CoMFA-Modelle

Mit diesem verbesserten Grundalignment konnten die Modelle mit der automatischen PLS optimiert werden. Die Anzahl der betrachteten CoMFA- und CoMSIA-Felder wurde erweitert. Zusätzlich zu den sterischen und elektrostatischen Feldern wurden auch Wasserstoffbrücken-, Donor-/Akzeptor- und hydrophobe Eigenschaften sowie das HINT-Feld [86] eingesetzt. Diese waren jedoch den kombinierten sterischen und elektrostatischen Feldern von CoMFA bzw. CoMSIA unterlegen.

Die Modelle wurden durch wiederholte Anwendung der automatischen PLS verbessert, wobei versucht wurde, das Overfitting zu vermeiden. Um dies sicherzustellen wurden Stabilitätstests durchgeführt. Die finalen Modelle sind mit ihren  $q^2$ -Werten und Vorhersagefehlern in Tabelle 3.17 dargestellt. Das Alignment dieser Modelle zeigt Abbildung 3.28.

Die Anwendung der Auto-PLS Methode für die Selektion von alternativen Alignments ist also in diesem Beispiel erfolgreich durchgeführt worden.

### Stabilität der CoMFA-Modelle

Die Stabilität dieser Modelle wurde mit der Random-Groups-PLS-Methode (s. Kapitel 3.2.3, Seite 46) überprüft. Es erfolgte eine Kreuzvalidierung mit fünf zufällig ausgewählten Gruppen und 100 Wiederholungen für jedes Modell. Abbildung 3.29 zeigt

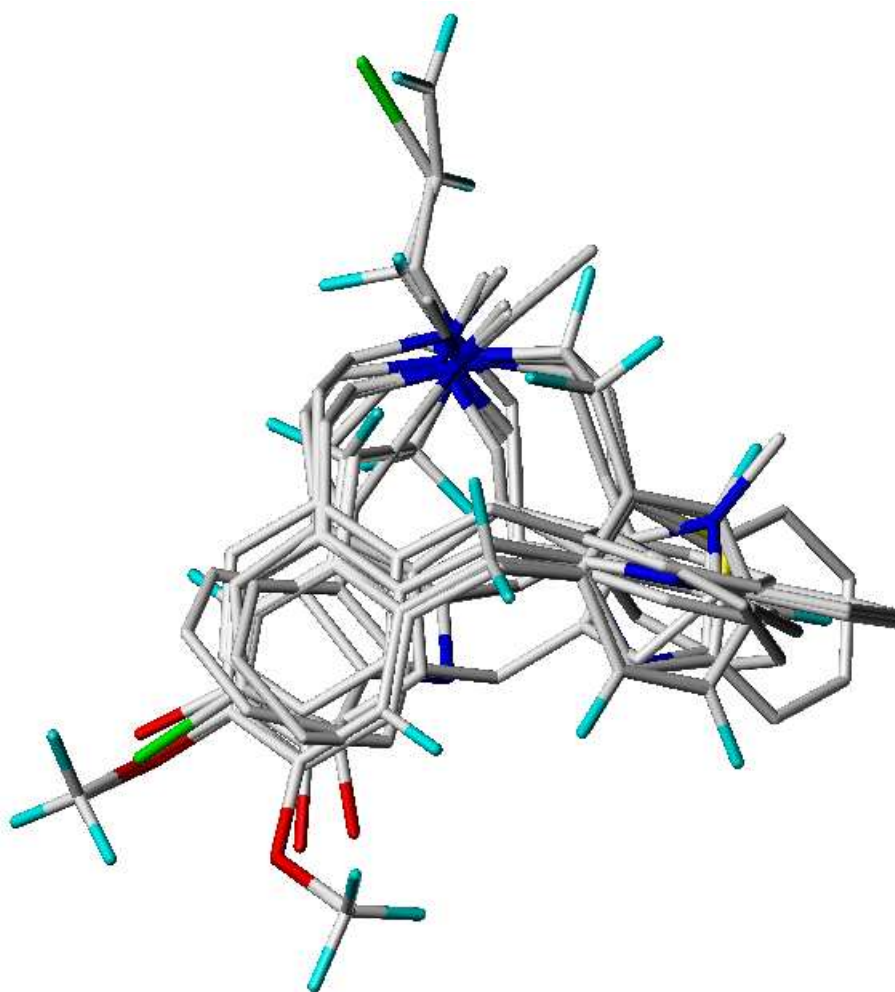


Abbildung 3.28: Alignment der finalen CoMFA- und CoMSIA-Modelle

Feld	Methode	SDEP	$q^2$
ster + ele	CoMFA	0,601	0,821
ster + ele	CoMSIA	0,495	0,879
hydrophob	CoMSIA	0,736	0,731

Tabelle 3.17: Ergebnisse der CoMF- und CoMSI-Analysen der finalen Modelle; Mindestschwankungsbreite (CF): 0,75 kcal; optimale Komponentenzahl: 3

die Verteilung der Vorhersagekraft und Tabelle 3.18 die durchschnittlichen  $q^2$ -Werte mit ihrer Standardabweichung. Daraus wird ersichtlich, dass es keinen „Einbruch“ bei der Vorhersagekraft gibt, was angesichts der kleinen Zahl der Verbindungen erstaunlich ist.

Die Random-Groups-PLS kann die Validierung durch einen externen Testdatensatz nicht ersetzen. Die kleine Anzahl an Verbindungen mit verlässlichen Daten machte in diesem Fall jedoch die Aufstellung einer solchen Testgruppe unmöglich. Eine weitere Möglichkeit wäre die Benutzung der hier aufgestellten QSAR-Modelle zur Vorhersage der Aktivität von weiteren Dopamin  $D_1$ -Antagonisten, für die nur Daten von M. Decker vorhanden sind. Da diese Daten mit einem anderen Zell- und Puffersystem erhoben wurden, müssten sie vorher allerdings mit Hilfe einer Regressionsgleichung umgerechnet werden (s. Kapitel 2.2.2, Abbildung 2.4, Seite 19).

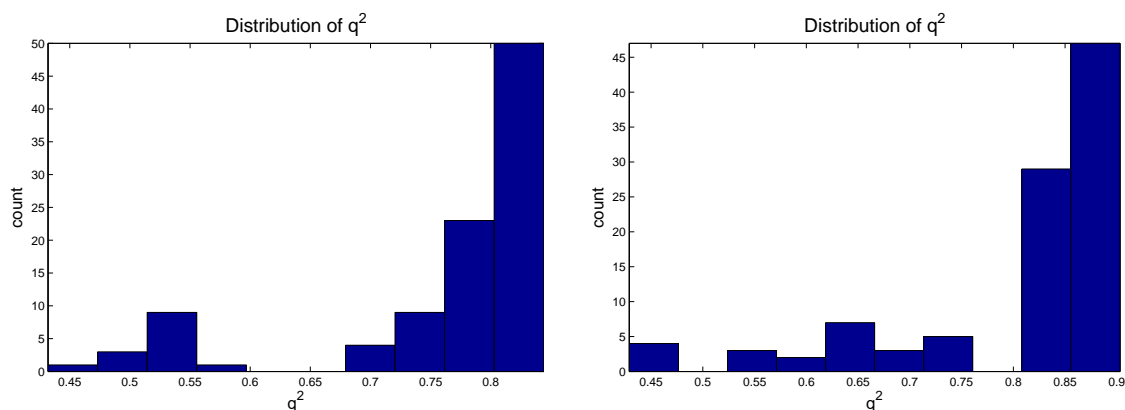


Abbildung 3.29:  $Q^2$ -Wertverteilung nach einem Stabilitätstest der endgültigen CoMFA- (links) und CoMSIA-Modelle (rechts) für das kombinierte (sterisch + elektrostatisch) Feld

Feld	Methode	$\bar{q}^2$	SD
ster + ele	CoMFA	0,757	0,101
ster + ele	CoMSIA	0,805	0,115
hydrophob	CoMSIA	0,618	0,187

Tabelle 3.18: Durchschnittliche  $q^2$ -Werte und Standardabweichung (SD) nach der Random-Groups-PLS.

## 4. COSMO

Das Ziel bei der Definition von Moleküldeskriptoren ist es, Moleküleigenschaften quantitativ zu beschreiben. Häufig verwendete Deskriptoren sind deshalb solche, die sterische Eigenschaften (Molekülausdehnung, Volumen, Moleküloberfläche), elektrostatische Eigenschaften (Ladungen) oder lipophile Eigenschaften (Löslichkeit, Verteilungskoeffizienten) beschreiben. Letztere lassen sich zu einem großen Teil auf elektrostatische Eigenschaften zurückführen.

Eine Möglichkeit, eine umfassende quantitative Beschreibung von einem Molekül zu erhalten, ist die Generierung eines Molekülfeldes für die jeweilige untersuchte Eigenschaft. Molekülfelder haben jedoch einen großen Nachteil. Will man sie für die verschiedenen Moleküle vergleichen, wie es die CoMFA-Methode macht (s. Kapitel 3.2 ab Seite 34), so müssen Ort und Orientierung im Feldgitter gleich sein. Die Verwendung von alignmentunabhängigen Deskriptoren umgeht dieses Problem. Jedoch ist der Informationsgehalt der Molekülfelder mit den bekannten Deskriptoren schwer zu erreichen. Die Vielzahl der verfügbaren Deskriptoren erschwert die Auswahl zusätzlich; die Dragon Software [87] bietet z. B. mehr als 1500 davon.

Eine verschiebungs- und rotationsunabhängige sterische bzw. elektrostatische Beschreibung von Molekülen gelingt beispielsweise mit Hilfe der DiP- und MaP-Deskriptoren (**D**istance **P**rofiles bzw. **M**apping **P**roperty distributions of molecular

surfaces) [88–91]. Die aus dem COSMO-Solvatationsmodell von A. Klamt [92] berechneten Sigma-Profile sind ein weiteres Beispiel für eine alignmentunabhängige und umfassende Beschreibung der elektrostatischen Eigenschaften eines Moleküls und deshalb ebenfalls vielfältig einsetzbar. Aus ihnen lassen sich physikochemische Eigenschaften wie Löslichkeit,  $pK_a$ -Werte,  $\log P$ -Werte und viele andere ableiten. Diese Teildisziplin der QSAR nennt sich QSPR – Quantitative Structure Property Relationship. Die Verwendung als QSAR-Deskriptoren liegt daher sehr nahe, ist aber bis jetzt eher selten untersucht worden. Aus diesem Grund geht diese Arbeit darauf ein.

## 4.1 Das COSMO-Solvatationsmodell

Noch heute werden viele Berechnungen zu Moleküleigenschaften in der Gasphase bzw. im Vakuum durchgeführt. Das Molekül wird dabei einzeln für sich betrachtet. Dieses Vorgehen ignoriert jedoch die Tatsache, dass viele Moleküleigenschaften von ihrer Umgebung beeinflusst werden, wenn die Moleküle in kondensierter Phase (d. h. meistens flüssig bzw. gelöst) vorliegen. Da das der wichtigste Fall in der medizinischen Chemie ist, kommt den Solvatationsmodellen eine große Bedeutung zu.

### 4.1.1 Problem der Berücksichtigung der Solvation

Regelmäßig kommen Solvatationsmodelle zum Einsatz, wenn versucht wird, makroskopische Eigenschaften von Stoffen wie Löslichkeit, Verteilungskoeffizienten,  $pK_a$ -Wert usw. mittels Computerberechnungen vorherzusagen, die auf den physikalisch-chemischen Grundprinzipien<sup>1</sup> beruhen. Trotzdem beruhen diese Methoden fast ausschließlich auf Einzelmolekülberechnungen, während für makroskopische Eigenschaften die Thermodynamik von Molekülensembeln eine sehr große Rolle spielt.

Moleküldynamiksimulationen oder Monte-Carlo-Methoden scheinen für diese Aufgabe besser geeignet zu sein, müssen aber aufgrund der großen Anzahl an Molekülen

---

<sup>1</sup>Die sehr häufig verwendeten Inkrementmethoden wie CLogP usw.. gehören nicht dazu.

bzw. Atomen die Berechnungen stark vereinfachen. Es kommen hierbei oft nur Kraftfeldmethoden zum Einsatz, deren Genauigkeit meist nicht ausreicht, um die oben erwähnten Eigenschaften vorherzusagen. Quantenmechanische Moleküldynamikmethoden sind bereits seit langem bekannt, konnten sich aber aufgrund der enormen Anforderungen an die Rechenkapazität noch nicht etablieren.

### 4.1.2 Solvatationsmodelle

Einen Kompromiss zwischen gewünschter Genauigkeit und erforderlicher Rechenkraft stellen Kontinuum-Solvatationsmodelle (CSM) wie das „Polarizable Continuum Model“ dar, die bei quantenmechanischen Einzelmolekülberechnungen eingesetzt werden. Dabei wird die Umgebung des Moleküls nicht durch weitere Einzelmoleküle, sondern durch ein dielektrisches Kontinuum repräsentiert.

Dieses Dielektrikum schirmt einerseits das Molekül ab und wird andererseits entsprechend seiner Dielektrizitätskonstante durch das elektrische Feld des Moleküls teilweise polarisiert. Das Molekül selbst wird wiederum von dem elektrischen Feld, welches von den lokalen Ladungen des polarisierten Dielektrikums ausgeht polarisiert.

Dabei werden an den Schnittstellenbereichen (zwischen der Oberfläche des Moleküls und dem angrenzenden dielektrischen Kontinuum) wechselseitig Abschirmungsladungen induziert. Die Abschirmungsladungen spiegeln den elektrostatischen Charakter des Moleküls in Lösung besser wider als die aus reinen Gasphasenberechnungen gewonnenen elektrostatischen Potenziale (ESP). Diese wechselseitige Polarisation ist bei vielen PCM-Modellen durch die Verwendung eines „Self Consistent Reaction Field“ (SCRF) berücksichtigt und erfordert einen entsprechend hohen zusätzlichen rechnerischen Aufwand.

1993 veröffentlichten Klamt und Schüürmann [92] das „Conductor like Screening Model“ (COSMO), welches eine vereinfachte aber nicht weniger genaue Berücksichtigung von Solvationseffekten ermöglicht. Die Vereinfachung besteht in der Betrachtung des das Molekül umgebenden Kontinuums als Leiter (Conductor) und nicht

als Dielektrikum. Die Abschirmungsladungen werden also gegen einen Leiter berechnet. Da dieser nicht polarisierbar ist, entfällt die aufwändige Berücksichtigung der wechselseitigen Polarisierung durch ein SCRF. Trotzdem werden die dielektrischen Eigenschaften der Umgebung durch eine entsprechende Skalierung über die Dielektrizitätskonstante mit einbezogen.

Im Prinzip ist dies der gegensätzliche Ansatz zu den Polarized Continuum Modellen, bei denen die Eigenschaften eines Leiters (Permittivität) auf einen Nichtleiter (totales Dielektrikum mit  $\varepsilon = 1$ ) übertragen werden. Dementsprechend gut funktioniert das COSMO-Modell auch bei Lösungsmitteln mit hoher Dielektrizitätskonstante wie z. B. Wasser ( $\varepsilon \approx 80$ ), welche mehr einem Leiter als einem Nichtleiter ähneln. Klamt et al. [93] konnten zeigen, dass auch bei Lösungsmitteln mit kleiner Dielektrizitätskonstante der Fehler sehr klein ist (kleiner als 10 Prozent).

Durch die Verknüpfung der COSMO-Daten mit Methoden der statistischen Thermodynamik gelingt schließlich auch die Vorhersage der bereits zuvor erwähnten makroskopischen Eigenschaften (Verteilungskoeffizienten, Löslichkeiten, Dissoziationskonstanten u.v.m.) [94].

### 4.1.3 COSMO-Daten als Deskriptoren für die QSAR

Die elektrostatischen Eigenschaften sind für eine Reihe von chemisch-physikalischen Merkmalen verantwortlich, wozu z. B. der Wasserstoffbrückenbindungscharakter und die Lipophilie bzw. Hydrophilie zählen. Selbst van-der-Waals-Kräfte (Londonsche Dispersionskräfte) werden mit Ungleichgewichten in der Ladungsverteilung erklärt. Fast alle Arten intermolekularer Wechselwirkungen lassen sich auf elektrostatische Einflüsse zurückführen.

Die über das COSMO-Solvatationsmodell berechneten Abschirmungsladungen spiegeln das elektrostatische Potenzial eines Moleküls gut wider. Also ist es sinnvoll, diese als Moleküldeskriptor — unter anderem auch für die QSAR — zu verwenden. Die genannten Abschirmungsladungen werden auf einer dreidimensionalen Moleküloberfläche berechnet und enthalten somit auch räumliche Informationen. Damit eignen



sie sich sogar für die 3D-QSAR. Allerdings ist die direkte Verwendung dieser Ladungen mit ihren Koordinaten für die 3D-QSAR nicht möglich. Da die Ladungen auf der Moleküloberfläche lokalisiert sind, kann man die Moleküle so auch nicht direkt vergleichen. Man müsste aus den Ladungen zunächst elektrostatische Felder ableiten, was wieder zum Problem der Alignmentabhängigkeit führen würde.

Aus diesen Gründen müssen die elektrostatischen von den sterischen Informationen abstrahiert werden, was durch die Erstellung einer Wahrscheinlichkeitsdichtefunktion der Ladungsdichte erfolgt. Diese Funktion wurde von A. Klamt „COSMO-Sigma-Profil“ genannt. Sie wird auch in den Berechnungen von COSMO-RS [95–97] zur Eigenschaftsvorhersage verwendet. Die COSMO-Sigma-Profile sind grundsätzlich unabhängig von der Molekülgröße und der Raumposition des Moleküls, was einen großen Vorteil für die QSAR darstellt.

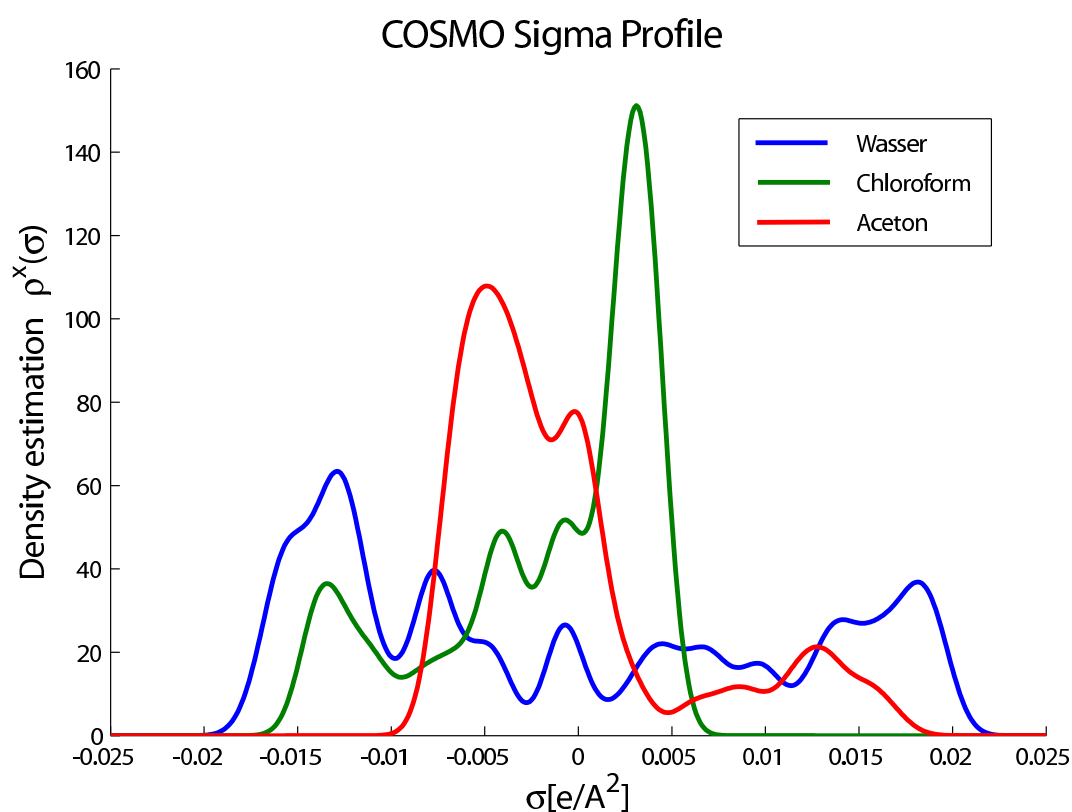


Abbildung 4.1: COSMO-Sigma-Profile der Verbindungen Wasser (blau), Chloroform (grün) und Aceton (rot)

Abbildung 4.1 zeigt exemplarisch die COSMO-Sigma-Profile von Wasser, Chloroform und Aceton. Dargestellt ist auf der Abszisse die Ladungsdichte in Elementarladungen pro  $\text{\AA}^2$  und auf der Ordinate die zugehörige Wahrscheinlichkeitsdichte. Klamt et al. konnten mehrfach zeigen, dass sich diese Profile zur quantitativen Struktureigenschaftsvorhersage (QSPR) eignen [98–103], die Verwendung für die QSAR ist dagegen bisher kaum dokumentiert. Der verwendete Algorithmus weicht im Detail etwas von dem Klamts ab, so dass die hier gezeigten und auch später verwendeten Sigma-Profile nicht ganz identisch mit denen des Originalautors sind. Auf die Verwendung als QSAR-Deskriptor dürfte dies aber kaum Auswirkungen haben.

## 4.2 Berechnung der COSMO-Sigma-Profile

Wie im vorigen Abschnitt bereits erläutert wurde, sind die Sigma-Profile eine Wahrscheinlichkeitsdichtefunktion der Ladungsdichten der COSMO-Abschirmungsladungen  $\sigma$ . Die Wahrscheinlichkeitsdichte ist die geschätzte Häufigkeit des Vorkommens eines bestimmten Wertes in einer Ansammlung von Zahlen. Die Zahlen stellen in diesem Fall die Ladungsdichte  $\sigma$  dar. Die Häufigkeit wird als geschätzt bezeichnet, da es sich bei der Gesamtheit der  $\sigma$ -Werte einer Verbindung lediglich um eine Stichprobe handelt.

Man kann mit verschiedenen Methoden die Wahrscheinlichkeitsdichtefunktion der wahren Verteilungsfunktion annähern, was umso besser funktioniert, je größer die Stichprobe ist. Die Stichprobengröße — also die Zahl der  $\sigma$ -Werte — hängt direkt von der Größe der Oberfläche des Moleküls und damit von der Molekülgröße ab, da die Größe der Segmente annähernd gleich ist. Je größer das Molekül ist, desto besser kann man somit die wahre Verteilungsfunktion approximieren. Das bedeutet aber auch, dass die ermittelten Sigma-Profile von kleinen Molekülen (z. B. Wasser) recht ungenau sind. Das spiegelt sich im „unruhigen“ Kurvenverlauf wider, welcher im Fall des Wassers schlicht die geringe Anzahl von Werten zurückzuführen ist.

Es gibt mehrere Möglichkeiten, Wahrscheinlichkeitsdichten zu bestimmen. Die einfachste Variante erstellt ein Histogramm und unterteilt die Spanne zwischen dem

größten und dem kleinsten Wert in gleichgroße Abschnitte. Anschließend werden die Werte, die in den jeweiligen Abschnitt fallen, gezählt. Das ist auch die Methode, die Klamt et al. sowie Oldland [104] verwenden. Zusätzlich verteilen sie jedoch den von der Intervallmitte abweichenden Teil der Ladungsdichte auf die benachbarten Intervalle, so dass die Gesamtladung erhalten bleibt. Abbildung 4.2(a) zeigt die auf diesem Weg berechneten Sigma-Profile. Möglich ist außerdem eine Kurvenglättung mittels B-Splines [105], so dass man das Bild von Abbildung 4.2(b) erhält.

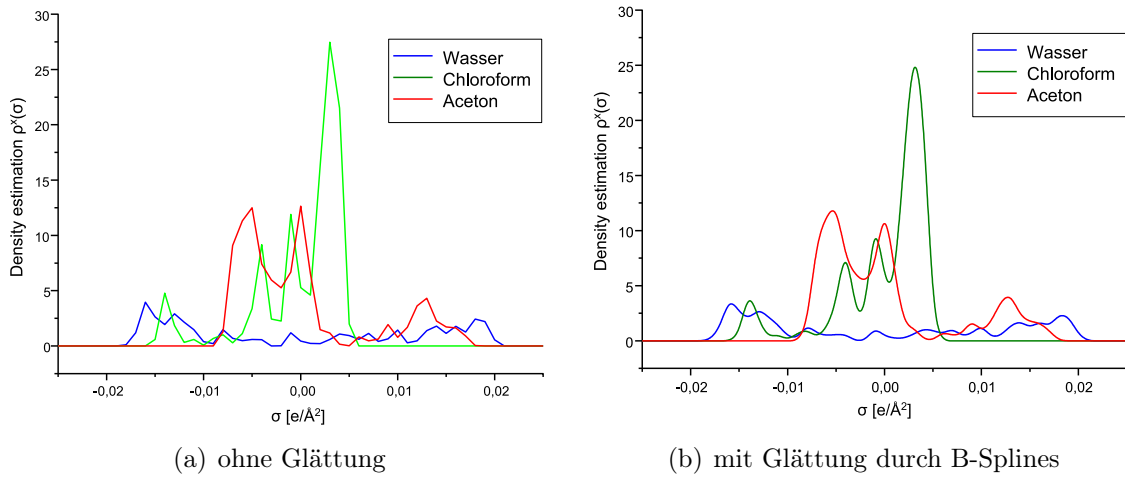


Abbildung 4.2: COSMO-Sigma-Profile erstellt mit der Histogrammmethode

Für diese Methode ist allerdings aus folgendem Grund eine hinreichend große Stichprobe von Nöten: Die Genauigkeit der resultierenden Funktion hängt von der gewählten Intervallbreite ab. Je kleiner das Intervall, desto genauer könnte man die wahre Verteilung abbilden, vorausgesetzt man hat genügend Werte zur Verfügung, um sie in die größere Anzahl von Intervallen einzusortieren. Klamt et al. benutzen ein Intervall von  $0,001 \text{ e}/\text{\AA}^2$ , was bei einer Spanne von  $-0,025 - 0,025$  zu 51 Werten führt. Bei dem häufig benutzten Beispiel Wasser liefert die COSMO-Berechnung jedoch nur 136  $\sigma$ -Werte, die dann auf die 51 Intervalle verteilt werden.

Die Parzen-Window-Methode [106] mit Gaußkernel scheint in diesem Fall geeigneter zu sein [107], da sie die einzelnen Werte innerhalb einer festgelegten Spannbreite als normalverteilt annimmt. Jeder Wert  $x_i$  wird über eine stetige Gaußkurve mit der erwähnten Spannbreite (auch Fensterbreite genannt) als Standardabweichung

(stdev) repräsentiert. Diese Gaußkurven werden schließlich zur Wahrscheinlichkeitsdichtefunktion aufsummiert. In Abbildung 4.3 sind die Einzelfunktionen mit unterbrochenen Linien und die resultierende Funktion mit einer durchgezogenen dickeren Linie dargestellt. Dort wo die Werte dichter zusammenliegen, ergibt sich eine höhere Wahrscheinlichkeitsdichte. Die Gleichungen 4.1 und 4.2 fassen das Verfahren zusammen.

$$\rho(x) = \frac{1}{N} \sum_{i=1}^N W(x - x_i) \quad (4.1)$$

$$\text{wobei} \quad W(x) = \frac{1}{\text{stdev} \sqrt{2\pi}} e^{-x^2/2\text{stdev}^2} \quad (4.2)$$

Die Genauigkeit der resultierenden Funktion hängt hier nicht von der Intervallbreite, sondern von der gewählten Fensterbreite ab. Die Güte der Näherung ist jedoch ebenso wie bei der Histogrammmethode von der Anzahl der Ausgangswerte abhängig.

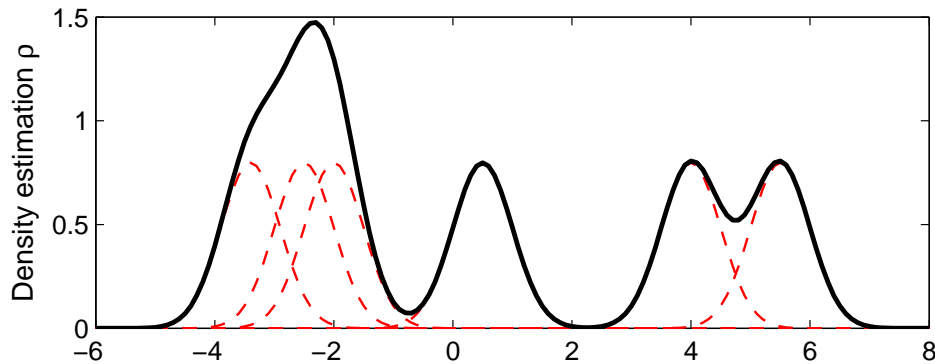


Abbildung 4.3: Die Parzen-Window-Dichte (durchgezogene dicke Linie) wird durch Aufsummierung von Gaußfunktionen (rote unterbrochene Linien) berechnet; die Werte  $x_i$  bzw.  $\mu$  betragen hier -3,4; -2,5; -2,0; 0,5; 4 und 5,5; die Fensterbreite beträgt 0,5.

## 4.3 Sigma-Profile als Moleküldeskriptoren

Eine wesentliche Anforderung an einen Moleküldeskriptor ist, dass sich mit seiner Hilfe Gemeinsamkeiten und Unterschiede zwischen Molekülen quantitativ beschreiben lassen. Das Gebiet der molekularen Ähnlichkeitsanalyse befasst sich intensiv mit Deskriptoren, mit denen sich die Ähnlichkeit bzw. Verschiedenheit von Molekülen quantitativ erfassen lässt. Solche Deskriptoren und Algorithmen zu ihrem Vergleich werden benötigt, um Substanzbibliotheken zu durchsuchen, z. B. um Verbindungen für das (Virtual) Screening zu finden. Auch die QSAR benötigt die Deskriptorwerte, um sie mit der biologischen Aktivität zu korrelieren. Die COSMO-Sigma-Profile repräsentieren die elektrostatischen Eigenschaften eines Moleküls und machen Moleküle auf dieser Basis vergleichbar.

### 4.3.1 Vergleichsverfahren

Da zum Vergleich der Ähnlichkeit eine einzelne Zahl am besten geeignet ist, muss eine Methode benutzt werden, die in Form einer Zahl ausdrückt, wieviel die Profile gemeinsam haben. Eine hierfür häufig benutzte Größe ist der Eigenwert. Das so genannte spezielle Eigenwertproblem ist die Bestimmung der nichttrivialen Lösungen der Gleichung:

$$\mathbf{A}\vec{x} = \lambda\vec{x} \quad (4.3)$$

Hierbei ist  $\mathbf{A}$  eine  $(n \times n)$  Matrix,  $\vec{x}$  der so genannte Eigenvektor mit Länge  $n$  und  $\lambda$  der Eigenwert, welcher eine skalare Größe ist. Eigenwerte charakterisieren ähnlich den Singulärwerten (aus einer Singulärwertzerlegung) Eigenschaften der Matrix  $\mathbf{A}$ . Im Fall der Sigma-Profile wäre der Eigenwert ein Maß für die Interkorrelation der Funktionswerte: Betrachtet man eine Matrix aus  $n$  Variablen ( $n$  Sigma-Profile), die nicht vollständig interkorreliert sind, so erhält man  $n$  Eigenwerte, deren Größe dem Anteil möglicher Hauptkomponenten an der Gesamtvarianz entspricht.

Ein Vorteil dabei ist, dass man mehrere Sigmaprofile miteinander vergleichen kann. Ist der erste Eigenwert sehr groß und alle weiteren Eigenwerte sehr klein, sind alle

Profile sehr ähnlich. Gibt es eine mehrstufige Gliederung der Eigenwerte, sind ein oder mehrere Profile weniger ähnlich. Man wüsste allein anhand der Eigenwerte allerdings nicht, welches Profil nicht zu den anderen passt, was man jedoch mit Hilfe einer Hauptkomponentenanalyse ermitteln könnte.

Eine andere Möglichkeit des Vergleichs zweier Funktionen ist die Berechnung ihres überlappenden Integrals — der gemeinsamen Fläche unter der Kurve (AUC). Der Anteil der gemeinsamen Fläche zweier Profile an der normierten Gesamtfläche ist der gesuchte Ähnlichkeitswert, welcher recht gut mit dem visuellen Empfinden bei Betrachtung der Profile übereinstimmt.

### 4.3.2 Vergleich von Konformationen eines Moleküls

Sieht man von flexiblen Molekülen mit sehr großen Dipolmomenten (z. B. flexible Betaine) einmal ab, dürften beim Konformationsvergleich nur sehr geringe Unterschiede zu Tage treten. Die wesentlichen funktionellen Gruppen und die Konfiguration des Moleküls sind identisch und die COSMO-Sigma-Profile enthalten keine direkten sterischen Informationen. Der Unterschied zwischen Konformationen eines Moleküls besteht jedoch zum größten Teil in der Sterik, die Elektrostatik ist nur indirekt und zu einem weitaus kleineren Teil davon betroffen.

Tabelle 4.1 zeigt die Ergebnisse des Vergleichs von Konformationen der Verbindung (R)-(+)-SCH 23390. Es sind drei repräsentative Konformationen, die bei der Konformationsanalyse der Verbindung gefunden wurden (s. auch Tabelle C.4 auf Seite 130). Die entsprechenden COSMO-Sigma-Profile sind in Abbildung 4.4 dargestellt. Die Konformation R490000 weicht sterisch von Konformation R460000 weniger ab als von Konformation R930000. Die höhere Ähnlichkeit der beiden erstgenannten Konformationen ist ebenso im Sigma-Profil zu beobachten, wobei die Gesamtcharakteristik gleich bleibt. Sterische Unterschiede haben somit kleine, aber sichtbare Auswirkungen auf das elektrostatische Sigma-Profil.

Konform. 1	Konform. 2	RMS	Eigenwert (%)	gemeins. AUC (%)
R460000	R490000	0,611	99,78	96,38
R490000	R930000	1,457	99,56	95,60

Tabelle 4.1: Vergleich der repräsentativen Konformationen von (R)-(+)-SCH 23390

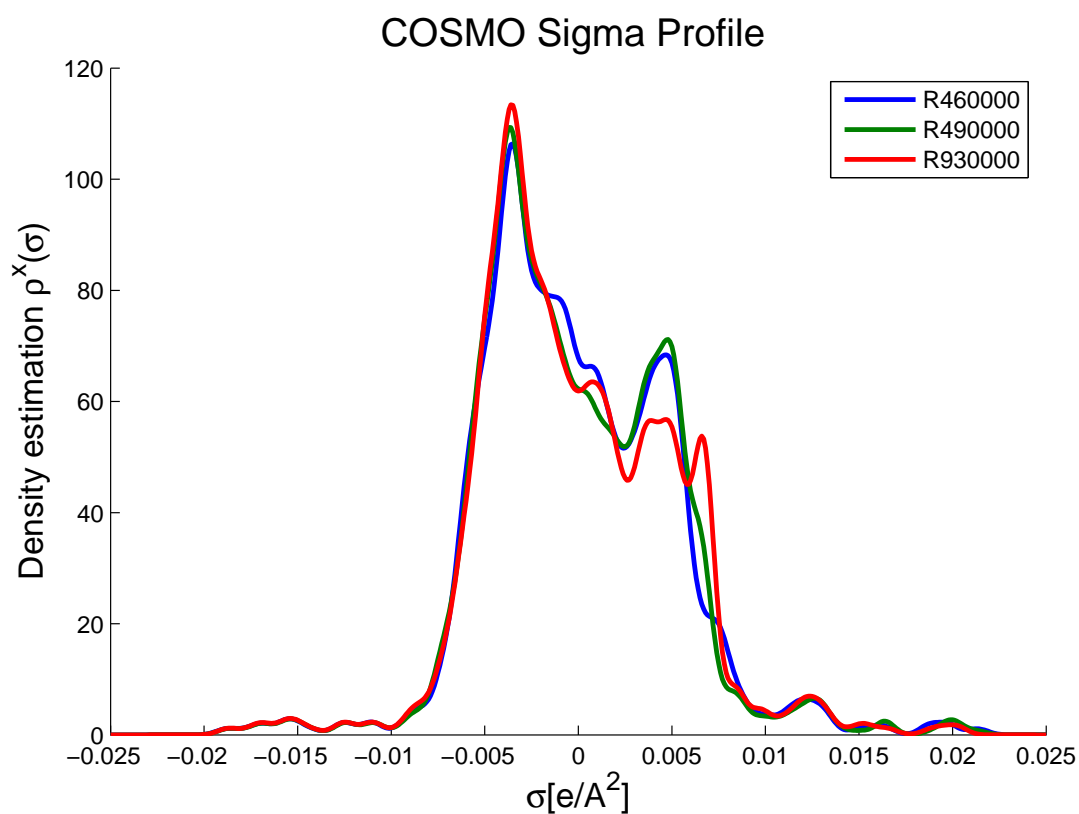


Abbildung 4.4: COSMO-Sigma-Profile der Konformationen von (R)-(+)-SCH 23390

### 4.3.3 Vergleich von verschiedenen Molekülen

Tabelle 4.2 und Abbildung 4.5 zeigen die COSMO-Sigma-Profile von unterschiedlichen Molekülen im direkten Vergleich. Während mit n-Hexan und n-Oktan sehr ähnliche Moleküle verglichen werden, stehen sich auf der anderen Seite mit n-Oktan und 1-Oktanol, Chloroform und Aceton sowie Wasser und n-Oktanol eher unterschiedliche Moleküle gegenüber. Dies ist der Gestalt der COSMO-Sigma-Profile auf Seite 91 direkt zu entnehmen.

Verbindung 1	Verbindung 2	Eigenwert (%)	gemeins. AUC (%)
Hexan	Oktan	99,99	98,82
Oktan	Oktanol	98,91	85,44
Chloroform	Aceton	66,04	43,31
Wasser	Oktanol	90,55	24,79

Tabelle 4.2: Vergleich von verschiedenen unterschiedlich ähnlichen Molekülen

Die Ähnlichkeit der ersten beiden Paare wird sowohl durch den sehr großen ersten Eigenwert als auch durch den Anteil der gemeinsamen AUC widerspiegelt. Bei den letzten beiden Paaren Chloroform/Aceton und Wasser/Oktanol ergibt sich jedoch eine Diskrepanz zwischen den beiden Ähnlichkeitsmaßen. Hier gibt der Wert der gemeinsamen AUC die Ähnlichkeit bzw. Verschiedenheit besser wieder, während für die Profile von Oktanol und Wasser ein sehr großer erster Eigenwert berechnet wurde.

Das mag daran liegen, dass der Eigenwert stärker die Form des Profils berücksichtigt und nicht die absoluten Unterschiede in den Amplituden. Das bedeutet, dass das Übereinanderliegen von Maximum und Minimum zweier Kurven den ersten Eigenwert stärker mindert als das Übereinanderliegen zweier Maxima mit sehr unterschiedlicher Höhe. Diese beiden Phänomene sind bei den Paaren Wasser/Oktanol (s. Abbildung 4.5(d)) und Chloroform/Aceton (s. Abbildung 4.5(c)) zu beobachten. Wahrscheinlich führt eine Kombination der beiden Parameter Eigenwert und



gemeinsame AUC zu einem noch robusteren Ähnlichkeitsmaß, was an dieser Stelle aber nicht weiter untersucht wurde.

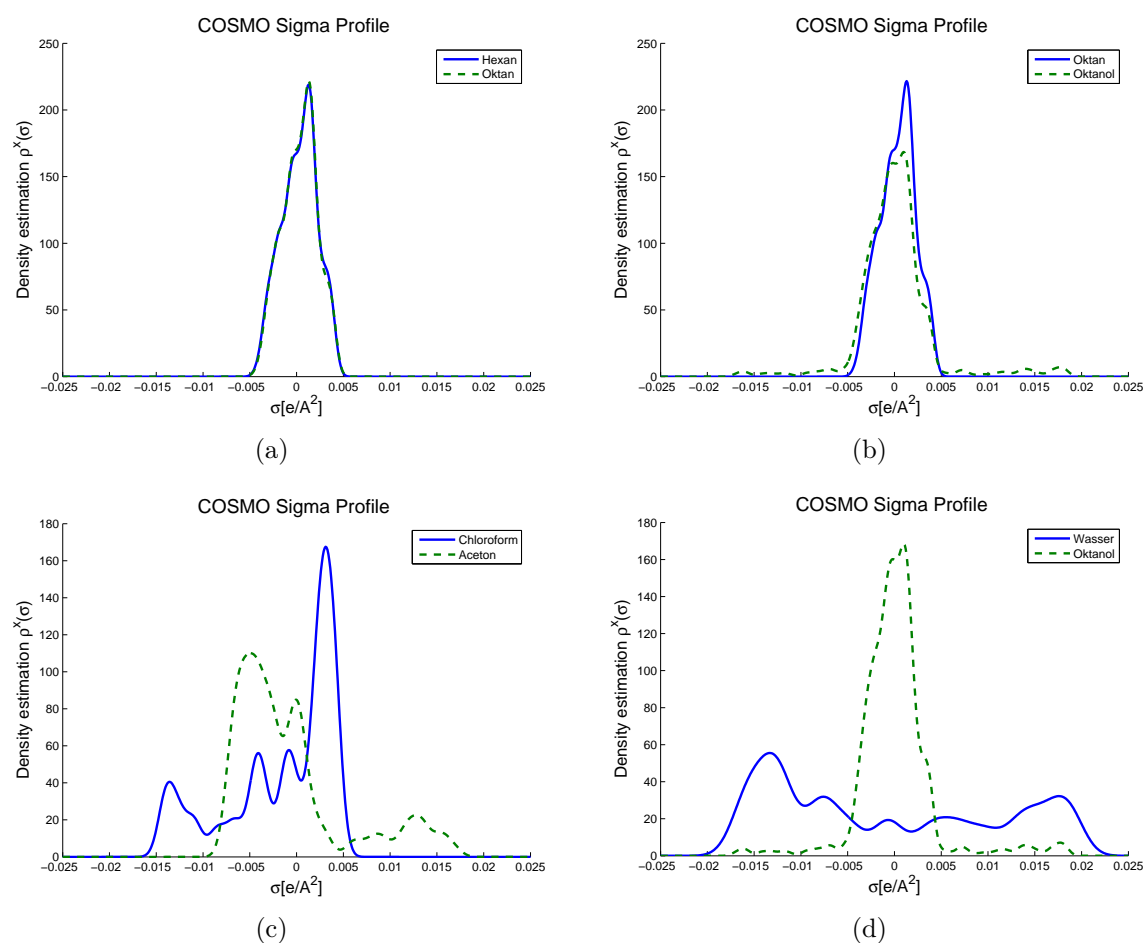


Abbildung 4.5: COSMO-Sigma-Profile einiger unterschiedlich ähnlicher Moleküle

## 4.4 QSAR-Anwendung der COSMO-Deskriptoren

Wenn mit den COSMO-Sigma-Profilen die elektrostatischen Eigenschaften von Molekülen charakterisiert werden können, sollten sie sich auch als QSAR-Deskriptoren nutzen lassen. Voraussetzung dafür ist jedoch, dass die elektrostatischen Eigenschaften einen Beitrag zur Aktivität leisten. Bei den Dopamin D<sub>1</sub>-Antagonisten, die bereits in Kapitel 3.4 vorgestellt und verwendet wurden, ist das der Fall. Im Folgenden wurde untersucht, ob ein Zusammenhang zwischen den COSMO-Sigma-Profilen und der biologischen Aktivität (der gemessenen Bindung am D<sub>1</sub>-Rezeptor) besteht.

#### 4.4.1 Verwendete Strukturen und Konformationen

Die 2D-Strukturen und die zugehörigen Radioligandbindungsdaten sind bereits in Abbildung 3.22 und Tabelle 3.15 dargestellt worden. Es wurden die aus dem Konformationsclustering (siehe Kapitel 3.1.3 und 3.4.2) erhaltenen globalen Minimumkonformationen, die im Anhang C abgebildet sind, verwendet.

Die für die Berechnung der COSMO-Sigma-Profile benötigten COSMO-Abschirmungsladungen wurden mit dem Programm TURBOMOLE [108] berechnet. Hierfür kam das Dichtefunktional BP-TZVP [109–112] mit einem Standardgitter zum Einsatz. Die COSMO-Parameter entsprachen ebenfalls den Programmvorgaben, mit Ausnahme der COSMO-Radien: hier wurden die optimierten Radien (siehe [113]) eingestellt. Mit diesen Einstellungen wurden alle Konformationen mit TURBOMOLE erneut optimiert. Die bei Erreichen der Konvergenz erhaltene COSMO-Datei enthält die Ladungsdichten der Abschirmungsladungen.

Mit dem Programm `cosmo_anA` (siehe Anhang D.7) wurden aus diesen Ladungsdichten die COSMO-Sigma-Profile mit der Parzen-Window-Methode (s. Gleichung 4.2) berechnet. Die Fensterbreite wurde für alle Verbindungen auf  $0,0005 \text{ e}/\text{\AA}^2$  festgelegt. Das Intervall betrug  $[-0,025 \dots 0,025]$  mit einer Unterteilung von  $0,0001 \text{ e}/\text{\AA}^2$ , so dass ein Profil aus 501 Werten bestand. Die berechneten COSMO-Sigma-Profile sind in Abbildung 4.6 dargestellt.

Da bei Verbindung B157 zwei Enantiomere möglich sind, wurde untersucht, ob sich die COSMO-Sigma-Profile der beiden Enantiomere unterscheiden. Sie wurden sowohl miteinander als auch mit den Profilen der anderen Verbindungen verglichen. Die Eigenwertidentität der Profile der Enantiomere betrug 99,98 %, die AUC-Übereinstimmung betrug 98,79 %, was durch kleine konformationelle Unterschiede der Enantiomere bedingt ist. Demnach zeigten die Profile eine sehr große Ähnlichkeit. hundertprozentig In der Tabelle 4.3 ist das Ergebnis des Vergleichs mit den übrigen Verbindungen abgebildet.

Eine lineare Regression der Enantiomere mit Fit durch den Koordinatenursprung ergab einen  $r^2$ -Wert von 0,999 mit einem Anstieg von 1,0 (Eigenwertmethode) bzw. einen  $r^2$ -Wert von 0,998 mit einem Anstieg von 1,0 (AUC-Methode). Demzufolge ist es ohne Bedeutung, welches Enantiomer für die QSAR mit den COSMO-Daten verwendet wird. In den folgenden Berechnungen wurde das S-Enantiomer von B 157 verwendet.

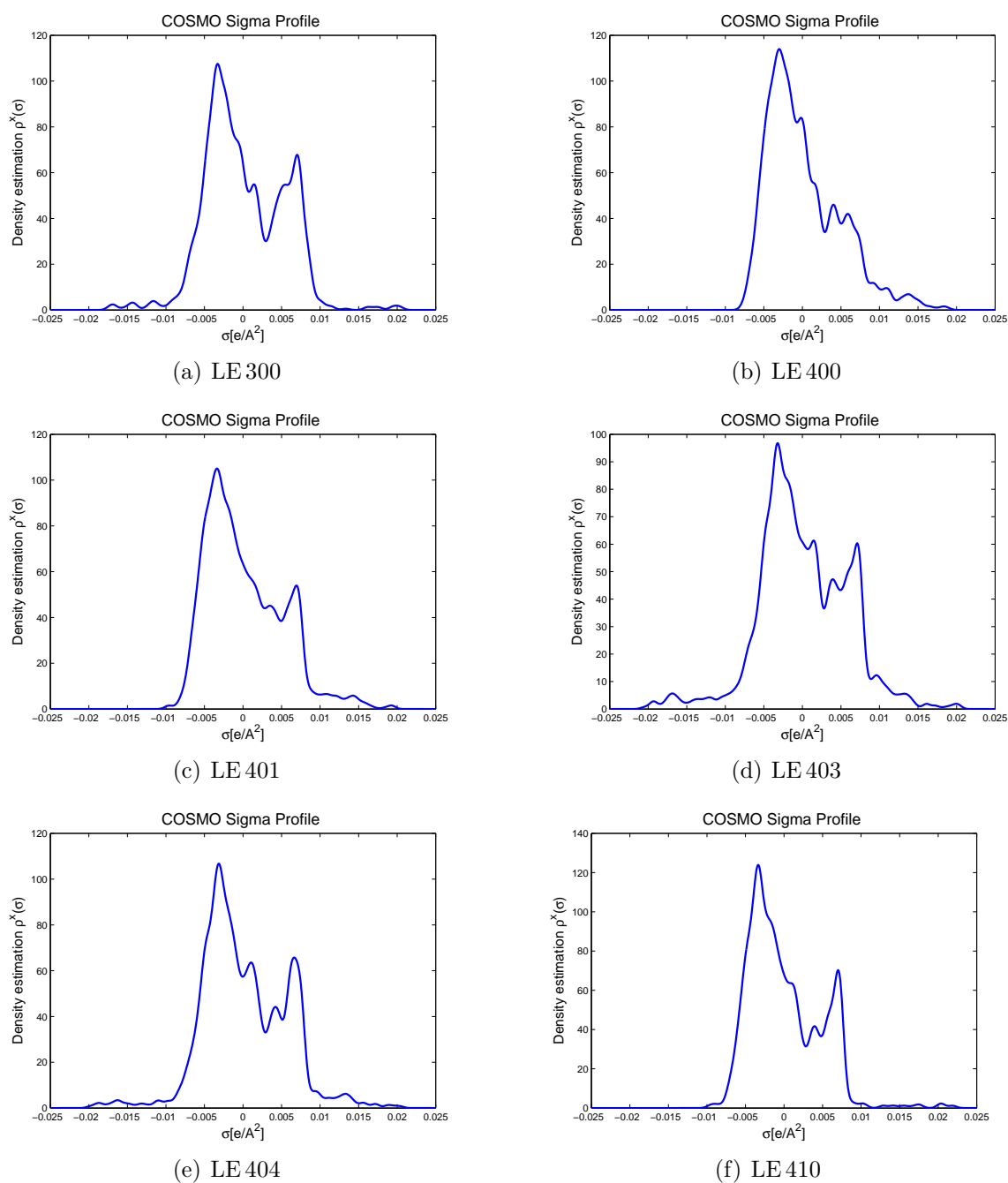
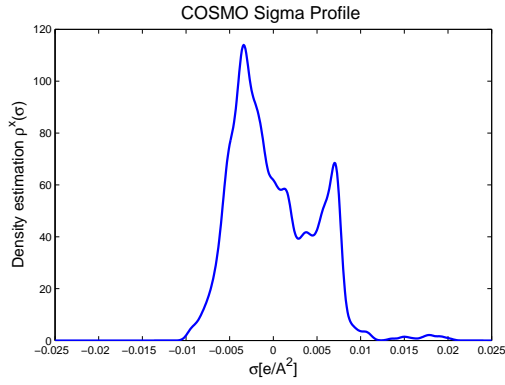
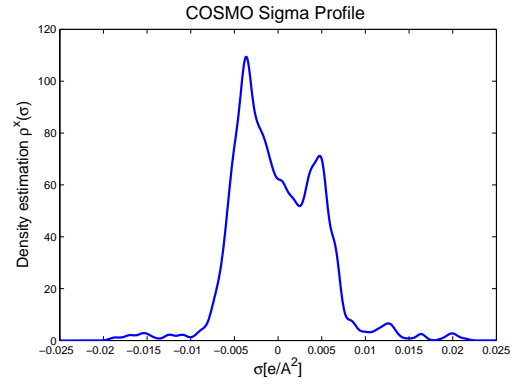


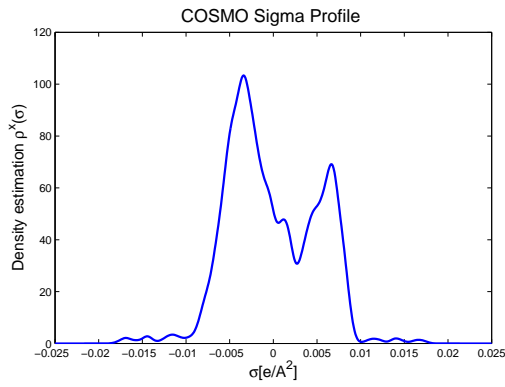
Abbildung 4.6a: COSMO-Sigma-Profile der Dopamin  $D_1$ -Antagonisten



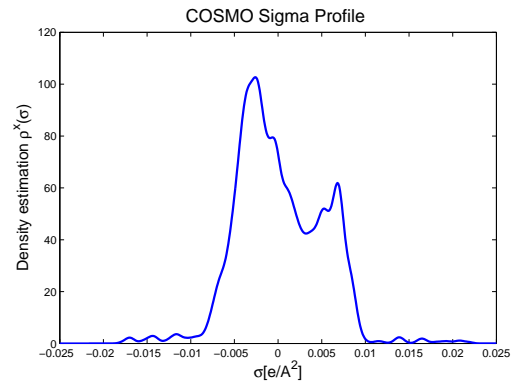
(a) LE 420



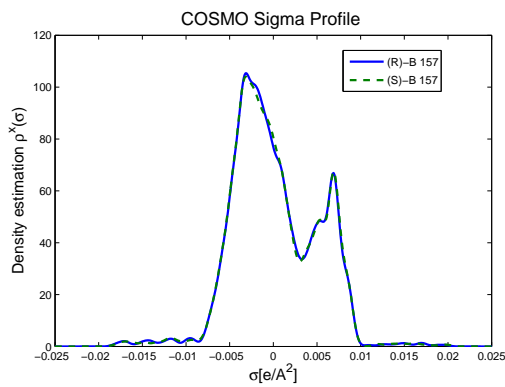
(b) (R)-(+)-SCH 23390



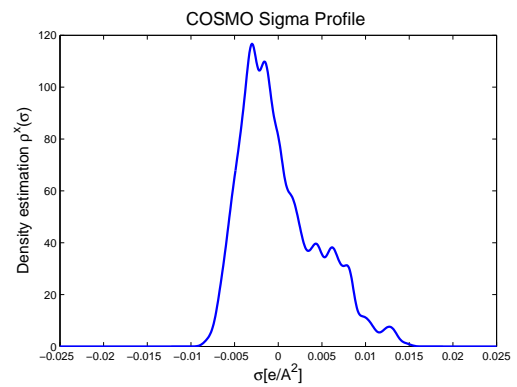
(c) AHAD11



(d) LERU 301



(e) B 157



(f) SH 3

Abbildung 4.6b: COSMO-Sigma-Profile der Dopamin  $D_1$ -Antagonisten

Enantiomer	(R)-B 157		(S)-B 157	
Verbindung	Eigenwert	AUC	Eigenwert	AUC
AHA D11	97,86	88,22	97,87	88,50
LE 400	98,77	89,77	98,75	89,70
LE 401	98,62	88,69	98,59	88,71
LE 403	99,35	90,43	99,33	90,39
LE 404	99,24	91,99	99,20	92,10
LE 410	99,18	92,14	99,17	92,24
LE 420	99,18	92,14	99,17	92,24
LE 300	99,45	93,66	99,44	93,87
LERU 301	99,77	95,72	99,75	95,50
SCH 23390	96,69	84,37	96,61	84,40
SH 3	99,00	90,81	98,93	90,63

Tabelle 4.3: Vergleich der COSMO-Sigma-Profile der Enantiomere von B 157 mit den übrigen Dopamin D<sub>1</sub>-Antagonisten

#### 4.4.2 PLS-Analyse der COSMO-Daten

Für die PLS-Regression der COSMO-Sigma-Profile kam das selbstentwickelte Programm **PLS-Toolbox** zum Einsatz. Dieses Programm ermöglicht verschiedene Varianten der Datenbehandlung (Zentrierung, Skalierung, Glättung, Variablenausschluss) und Modellvalidierung (LOO- und LMO-Kreuzvalidierung und Scramble-Test). Im Anhang D.1 ist das Programm näher beschrieben.

Da einige Werte aufgrund der hohen Rechengenauigkeit von MATLAB extrem klein waren ( $10^{-243}$  und kleiner), wurden die Daten als erstes geglättet, indem alle Werte auf  $10^{-10}$  genau gerundet wurden. Die X- und Y-Variablen wurden vor und während der Kreuzvalidierung zentriert. Die für die Berechnung der Kreuzvalidierungskennzahlen erforderlichen Erwartungswerte (Mittelwerte der Variablen) wurden ebenfalls ständig aktualisiert. Eine Skalierung auf eine gemeinsame Standardabweichung wurde unterlassen, da dabei die relative Bedeutung der Variablen nivelliert worden wäre. Die für alle Berechnungen eingesetzte PLS-Methode war die der orthogonalen Scores.

Zunächst wurden die wichtigsten X-Variablen anhand der minimalen Korrelation jeder einzelnen Variable mit den Y-Werten bestimmt. Dies geschah durch eine systematische Berechnung aller PLS-Modelle mit Leave-One-Out-Kreuzvalidierung bei steigendem Schwellenwert für die minimale Korrelation mit den Aktivitätswerten. In Abbildung 4.7 sind die  $q^2$ -Werte dieser Korrelationsschwellenwertabtastung dargestellt. Der Median der  $q^2$ -Werte betrug 0,556, d. h. mehr als die Hälfte der  $q^2$ -Werte lag über 0,5. Der maximale  $q^2$ -Wert betrug 0,733 (SPRESS = 0,71; SDEP = 0,65; NoC = 1) und ergab sich für ein Modell mit dreißig X-Variablen bei einem minimalen quadrierten Korrelationskoeffizienten von  $r^2 = 0,39$ .

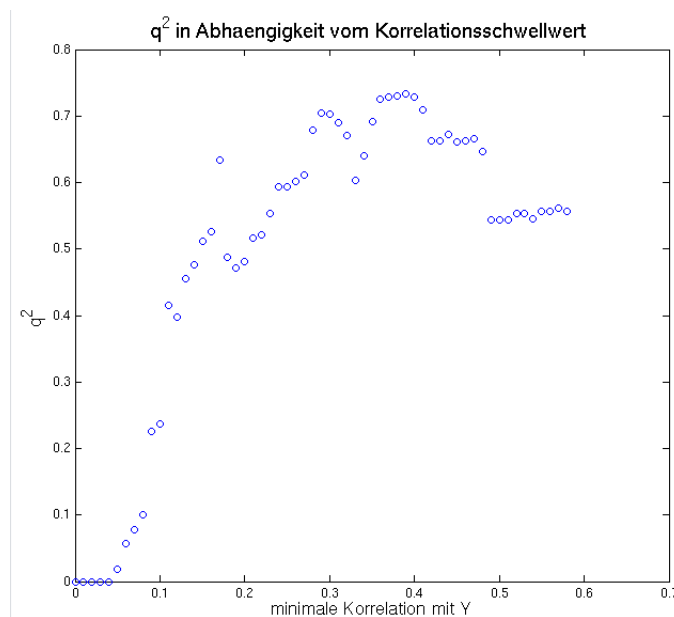


Abbildung 4.7: Die  $q^2$ -Werte der PLS-Analysen mit zunehmendem Ausschluss von X-Variablen durch systematische Erhöhung der minimalen Korrelation mit Y

Um zu überprüfen, ob der gefundene Zusammenhang nicht zufällig ist, wurde ein Scramble-Test durchgeführt. Dabei werden die Y-Werte durcheinander gewürfelt, und eine anschließende Korrelationsschwellenwertabtastung sollte nur sehr schlechte PLS-Modelle ergeben. Diese Prozedur muss mehrfach durchgeführt werden, da es theoretisch auch möglich ist, dass sich die Reihenfolge der Y-Werte nicht oder nur wenig ändert. Durch die Betrachtung des Durchschnitts der  $q^2$ -Werte einer ausreichend großen Anzahl von Abtastungsvorgängen können solche „Ausreißer“ kompensiert werden.

Der Scramble-Test für die COSMO-Daten der Dopamin D<sub>1</sub>-Antagonisten wurde fünfzigmal durchgeführt. Der anfängliche Initialisierungswert für den Zufallsgenerator (Random State) wurde auf 208165,984 gesetzt, um die Ergebnisse reproduzieren zu können. Der Mittelwert der  $q^2$ -Werte betrug 0,177 bei einer Standardabweichung von 0,296. Der durchschnittliche Median der  $q^2$ -Werte lag etwas niedriger bei 0,166 (SD = 0,316). Die Annahme, dass auch zufällig zugeordnete Aktivitätswerte gute PLS-Modelle ergeben können, wurde somit widerlegt.

Anzahl ausgelass. Verb.	$\bar{q}^2$	SD	$\bar{\text{SPRESS}}$	$\bar{\text{SDEP}}$
2	0,717	0,030	0,736	0,672
3	0,699	0,049	0,764	0,696
4	0,702	0,041	0,771	0,701
5	0,686	0,061	0,882	0,800

Tabelle 4.4: Ergebnisse der Leave-Many-Out-Kreuzvalidierungen für das beste PLS-Modell

Die Stabilität des bei der Korrelationsschwellenwertabtastung gefundenen PLS-Modells lässt sich mit Hilfe mehrfacher Durchführung einer Leave-Many-Out-Kreuzvalidierung überprüfen. Diese Methode wurde bereits zur Überprüfung der QSAR-Modelle aus der automatischen PLS (siehe Abschnitt 3.2.3) eingesetzt und eignet sich auch für die Modelle aus den COSMO-Sigma-Profil-Deskriptoren. Diese LMO-Kreuzvalidierung wurde ebenfalls bei einem anfänglichen Random State von 208165,984 fünfzigmal durchgeführt. Bis zu einer Gruppengröße von vier Verbindungen, was dem Auslassen von 33 % entspricht, verringerte sich der  $q^2$ -Wert nur wenig. Die Tabelle 4.4 fasst die Ergebnisse dieser Random-Groups-PLS-Analysen zusammen.

Der nichtvalidierte Regressionskoeffizient betrug 0,82. In Abbildung 4.8 sind die vom (nichtvalidierten) Modell vorhergesagten gegen die gemessenen Bindungsdaten der Dopamin D<sub>1</sub>-Antagonisten aufgetragen worden. Die 30 Variablen dieses Modells sind in Abbildung 4.9 im COSMO-Sigma-Profil der Verbindung (R)-(+)-SCH 23390 markiert worden. Für die Erklärung der Aktivität genügen offensichtlich nur höhere

positive bzw. negative Ladungsdichten. Der gesamte Bereich kleinerer Ladungsdichten bzw. ohne Ladung ( $\sigma = 0$ ) wurde vom Modell ausgeschlossen, da die Korrelation dieser Variablen mit der Aktivität zu gering war ( $r^2 < 0,39$ ).

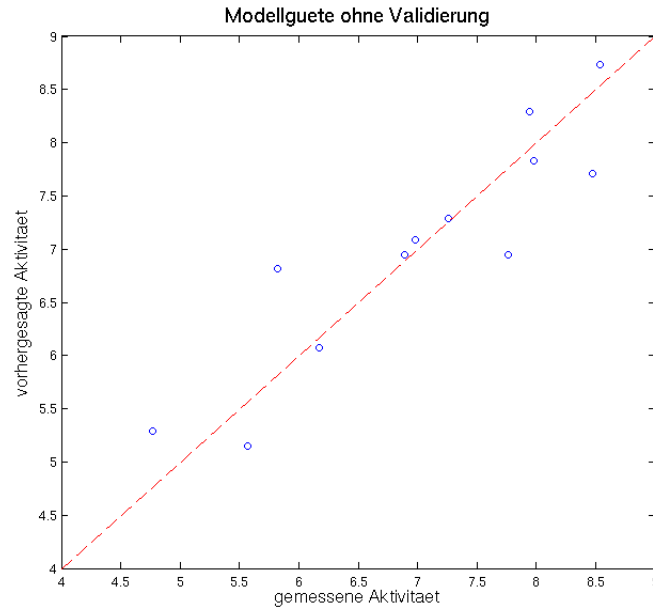


Abbildung 4.8: Vergleich der gemessenen mit der vom (nichtvalidierten) Modell berechneten Aktivität.

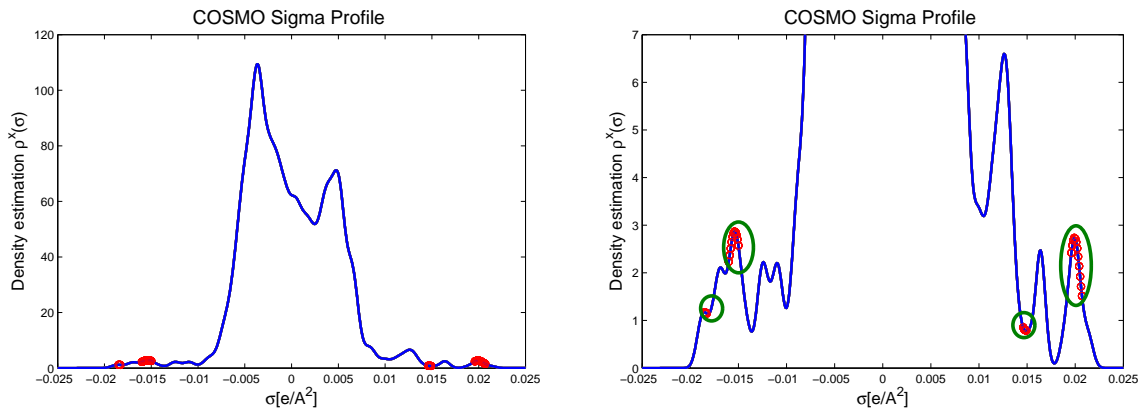


Abbildung 4.9: Vom Modell benutzte Bereiche der Ladungsdichte am Beispiel von (R)-(+)-SCH 23390.



## 5. Zusammenfassung und Ausblick

Die Entwicklung neuartiger Methoden zur Ableitung quantitativer Struktur-Wirkungsbeziehungen (QSAR) stand im Fokus dieser Arbeit. Diese Methoden wurden an Dopaminrezeptorantagonisten der Subtypen D<sub>1</sub>, D<sub>2</sub> und D<sub>3</sub> angewandt und evaluiert.

Dopaminrezeptorantagonisten zeichnen sich durch eine ausgeprägte strukturelle Heterogenität mit nur wenigen gemeinsamen Merkmalen aus, welche die Suche nach einem geeigneten Alignment erheblich erschwert. Besonders bei 3D-QSAR Verfahren wie der CoMFA entsteht durch die fehlende strukturelle Homogenität das Problem, die Verbindungen in einer geeigneten Konformation in korrekter Art und Weise zu überlagern.

Zur Lösung des Alignmentproblems wurden unterschiedliche Strategien angewandt:

1. Reduktion der Konformationsvielfalt durch Konformationsclustering
2. Automatisierte PLS zur Auswahl geeigneter Konformationen und Optimierung des Alignments
3. COSMO-Sigma-Profil als neuer alignmentunabhängiger Deskriptor für die 3D-QSAR

Die Kenntnis über die möglichen Konformationen einer Verbindung ist eine elementare Voraussetzung für die Durchführung von 3D-QSAR-Analysen. Hierzu wurde in dieser Arbeit ein Verfahren entwickelt, welches unabhängig von der Struktur der Verbindungen die konformationelle Flexibilität ermittelt und quantifiziert. Dieses Verfahren bestimmt und vergleicht zunächst alle sinnvollen Konformationen. Anschließend wird ein repräsentativer Querschnitt aus mehreren Konformationsclustern gebildet, die nach ihrer wahrscheinlichen, auf den Kriterien Energie und Häufigkeit basierenden, Relevanz klassifiziert werden.

Für die Realisierung der CoMF-Analyse wurden so nur die relevanten Konformationen berücksichtigt, die aber dennoch zu einer beachtlichen Zahl an möglichen QSAR-Modellen unter Einbeziehung alternativer Alignments führten. Die während dieser Promotion entwickelte automatisierte PLS ermöglicht die Modellselektion auf objektiver Basis. Die Berechnung und Auswertung zehntausender QSAR-Modelle hatte zum Ziel, die besten Konformationen für das optimale Alignment zu isolieren.

Die geeignete Konformation war auch für andere aus der 3D-Struktur abgeleitete Deskriptoren relevant. Die Sigma-Profile aus dem COSMO-Solvatationsmodell von Andreas Klamt [92, 95] fassen elektronische Oberflächeneigenschaften von Molekülen zusammen. Bei diesem Deskriptor werden 3D-Informationen aus quantenmechanischen Berechnungen in den eindimensionalen Raum projiziert, wobei die Abhängigkeit von der Orientierung im Raum eliminiert wird. Basierend auf den COSMO-Sigma-Profilen der Dopaminrezeptorantagonisten wurde ein alignmentunabhängiges QSAR-Modell erstellt. Mit Hilfe des selbstentwickelten Programms PLS-Toolbox wurden hierfür die relevanten Bereiche der COSMO-Sigma-Profile identifiziert und mit den Radioligandbindungsdaten am Dopamin D<sub>1</sub>-Rezeptor korreliert.

QSAR-Modelle für Liganden an Dopaminrezeptoren aller Subtypen sind von besonderem Interesse, da sie die Entwicklung neuer potenter Verbindungen positiv beeinflussen können. Die Verfügbarkeit zusätzlicher Aktivitätsdaten kann dazu genutzt werden, die erstellten Modelle zu verbessern und auszubauen. Klinisch relevant ist stets das gesamte pharmkologische Profil, weshalb in Ergänzung zu den QSAR-

Modellen für den D<sub>1</sub>-Rezeptor auch Modelle für die Bindung an den D<sub>2</sub>-artigen Dopaminrezeptorsubtypen wünschenswert sind.

Mit Ausnahme des Rinderrhodopsins liegen für G-Protein gekoppelte Rezeptoren keine Röntgenkristallstrukturdaten vor. Auf der Basis des Rinderrhodopsins könnte in Zukunft ein Homologiemodell erstellt werden, anhand dessen mittels Dockingmethoden die vorgeschlagenen Alignments und Pharmakophormodelle überprüft werden könnte.

Bei der automatisierten PLS steht der  $q^2$ -Wert als wichtigstes Kriterium für die Optimierung der Modelle im Vordergrund. In Zukunft könnte diese Optimierung auf eine breitere Basis gestellt werden, was z. B. durch Integration eines Testdatensatzes und einer Leave-Many-Out-Kreuzvalidierung (LMO-CV) in den Prozess der Optimierung und Modellauswahl erfolgen könnte.

Das im Rahmen dieser Arbeit entwickelte Programm `PLS-Toolbox` bietet sich für die Implementierung solcher und weiterer Funktionen an, jedoch können Erweiterungen, die das CoMFA-Patent [67] verletzen würden, bis zum Ablauf des Patents nicht aufgenommen werden. Geplant sind allerdings Verbesserungen der Validierungsmöglichkeiten sowie die Aufnahmen weiterer Regressionstechniken, um den Entwicklungsprozess für QSAR-Modelle zu beschleunigen.

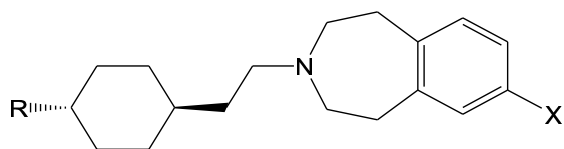
Die COSMO-Sigma-Profile offerieren ebenfalls noch vielfältige Möglichkeiten der Erforschung und Anwendung. In dieser Dissertation konnte die Eignung dieser Profile als alignmentunabhängiger Deskriptor für die QSAR-Analyse gezeigt werden. Diese Deskriptoreigenschaften sollten unbedingt an weiteren Datensätzen untersucht und evaluiert werden.

Die mit Hilfe der COSMO-Sigma-Profile erstellten QSAR-Modelle ermöglichen die Vorhersage von Aktivitäten neuer Verbindungen, allerdings können die Aktivitätsunterschiede der verschiedenen Verbindungen mit diesem Modell nur unzureichend auf konkrete Strukturunterschiede zurückgeführt werden. Hier könnte eine Rückprojektion der Modellvariablen auf die Oberfläche der Moleküle hilfreich sein. Da bei

der Generierung der Profile eine Dimensionsreduktion stattfindet, ist diese Rückprojektion jedoch nur in gewissen Grenzen möglich. Die COSMO-Sigma-Profile sind ein vielversprechender Deskriptor für die Verwendung in 3D-QSAR-Modellen und besitzen großes Potenzial für weitere Untersuchungen.

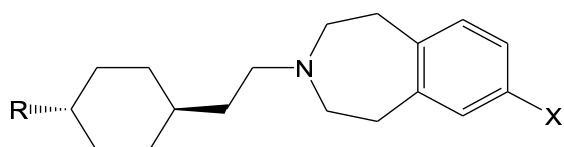
Die entwickelten und erprobten Methoden und Programme sind nicht nur auf Liganden des Dopaminrezeptors anwendbar. Sie offerieren neue Möglichkeiten bei der Erstellung und Verbesserung von QSAR-Modellen im Allgemeinen.

## A. Anhang – Tabellen zu den D<sub>2</sub>- und D<sub>3</sub>-Rezeptorantagonisten

MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)MS = Methylsulfonyl (-SO<sub>2</sub>Me)

Nr.	X	R	a	b	c
18	MSO				
19	MSO				
20	MSO				
21	MSO				
22	MSO				
23	MSO				
24	MSO				
25	MSO				
26	MSO				

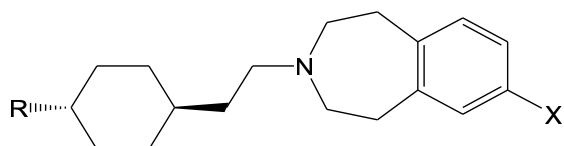
Tabelle A.1a: gewählte Konformationen



MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)  
 MS = Methylsulfonyl (-SO<sub>2</sub>Me)

Nr.	X	<i>R</i>	<i>a</i>	<i>b</i>	<i>c</i>
27	MSO				
28	MSO				
29	MSO				
30	MSO				
31	MSO				
32	MSO				
33	MSO				

Tabelle A.1b: gewählte Konformationen

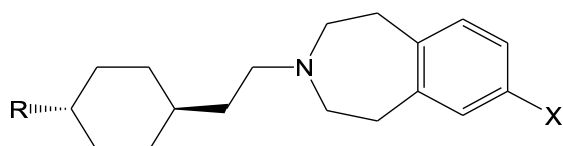


MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)  
 MS = Methylsulfonyl (-SO<sub>2</sub>Me)

Nr.	X	R	a	b	c
34	MSO				
43	MS				
44	MS				
45	MS				
46	MS				
47	MS				
48	MS				
49	MS				
50	MS				
51	MS				

Tabelle A.1c: gewählte Konformationen



MSO = Methylsulfonyloxy (-OSO<sub>2</sub>Me)MS = Methylsulfonyl (-SO<sub>2</sub>Me)

Nr.	X	<i>R</i>	<i>a</i>	<i>b</i>	<i>c</i>
52	MS				
53	MS				
54	MS				
55	MS				
56	MS				
57	MS				
58	MS				

Tabelle A.1d: gewählte Konformationen

CoMFA $D_2$ Selektion 1				
Feld	SDEP	$q^2$	NoC	Ausreißer
ster+ele	0,182	0,757	4	47(-0,41); 53(-0,39)
H-Brücken	0,180	0,746	2	27(-0,48)
sterisch	0,188	0,740	4	47(-0,45)
elektrostat.	0,160	0,845	9	57a(-0,282)

CoMFA $D_2$ Selektion 2				
Feld	SDEP	$q^2$	NoC	Ausreißer
ster+ele	0,196	0,718	4	53(-0,41)
H-Brücken	0,177	0,786	6	27(-0,36)
sterisch	0,202	0,711	5	47(-0,45)
elektrostat.	0,154	0,851	8	53(-0,31)

CoMFA $D_2$ Selektion 3				
Feld	SDEP	$q^2$	NoC	Ausreißer
ster+ele	0,192	0,730	4	47(-0,39); 53(-0,43)
H-Brücken	0,182	0,740	2	27(-0,48)
sterisch	0,199	0,730	6	47(-0,48)
elektrostat.	0,173	0,820	9	54c(-0,42)
elektrostat.	0,161	0,797	2	27(-0,37)

CoMFA $D_2$ Selektion 4				
Feld	SDEP	$q^2$	NoC	Ausreißer
ster+ele	0,196	0,716	4	47(-0,40); 53a(-0,46)
H-Brücken	0,182	0,740	2	27a(-0,47)
sterisch	0,193	0,726	4	47(-0,45); 53a(-0,41)
elektrostat.	0,187	0,726	2	27a(-0,51); 52a(0,39)

Tabelle A.2: CoMFA-Ergebnisse für den  $D_2$ -Rezeptor mit den verschiedenen Selektionen. Der Filterwert betrug 2,0.

CoMFA D <sub>3</sub> Selektion 1					
Feld	ohne	SDEP	q <sup>2</sup>	NoC	Ausreißer
ster+ele		0,233	0,475	5	27(-0,83)
ster+ele	27	0,184	0,574	7	29a(0,42)
H-Brücken		0,264	0,426	9	27(-0,96)
hydrophob	27	0,152	0,661	3	22(-0,31); 51(-0,30)
sterisch		0,230	0,485	5	27(-0,79)
sterisch	27	0,172	0,551	2	22(-0,34); 29a(0,38)
elektrostat.		0,215	0,584	7	27(-0,69)
elektrostat.	27	0,176	0,608	7	54a(-0,38)

CoMFA D <sub>3</sub> Selektion 2					
Feld	ohne	SDEP	q <sup>2</sup>	NoC	Ausreißer
ster+ele		0,258	0,355	5	27(-0,85); 29b(0,54)
ster+ele	27	0,169	0,566	2	29b(0,36)
H-Brücken		0,248	0,450	7	27(-0,90)
H-Brücken	27	0,171	0,672	7	
sterisch		0,266	0,316	5	27(-0,83); 29b(0,58)
sterisch	27	0,173	0,545		29b(0,36)
elektrostat.		0,261	0,391	7	27(-0,78)
elektrostat.	27	0,171	0,554	2	29b(0,40); 56b(-0,35)

CoMFA $D_3$ Selektion 3					
Feld	ohne	SDEP	$q^2$	NoC	Ausreißer
ster+ele		0,243	0,340	1	27(-0,98)
ster+ele	27	0,166	0,650	7	29c(0,31)
H-Brücken		0,310	0,381	14	27(-0,99)
H-Brücken		0,247	0,323	1	27(-0,90)
H-Brücken	27	0,174	0,540	2	—
sterisch		0,262	0,432	9	27(-0,77)
sterisch	27	0,174	0,618	7	—
elektrostat.		0,246	0,347	2	27(-0,92)
elektrostat.	27	0,164	0,591	2	22(0,34); 29c(-0,33)

CoMFA $D_3$ Selektion 4					
Feld	ohne	SDEP	$q^2$	NoC	Ausreißer
ster+ele		0,240	0,442	5	27a(-0,84)
ster+ele	27a	0,171	0,615	6	29c(0,35)
H-Brücken		0,249	0,333	2	27a(-0,97)
H-Brücken	27a	0,173	0,543	2	—
sterisch		0,242	0,433	5	27a(-0,84)
sterisch	27a	0,181	0,570	6	29c(0,42)
elektrostat.		0,253	0,289	1	27a(-1.0)
elektrostat.	27a	0,167	0,560	1	51a(-0,35)

Tabelle A.3: CoMFA-Ergebnisse für den  $D_3$ -Rezeptor mit den verschiedenen Selektionen. Der Filterwert betrug 2,0.

CoMFA Selektivität Selektion 1					
Feld	ohne	SDEP	q <sup>2</sup>	NoC	Ausreißer
ster+ele	27	0,186	0,177	5	52(-0,41); 53(-0,36)
H-Brücken	27	0,170	0,231	2	—
sterisch	27	0,177	0,192	3	29a(0,41); 52(-0,38)
elektrostat.	27	0,172	0,208	2	52(-0,37)

CoMFA Selektivität Selektion 2					
Feld	ohne	SDEP	q <sup>2</sup>	NoC	Ausreißer
ster+ele	27	0,185	0,156	4	52(-0,39); 53(0,37)
H-Brücken	27	0,173	0,206	2	—
sterisch	27	0,173	0,255	4	29b(0,36); 52(-0,37)
elektrostat.	27	0,172	0,185	1	—

CoMFA Selektivität Selektion 3					
Feld	ohne	SDEP	q <sup>2</sup>	NoC	Ausreißer
ster+ele	27	0,189	0,114	4	29c(0,43); 52(-0,38); 53(0,40)
H-Brücken	27	0,171	0,190	1	—
sterisch	27	0,185	0,214	6	29c(0,42); 52(-0,46)
elektrostat.	27	0,175	0,184	2	52(-0,38)

CoMFA Selektivität Selektion 4					
Feld	ohne	SDEP	q <sup>2</sup>	NoC	Ausreißer
ster+ele	27	0,176	0,147	1	—
H-Brücken	27	0,171	0,191	1	53a(0,34)
sterisch	27	0,181	0,215	5	29c(0,42); 52a(-0,46); 53a(0,35)
elektrostat.	27	0,173	0,172	1	—

Tabelle A.4: CoMFA-Ergebnisse für die Selektivität mit den verschiedenen Selektionen. Der Filterwert betrug 2,0.

Konformation	Anzahl	Ø Residuen	Konformation	Anzahl	Ø Residuen
18	239	0,105	18	602	0,140
19	239	0,102	19	377	0,081
20	239	0,080	20	602	0,044
21	239	0,159	21	571	0,119
22	207	0,069	22	602	0,078
23	239	0,050	23	602	0,020
24	239	0,147	24	418	<i>0,105</i>
25	239	0,046	25	602	0,130
26A	125	0,122	26	267	0,024
28A	239	0,125	28	186	<i>0,064</i>
29C	101	0,185	29C	390	<i>0,167</i>
30A	239	0,087	30A	602	0,072
31	61	0,158	31	161	<i>0,065</i>
32A	239	0,092	32A	602	0,052
33A	239	0,075	33A	602	0,087
34A	239	0,176	34A	403	<i>0,137</i>
43	239	0,038	43	602	0,054
44	239	0,075	44	602	0,103
45	239	0,041	45	602	0,037
46	239	0,175	46	602	0,161
47	239	0,249	47	602	0,218
48	239	0,026	48	602	0,039
49	239	0,041	49	602	0,032
50	239	0,034	50	602	0,023
51	239	0,142	51A	314	0,086
52	157	0,262	52A	265	0,173
53A	150	0,205	53A	337	0,112
54A	68	0,033	54A	602	0,038
55A	239	0,072	55A	602	0,082
56A	239	0,185	56	282	0,085
57A	239	0,109	57C	224	0,060
58A	239	0,032	58A	602	0,036

Tabelle A.5: Durchschnittliche Residuen des Feldes „ele+ster“ vor (links) und nach (rechts) Optimierung für die Selektivität. Die Gesamtzahl der ausgewerteten CoMF-Analysen betrug 239 (links) bzw. 602 (rechts).

Konformation	Anzahl	Ø Residuen	Konformation	Anzahl	Ø Residuen
18	380	0,032	18	1393	0,053
19	380	0,110	19	796	0,091
20	380	0,042	20	1393	0,026
21	380	0,127	21	1123	0,097
22	292	0,061	22	1393	0,060
23	380	0,060	23	1393	0,031
24	380	0,133	24	870	0,114
25	380	0,028	25	1393	0,105
26A	196	0,139	26	642	0,024
28A	380	0,120	28	457	0,076
29C	134	0,187	29C	696	0,165
30A	380	0,077	30A	1393	0,061
31A	113	0,121	31A	587	0,096
32A	380	0,085	32A	1393	0,060
33A	380	0,102	33A	1393	0,092
34A	380	0,167	34A	797	0,140
43	380	0,083	43	1393	0,092
44	380	0,086	44	1393	0,101
45	380	0,014	45	1393	0,014
46	380	0,127	46	1393	0,124
47	380	0,230	47	1393	0,221
48	380	0,032	48	1393	0,051
49	380	0,016	49	1393	0,018
50	380	0,022	50	1393	0,033
51	380	0,131	51	773	0,096
52	241	0,257	52A	634	0,156
53A	227	0,176	53A	759	0,111
54A	108	0,043	54A	1393	0,041
55A	380	0,051	55A	1393	0,070
56A	380	0,145	56	611	0,111
57A	380	0,106	57C	459	0,059
58A	380	0,030	58A	1393	0,031

Tabelle A.6: Durchschnittliche Residuen des sterischen Feldes vor (links) und nach (rechts) Optimierung für die Selektivität. Die Gesamtzahl der ausgewerteten CoMF-Analysen betrug 380 (links) bzw. 1393 (rechts).

Konformation	Anzahl	Ø Residuen	Konformation	Anzahl	Ø Residuen
18	213	0,175	18	1679	0,155
19	213	0,034	19	1679	0,035
20	213	0,070	20	1679	0,090
21	213	0,221	21A	1047	0,149
22	105	0,203	22A	602	0,181
23	213	0,054	23	1679	0,052
24	213	0,137	24A	987	0,052
25	213	0,073	25	1679	0,105
26A	130	0,072	26A	1679	0,106
28A	213	0,138	28B	513	0,090
29C	106	0,133	29C	473	<i>0,144</i>
30A	213	0,095	30A	1679	0,076
31	47	0,208	31B	404	0,161
32A	213	0,056	32A	1679	0,056
33A	213	0,047	33A	1679	0,088
34A	213	0,251	34A	1197	<i>0,182</i>
43	213	0,029	43	1679	0,019
44	213	0,051	44	1679	0,068
45	213	0,047	45	1679	0,017
46	213	0,201	46	1679	0,110
47	213	0,106	47	1679	0,085
48	213	0,084	48	1679	0,087
49	213	0,080	49	1679	0,061
50	213	0,045	50	1679	0,033
51	213	0,167	51	1679	0,083
52	138	<i>0,232</i>	52	1679	0,130
53A	132	0,191	53A	1675	0,187
54A	59	0,052	54A	1679	0,060
55A	213	0,086	55A	1679	0,093
56A	213	0,176	56A	1089	0,114
57A	213	0,135	57C	438	0,068
58A	213	0,067	58A	1679	0,070

Tabelle A.7: Durchschnittliche Residuen des elektrostatischen Feldes „ele“ vor (links) und nach (rechts) Optimierung für die Selektivität. Die Gesamtzahl der ausgewerteten CoMF-Analysen betrug 213 (links) bzw. 1679 (rechts).



Konformation	Anzahl	Ø Residuen	Konformation	Anzahl	Ø Residuen
18	6	0,155	18	410	0,160
19	6	0,055	19	410	0,054
20	6	0,157	20A	282	0,044
21	6	0,280	21	364	0,267
22C	6	0,266	22C	364	0,244
23	6	0,136	23A	260	0,093
24	6	0,052	24	410	0,060
25	6	0,140	25	274	0,124
26A	3	0,091	26A	410	0,123
28A	6	0,026	28A	410	0,045
29	4	0,113	29	164	0,069
30A	6	0,050	30A	410	0,058
31A	6	0,065	31A	410	0,067
32A	6	0,063	32A	410	0,078
33A	6	0,028	33A	410	0,053
34A	6	0,206	34	366	0,073
43	6	0,087	43	410	0,091
44	6	0,050	44	410	0,031
45	6	0,069	45	410	0,081
46	6	0,333	46	410	0,330
47	6	0,051	47	410	0,051
48	6	0,075	48	410	0,066
49	6	0,030	49	410	0,030
50	6	0,114	50A	272	0,049
51	6	0,186	51/51A	205	0,169
52	3	0,269	52	242	0,228
53A	4	0,244	53A	206	0,235
54A	4	0,063	54A	410	0,077
55A	6	0,123	55	158	0,030
56A	6	0,169	56A	410	0,145
57A	6	0,102	57A	410	0,115
58A	6	0,018	58A	410	0,015

Tabelle A.8: Durchschnittliche Residuen des Feldes „H-Brücken“ vor (links) und nach (rechts) Optimierung für die Selektivität. Die Gesamtzahl der ausgewerteten CoMF-Analysen betrug 6 (links) bzw. 410 (rechts).

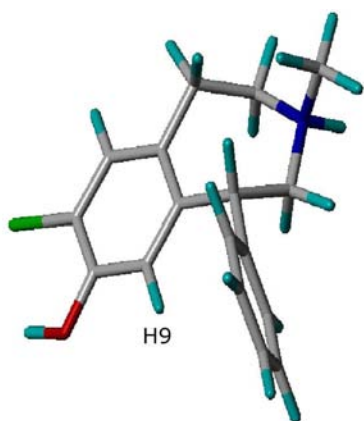


## B. Anhang – NMR-Konformationsuntersuchungen

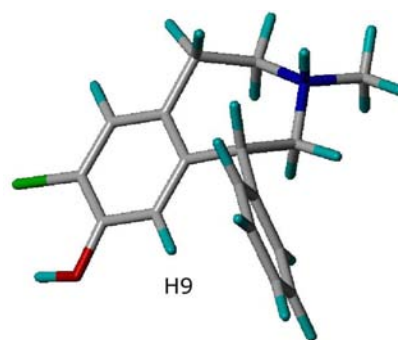
Das Konformationsclustering von (R)-(+)-SCH 23390 ergab zwei Hauptrepräsentanten - die Konformationen R490000 und R300000 (s. Tabelle C.4 auf Seite 130). Dabei stellt R490000, bei welchem der siebengliedrige heterozyklische Ring eine Sesselkonformation einnimmt, das globale Energieminimum dar. Der angeknüpfte Phenylring ist in äquatorialer Position. Im Konformer R300000 ist das Tetrahydroazepingerüst ebenfalls in Sesselkonformation, aber in die entgegengesetzte Richtung (spiegelbildlich zur Ebene des annelierten Chlorphenolrings) gefaltet. Dadurch ist der Phenylring bei gleicher Konfiguration des Moleküls in axialer Position angeknüpf (s. Abbildung B.1).

Die Energiediagramme für die Rotation um die Bindung, welche den Phenylring und das Tetrahydroazepingerüst verknüpft, sind für beide Konformationen in Abbildung 3.1 auf Seite 25 dargestellt. Die Energieminima liegen bei  $48^\circ$  und  $228^\circ$  für Konformation R490000 bzw. bei  $136^\circ$  und  $316^\circ$  für Konformation R300000, wobei diese Minima aufgrund der  $C_{2v}$ -Symmetrie des unsubstituierten Phenylrings identisch sind.

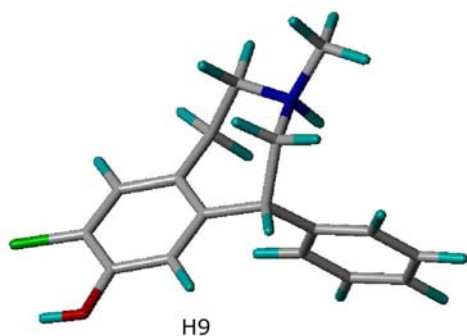
Befinden sich Atome (bei  $^1\text{H}$ -NMR-Spektroskopie besonders H-Atome) in räumlicher Nähe zu  $\pi$ -Elektronensystemen (Aromaten, C-C-Doppel- und Dreifachbindungen), so wird deren Elektronenhülle polarisiert. Der Atomkern kann — je nach Lage zum  $\pi$ -



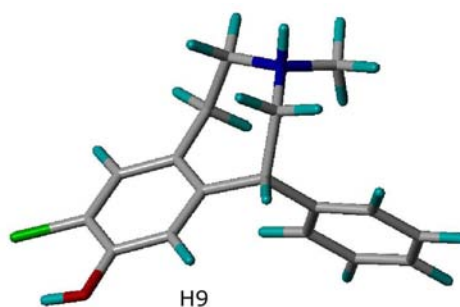
(a) R490000a: Axiale Stellung der N-Alkylgruppe



(b) R490000b: Äquatoriale Stellung der N-Alkylgruppe



(c) R300000a: Axiale Stellung der N-Alkylgruppe



(d) R300000b: Äquatoriale Stellung der N-Alkylgruppe

Abbildung B.1: (R)-(+)-SCH 23390: Die für die NMR-Berechnungen berücksichtigten am Benzazepinstickstoffatom protonierten repräsentativen Konformationen.

Elektronensystem — entweder zusätzlich abgeschirmt oder entschirmt werden. Der räumliche Einfluss von  $\pi$ -Elektronensystemen auf die chemische Verschiebung der NMR-Signale von Wasserstoffkernen wird in [114] sehr anschaulich beschrieben.

Die beiden repräsentativen Konformationen sollten sich bezüglich des  $^1\text{H}$ -NMR-Signals von Atom H9 unterscheiden, da dieses in jeweils unterschiedlicher Position zum Anisotropiekegel des Phenylrings steht. Nur bei Konformation R4900000 befindet es sich innerhalb dieses Kegels, was im  $^1\text{H}$ -NMR-Spektrum eine Hochfeldverschiebung des Signals dieses isolierten Kerns zur Folge haben müsste.

Chemische Verschiebungen einzelner  $^1\text{H}$ -NMR-Signale lassen sich mittlerweile recht gut mit quantenchemischen Methoden berechnen. In [115] sowie [116] werden die Grundlagen und praktischen Aspekte dieser Berechnungen erläutert. Berechnet man

außerdem noch die Kopplungskonstanten, lassen sich sogar vollständige Spektren ableiten [117] (wenn man alle möglichen Konformere sowie deren Anteil im zeitlichen Mittel des Messvorgangs betrachtet).

Für die beiden Wasserstoffkerne wurde mit der in GAUSSIAN implementierten DFT-Hybridmethode B3LYP mit dem Basissatz 6-31G(d,p) ein Unterschied in der chemischen Verschiebung von 0.35 (bzw. 0.3 für die protonierte Form) berechnet. Die Abbildung B.1 zeigt die für die Berechnung verwendeten Konformationen und Tabelle B.1 die detaillierten Ergebnisse. Die Berechnungen wurden ohne die Einbeziehung eines Lösungsmittels durchgeführt. Es wurde auch versucht, den Lösungsmiteleinfluss von D<sub>2</sub>O zu berücksichtigen. Die Berechnungen der NMR-Verschiebungen mit der in GAUSSIAN implementierten SCRF-Methode (Self Consistent Reaction Field) führten jedoch zu unsinnigen Ergebnissen.

Im 200 MHz NMR-Spektrum, welches bei Raumtemperatur (298 K) aufgenommen wurde, zeigte sich zunächst nur ein sehr breites Signal für das H-Atom H9 (s. Abbildung B.5). Nach dem Übergang zum 500 MHz NMR-Spektrometer wurde bei gleicher Temperatur eine sehr geringfügige Trennung der Signale sichtbar (s. Abbildungen B.2 und B.3). Erst nach Aufnahme eines NMR-Spektrums bei 280 K (s. Abbildungen B.4 und B.5) trennte sich das Signal fast vollständig, so dass ein zweites Signal mit einer Hochfeldverschiebung von ca. 0,67 ppm zu sehen war.

Dieses beobachtete Phänomen nennt man Koaleszenz. Oberhalb der Koaleszenztemperatur  $T_C$  (coalescence temperature) fließen die NMR-Signale zusammen. Diese Temperatur hängt von der Messfrequenz ab. Je höher die Messfrequenz, desto höher ist auch die Koaleszenztemperatur. Laut [118] hat eine Verdopplung der Messfrequenz eine Erhöhung der  $T_C$  um ungefähr 10°C zur Folge.

Bezogen auf die gemessenen Spektren bedeutet das, dass sich beide Konformationen bei Zimmertemperatur so schnell ineinander umwandeln, dass es im NMR-Spektrum zu einer Koaleszenz der Signale kommt. Erst unterhalb der Koaleszenztemperatur kann man mit einem empfindlichen NMR-Spektrometer beide Konformationen ne-

Name	Energie (a.u.)	$\sigma$ -H9	$\delta$ -H9	$\sigma$ -H31	$\sigma$ -H32	$\sigma$ -H33	$\sigma$ -NCH <sub>3</sub>	$\delta$ -NCH <sub>3</sub>
R490000	-1248.83786	25.438	<b>6.062</b>	28.379	29.577	29.622	29.193	<b>2.307</b>
R300000	-1248.8355	25.085	<b>6.415</b>	29.816	29.917	29.475	29.736	<b>1.764</b>
$\Delta$ E unprot.	-1.47462 kcal/mol	$\sigma$ -Diff	0.353					
R490000a	-1249.22465	25.089	6.412	27.705	28.970	28.964	28.546	2.954
R490000b	-1249.22879	25.069	6.432	28.761	28.732	28.779	28.756	2.743
Mean	-1249.22672		<b>6.422</b>					2.848
R300000a	-1249.22429	24.745	6.756					
R300000b	-1249.22479	24.821	6.679	29.240	29.423	28.671	29.111	2.389
Mean	-1249.22954		<b>6.717</b>					
$\Delta$ E (mean) prot.	-1.76955 kcal/mol	$\sigma$ -Diff	0.296					

Tabelle B.1: Berechnete chemische Verschiebungen für H9 und H31-33 (NCH<sub>3</sub>) der protonierten (a, b) und unprotonierten (ohne Suffix) Minimum-Energie-Konformationen R490000 und R300000 von (R)-(+)-SCH23390,  $\sigma$ -Werte stellen die absoluten berechneten Verschiebungen dar,  $\delta$ -Werte sind auf  $\sigma_{TMS}$  bezogen, Werte für NCH<sub>3</sub> wurden aus den Werten von H31-H33 gemittelt.

beneinander anhand der H9-Signale erkennen. Die Tatsache, dass die übrigen Banden breiter erscheinen, hängt mit der bei 280 K höheren Viskosität von D<sub>2</sub>O zusammen.

Die NMR-Messungen wurden mit der am Stickstoffatom protonierten Form von (R)-(+)-SCH 23390 durchgeführt. Diese besitzt strenggenommen am N-Atom ein weiteres Chiralitätszentrum. Die energetische Inversionsbarriere dafür ist in wässriger Lösung jedoch zu gering bzw. ist die Protonierung und Deprotonierung für ein 500 MHz NMR-Spektrometer zu schnell. Sonst hätte man im NMR-Spektrum ebenso eine Signalaufspaltung (im geringeren Verhältnis) für die N-Methylgruppe feststellen können, da deren H-Atome bei Konformation R490000 in äquatorialer Position ebenfalls im Einfluss der Anisotropie des Phenylrings stehen kann. Die Verschiebungen dafür wurden ebenfalls vorausberechnet (s. Tabelle B.1), ließen sich jedoch nicht im NMR-Spektrum nachweisen. Die experimentell bestimmten Verschiebungen sind in Tabelle B.2 zusammengefasst.

Frequenz (MHz)	Temp. (K)	$\delta$ -H9-1	$\delta$ -H9-2	$\sigma$ - Diff.	$\delta$ -NCH <sub>3</sub>
200	298	6.53		0	2.99
500	298	6.35	6.77	0.42	2.99
500	280	6.23	6.9	0.67	3.00

Tabelle B.2: Experimentelle chemische Verschiebungen im 200 u. 500 MHz <sup>1</sup>H-NMR.

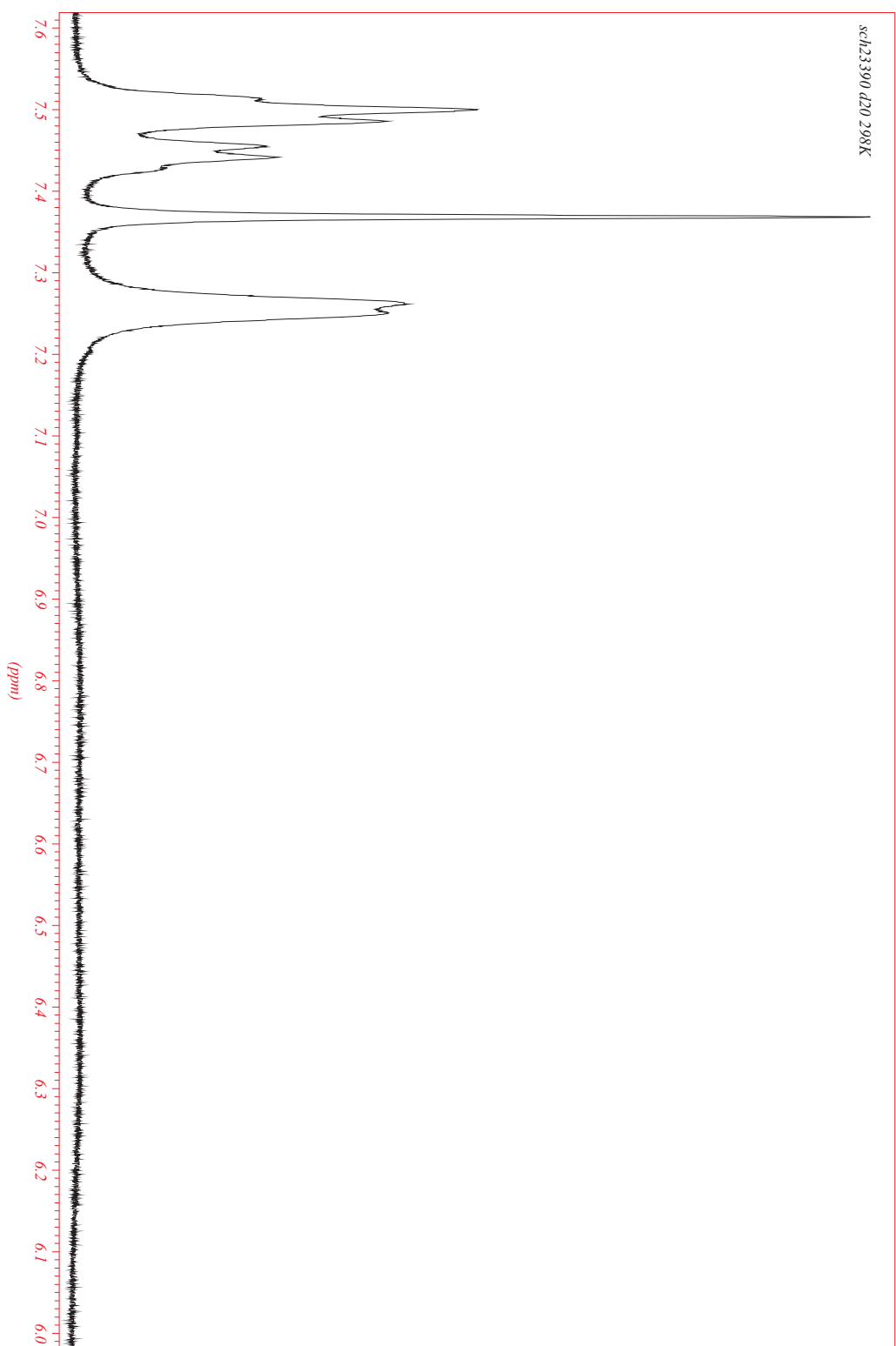


Abbildung B.2: 500 MHz  $^1\text{H}$ -NMR Spektrum von (R)-(+)-SCH23390 bei 298K: Übersicht über den Aromatenbereich



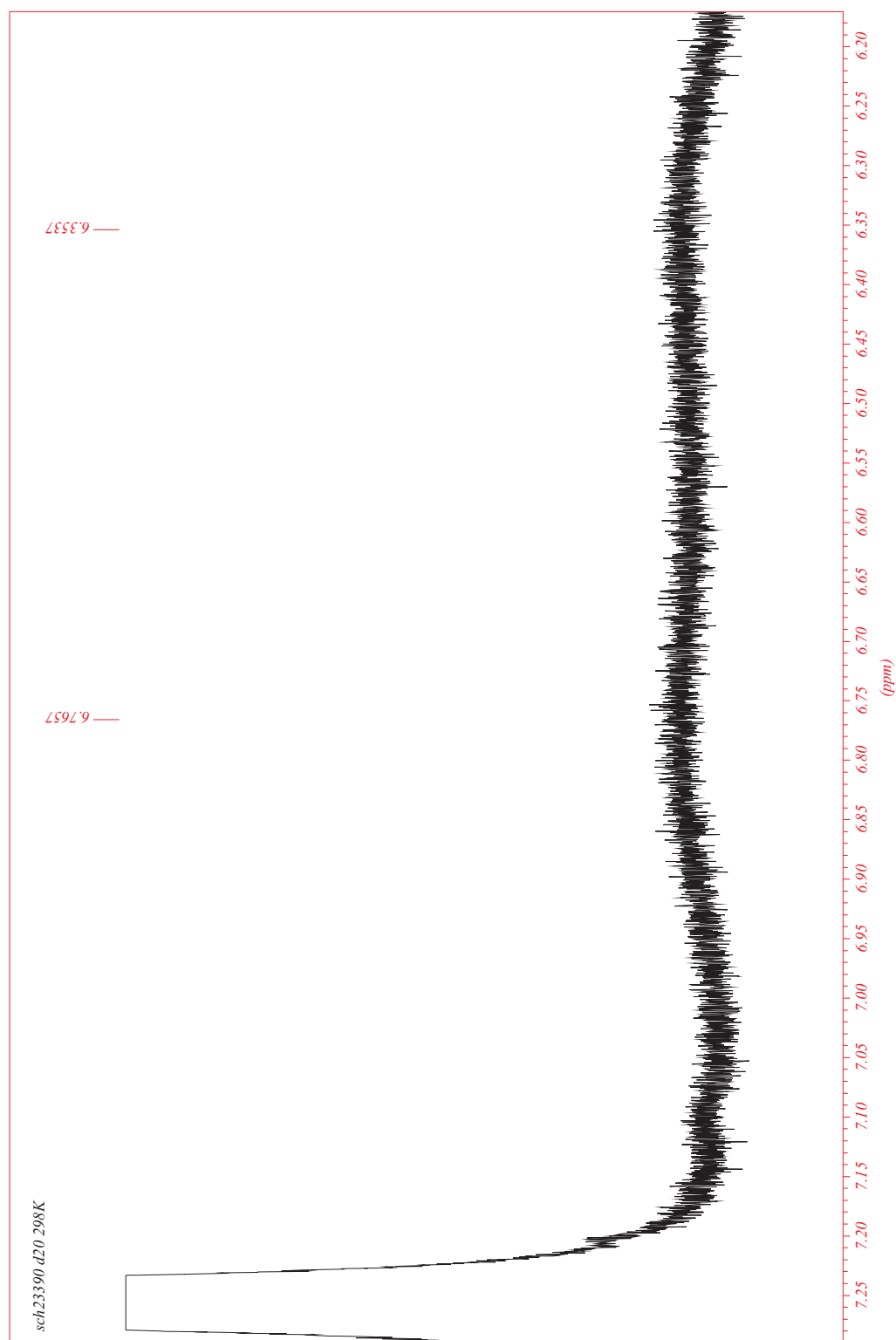


Abbildung B.3: 500 MHz  $^1\text{H}$ -NMR Spektrum of (R)-(+)-SCH 23390 bei 298K: Vergrößerung des Bereichs von H9

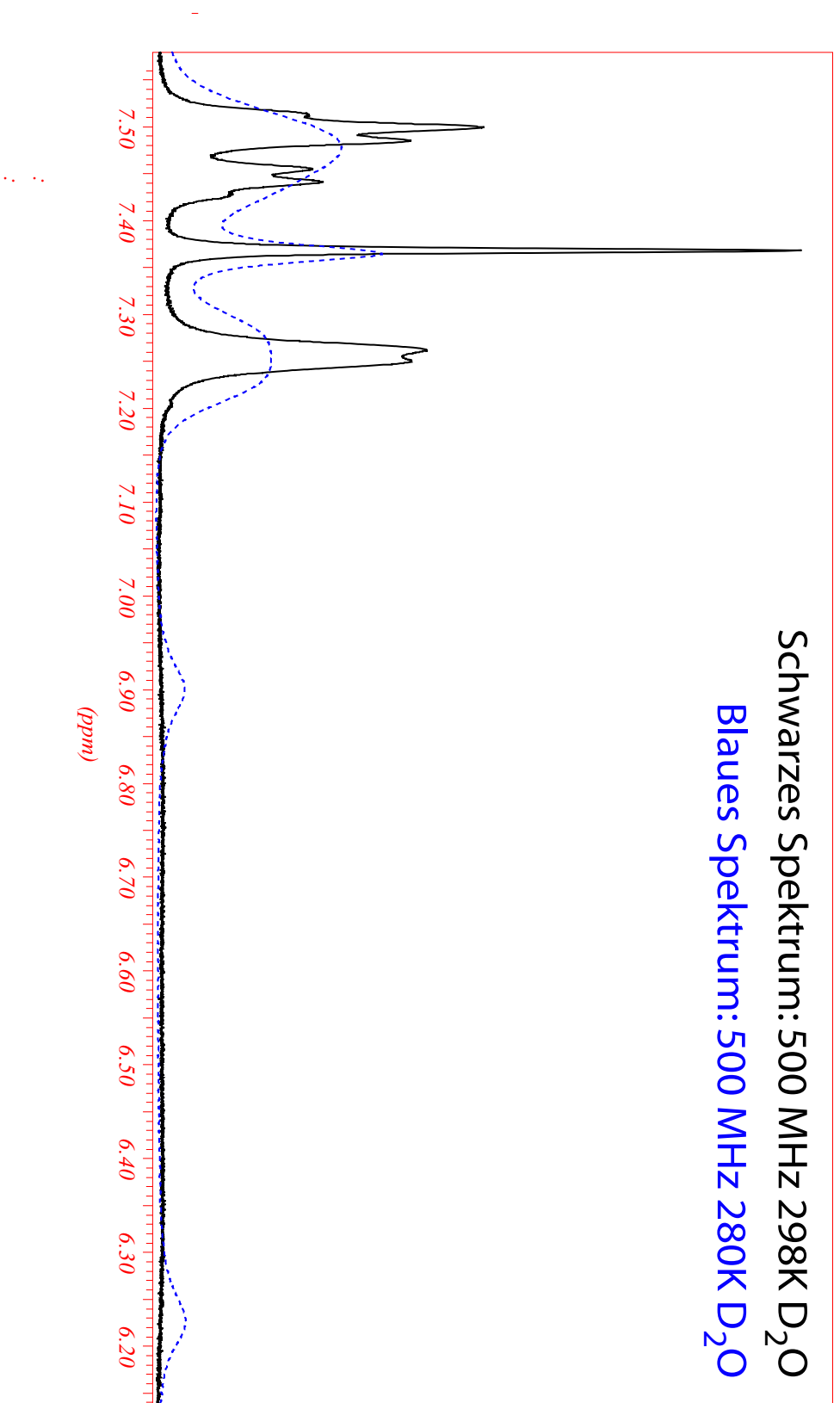


Abbildung B.4: 500 MHz <sup>1</sup>H-NMR Spektren von (R)-(+)-SCH 23390 bei 298 K (schwarz) und 280 K (blau, gestrichelt) überlagert

Schwarzes Spektrum: 200 MHz, 298 K, D<sub>2</sub>O  
Blaues Spektrum: 500 MHz, 280 K, D<sub>2</sub>O

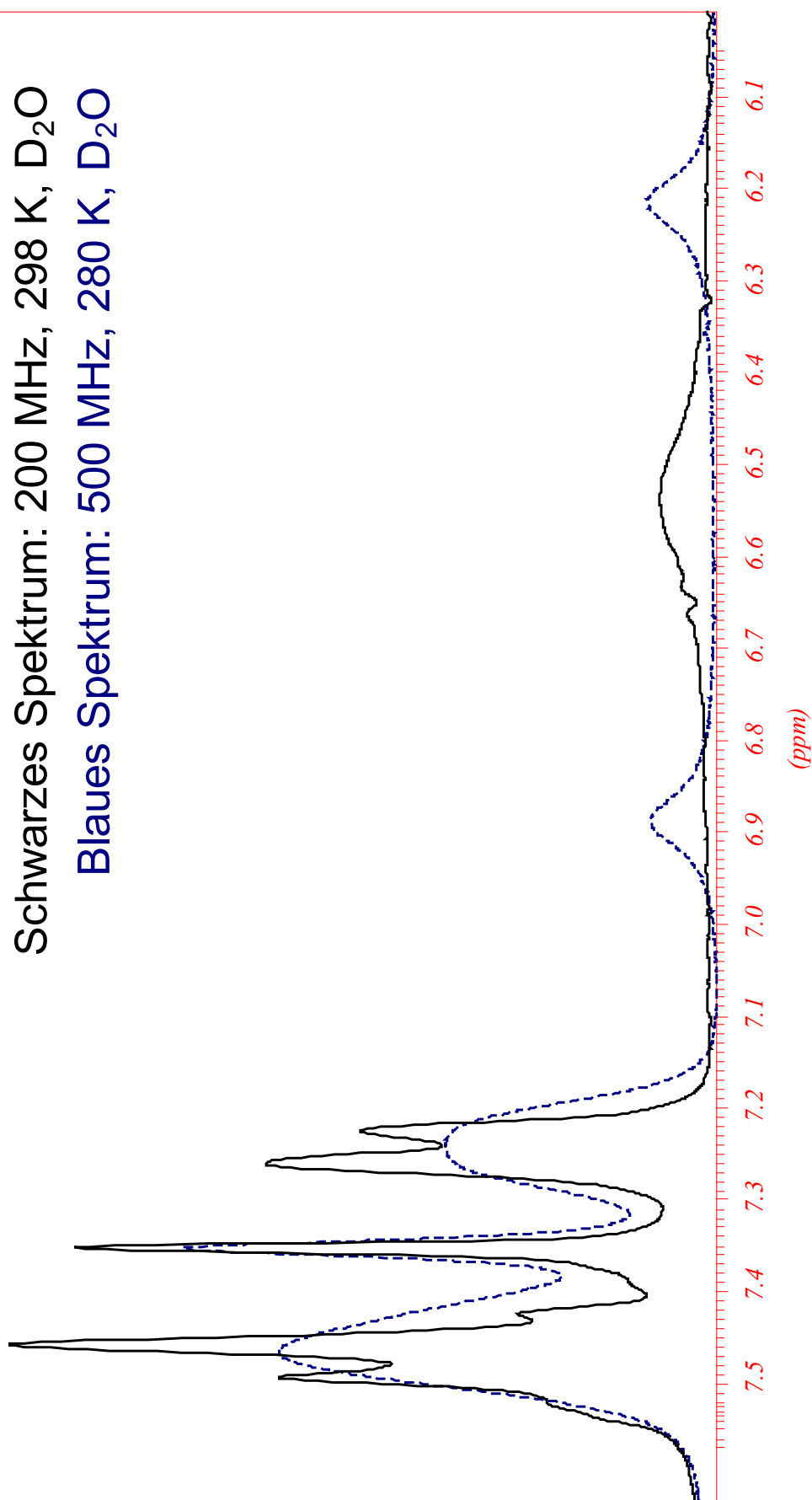


Abbildung B.5: 500 u. 200 MHz <sup>1</sup>H-NMR Spektrum von (R)-(+)-SCH 23390 bei 298 K und 280 K überlagert



C. Anhang – Daten der  
Dopamin-D<sub>1</sub>-, D<sub>2</sub>-, D<sub>4</sub>- und  
D<sub>5</sub>-Rezeptorantagonisten

$D_1$				$D_5$		
Name	$Ca^{2+}$ pK <sub>i</sub>	$\pm$	RLB pK <sub>i</sub>	$Ca^{2+}$ pK <sub>i</sub>	$\pm$	RLB pK <sub>i</sub>
le_300	7,22	0,13	8,64	7,9	0,18	8,11
le_400	-	-	6,29	-	-	5,58
le_403	7,56	0,1	6,47	8,63	0,26	5,97
le_404	8,2	0,14	9,47	8,77	0,33	7,82
le_410	7,39	0,07	8,13	8,51	0,31	7,89
le_420	6,73	0,08	7,77	7,47	0,15	7,13
r-sch23390	9,33	0,21	9,24	9,1	0,11	8,85
aha_d11	-	-	6,58	6,97	0,2	-
b157	6,13	0,13	-	6,95	0,16	-
lan_d5	-	-	6,74	-	-	-
le_ru_301	6,77	0,14	7,55	8,65	0,18	7,89
sh_3	6	0,3	7,21	6,47	0,15	6,44
sh_4	-	-	7	-	-	-

Tabelle C.1: Aktivitätsdaten aus Calcium- ( $Ca^{2+}$ ) und Radioligandbindungsassay (RLB) für  $D_1$ -artige Dopaminrezeptoren

$D_2$				$D_4$		
Name	$Ca^{2+}$ pK <sub>i</sub>	$\pm$	RLB pK <sub>i</sub>	$Ca^{2+}$ pK <sub>i</sub>	$\pm$	RLB pK <sub>i</sub>
le_300	7,93	0,11	7,22	7,9	0,18	7,15
le_400	-	-	-	-	-	5,6
le_403	7,13	0,08	-	8,63	0,26	6,78
le_404	7,71	0,09	7,72	8,77	0,33	7,92
le_410	8,13	0,1	7,16	8,51	0,31	6,94
le_420	7,08	0,11	6,76	7,47	0,15	6,6
r-sch23390	5,59	0,12	-	-	-	-
b157	6,03	0,08	-	-	-	-
lan_d5	6,39	0,12	7,01	-	-	-
le_ru_301	7,08	0,11	6,53	-	-	-
sh_3	6,58	0,07	6,15	-	-	-

Tabelle C.2: Aktivitätsdaten aus Calcium- ( $Ca^{2+}$ ) und Radioligandbindungsassay (RLB) für  $D_2$ -artige Dopaminrezeptoren

AHA D11				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
85	R850000	72.946	35	1.336
54	R570000	73.489	26	1.393
40	R440000	75.280	19	1.455
27	R320000	75.764	17	1.637
58	R600000	77.006	17	1.459
46	R50000	77.096	13	1.565
41	R450000	77.417	9	1.766
68	R70000	77.488	7	1.651
100	R990000	77.805	11	1.603
12	R190000	82.234	14	1.599
86	R860000	82.882	4	1.686
59	R610000	83.259	5	1.760

Tabelle C.3: Die repräsentativen Konformationen von AHA D11 (Schwelle = 1.2 Å)

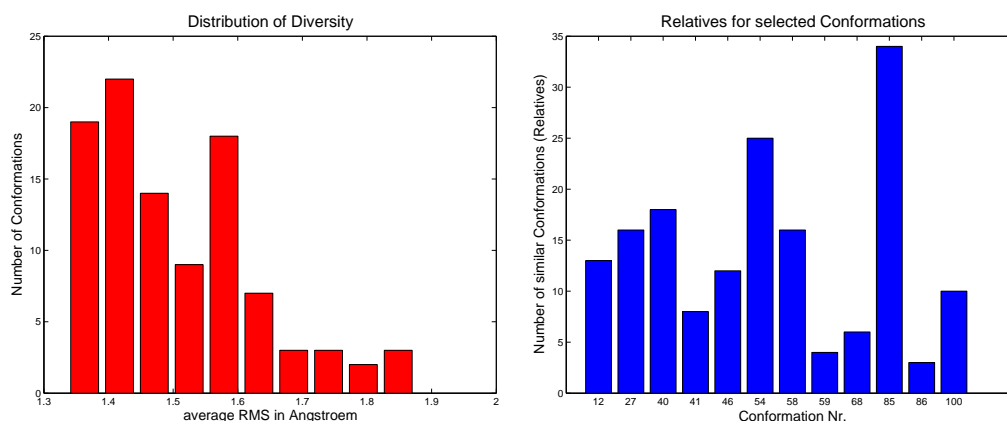


Abbildung C.1: Diversitätsverteilung und ausgewählte Repräsentanten der Verbindung AHA D11

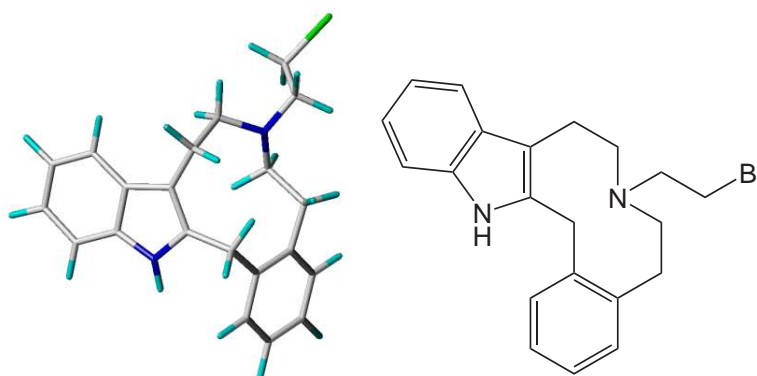


Abbildung C.2: Globale Energieminimumkonformation von AHA D11

(R)-(+)-SCH 23390				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
45	R490000	-1.312	55	0.667
25	R300000	1.190	32	0.919
42	R460000	2.056	27	0.672
94	R930000	4.281	4	1.098

Tabelle C.4: Die repräsentativen Konformationen von (R)-(+)-SCH 23390 (Schwelle = 0.6 Å)

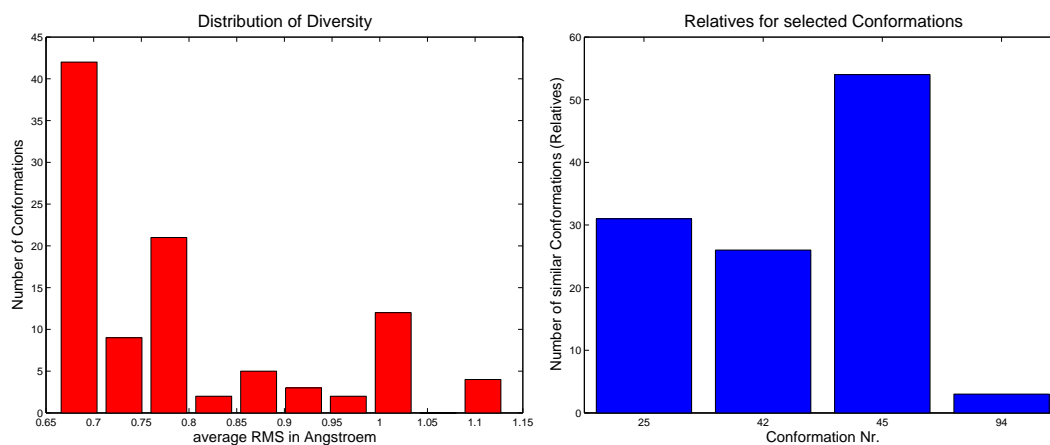


Abbildung C.3: Diversitätsverteilung und ausgewählte Repräsentanten von (R)-(+)-SCH 23390

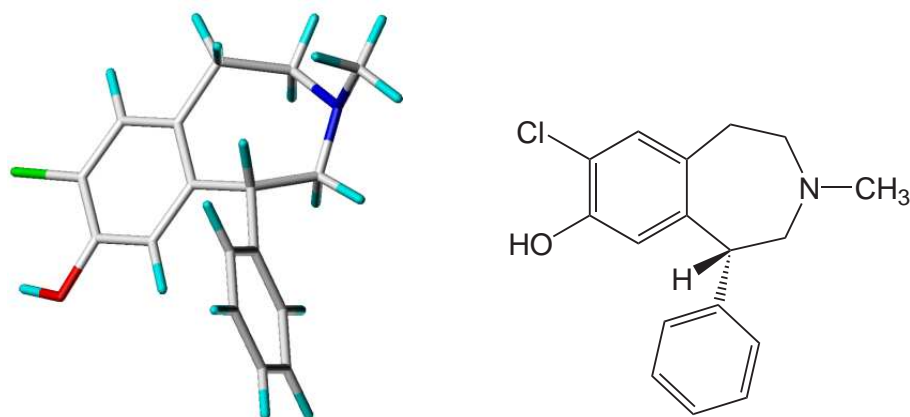


Abbildung C.4: Globale Energieminimumkonformation von (R)-(+)-SCH 23390



LE 300				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
38	R420000	73.695	34	1.110
34	R390000	73.695	42	1.065
93	R920000	76.827	23	1.271
28	R320000	76.843	23	1.223
8	R150000	84.050	14	1.397

Tabelle C.5: Die repräsentativen Konformationen von LE 300 (Schwelle = 1.0 Å)

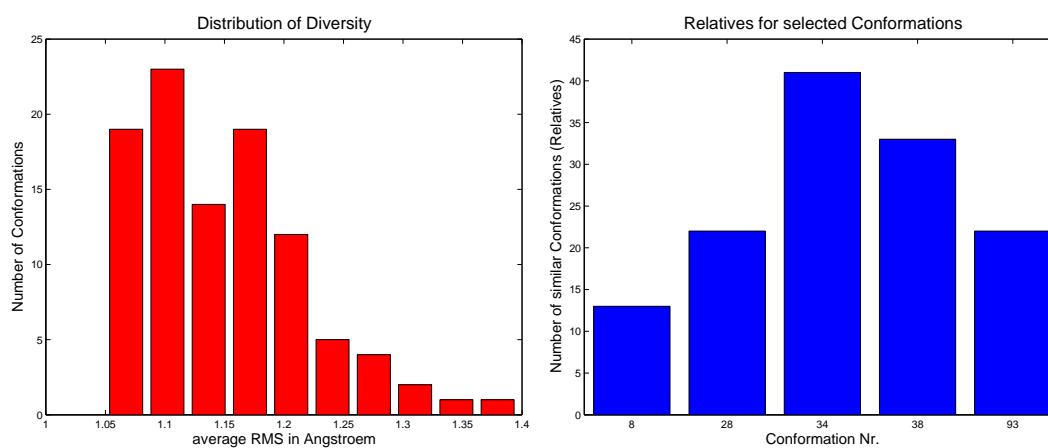


Abbildung C.5: Diversitätsverteilung und ausgewählte Repräsentanten von LE 300

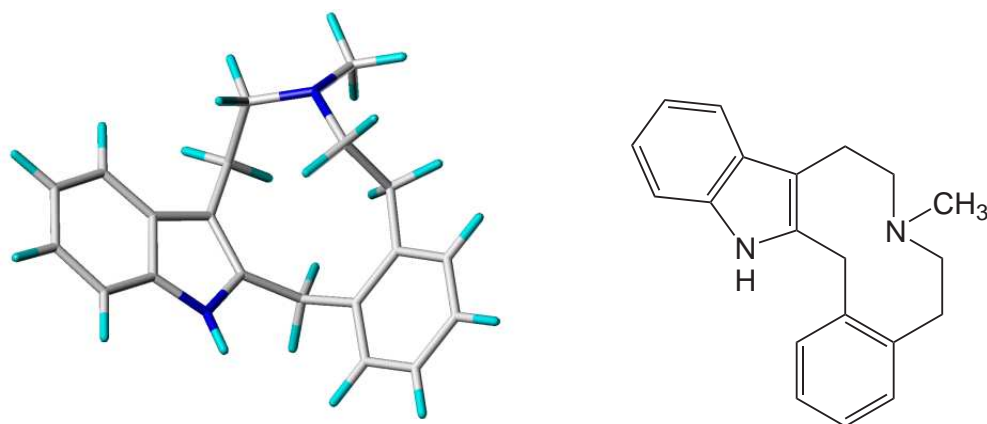


Abbildung C.6: Globale Energieminimumkonformation von LE 300

LE 404				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
45	R490000	0.546	36	1.109
93	R920000	0.546	42	1.060
47	R500000	3.265	25	1.205
84	R840000	3.272	29	1.130
99	R980000	8.103	3	1.476
56	R590000	8.149	7	1.471

Tabelle C.6: Die repräsentativen Konformationen von LE 404 (Schwelle = 1.0 Å)

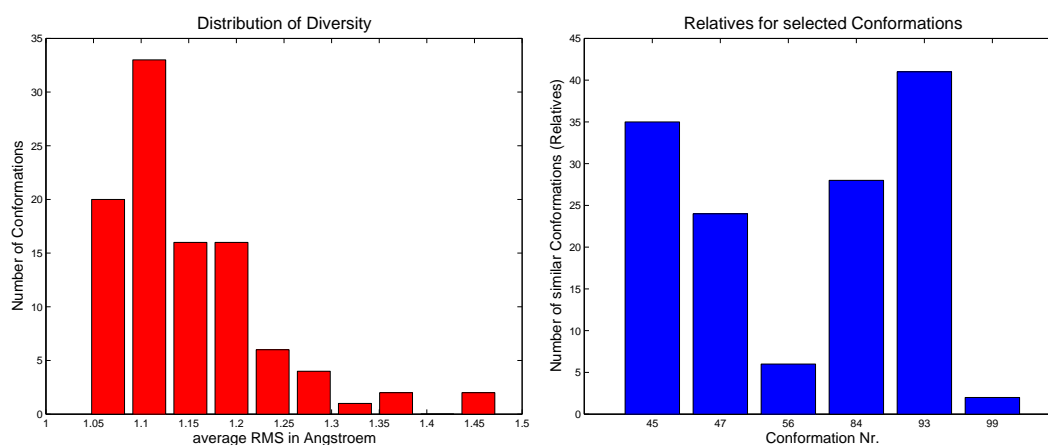


Abbildung C.7: Diversitätsverteilung und ausgewählte Repräsentanten von LE 404

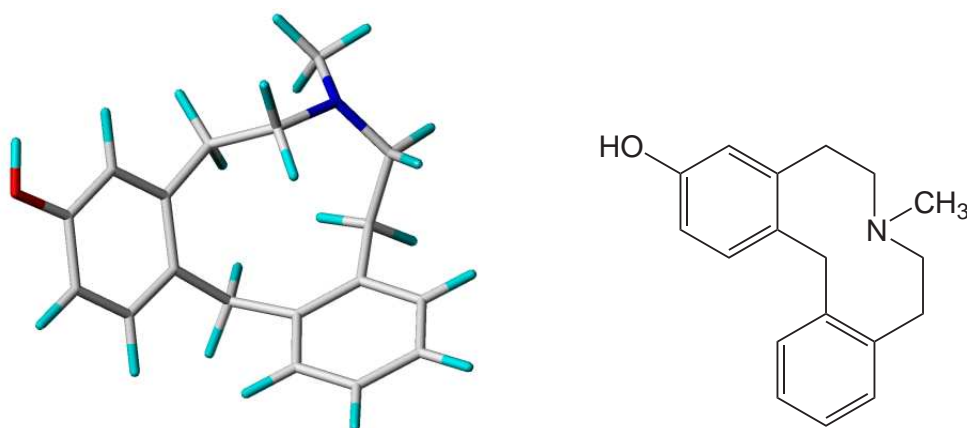


Abbildung C.8: Globale Energieminimumkonformation von LE 404

LE 410				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
38	R420000	44.716	55	0.712
89	R890000	46.261	45	0.786
94	R930000	48.279	44	0.892
57	R60000	52.259	5	1.430
25	R300000	52.276	4	1.423

Tabelle C.7: Die repräsentativen Konformationen von LE 410 (Schwelle = 0.7 Å)

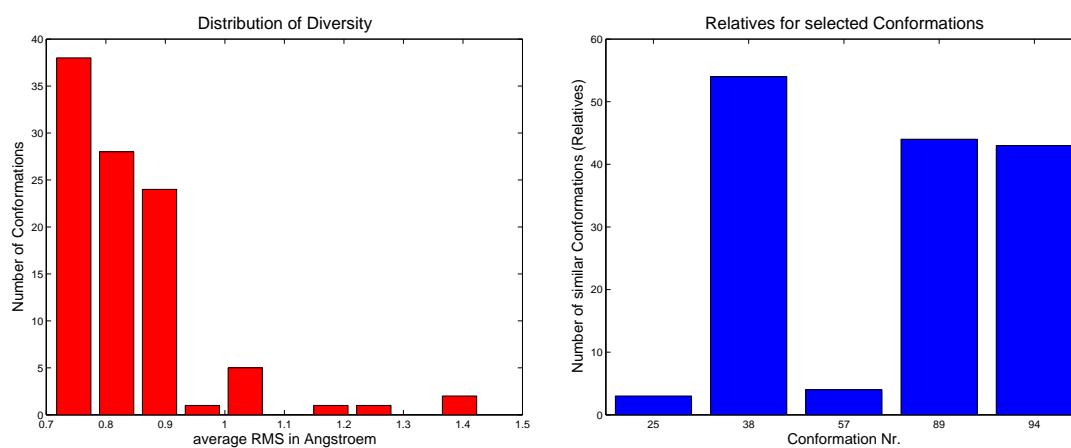


Abbildung C.9: Diversitätsverteilung und ausgewählte Repräsentanten von LE 410

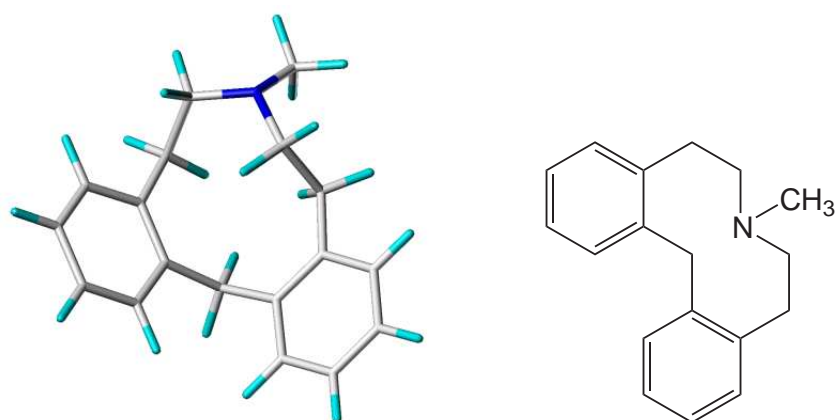


Abbildung C.10: Globale Energieminimumkonformation von LE 410

LE 420				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
36	R400000	47.453	37	0.982
89	R890000	47.453	33	1.021
74	R750000	49.224	30	1.120
96	R950000	50.213	35	1.052
10	R170000	53.660	9	1.231

Tabelle C.8: Die repräsentativen Konformationen von LE 420 (Schwelle = 0.9 Å)

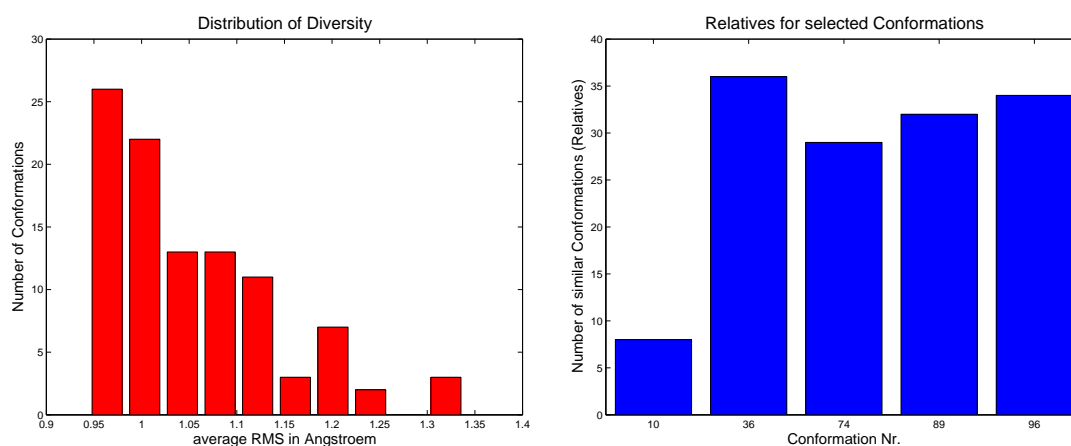


Abbildung C.11: Diversitätsverteilung und ausgewählte Repräsentanten von LE 420

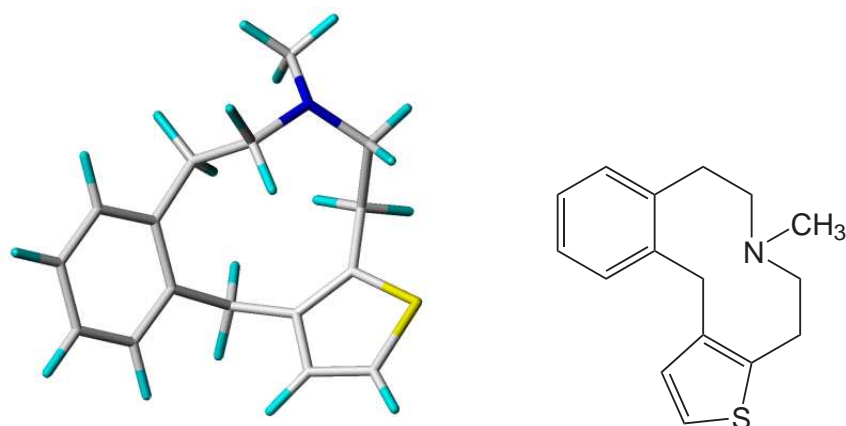


Abbildung C.12: Globale Energieminimumkonformation von LE 420

LERU 301				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
55	R580000	68.372	36	1.239
82	R820000	68.373	42	1.186
84	R840000	71.053	28	1.264
19	R250000	71.061	18	1.379
58	R600000	73.410	16	1.408
18	R240000	77.557	16	1.396
88	R880000	77.903	16	1.460
48	R510000	78.500	3	1.667

Tabelle C.9: Die repräsentativen Konformationen von LERU 301 (Schwelle = 1.1 Å)

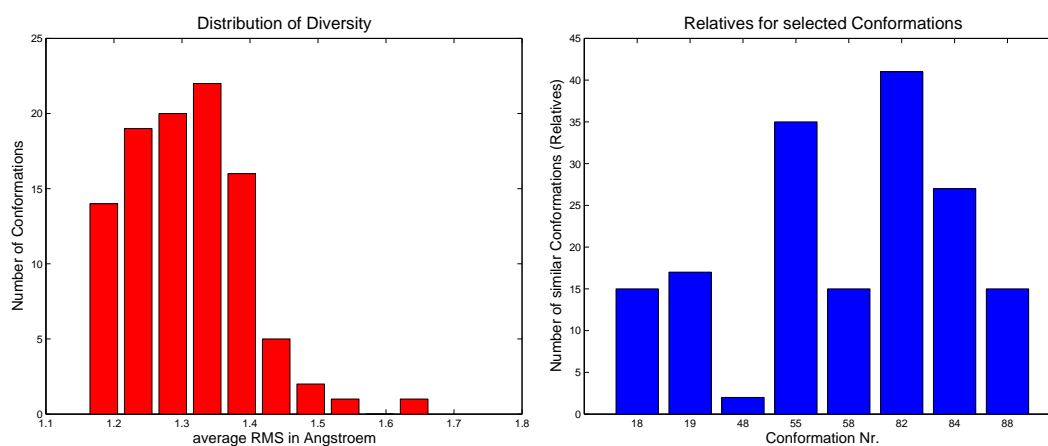


Abbildung C.13: Diversitätsverteilung und ausgewählte Repräsentanten von LERU 301

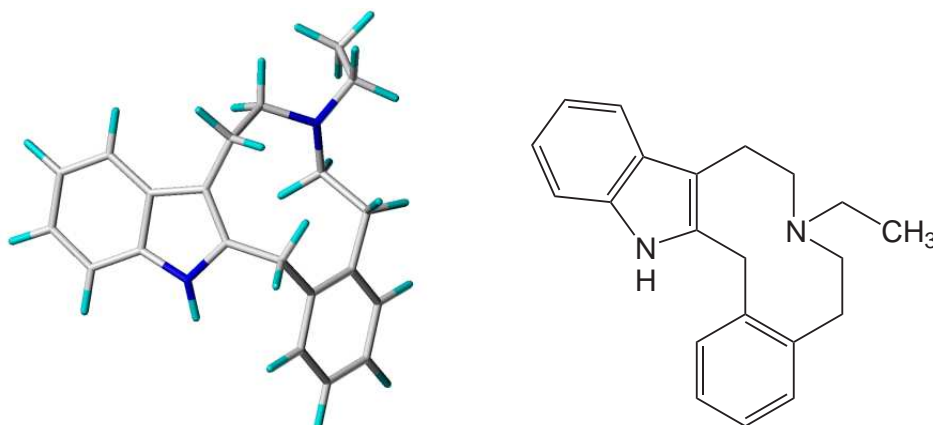


Abbildung C.14: Globale Energieminimumkonformation von LERU 301

SH 3				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
19	R250000	63.742	38	0.997
11	R180000	63.747	37	0.985
58	R600000	65.097	32	1.118
86	R860000	65.111	22	1.243
66	R680000	72.739	17	1.259

Tabelle C.10: Die repräsentativen Konformationen von SH 3 (Schwelle = 0.9 Å)

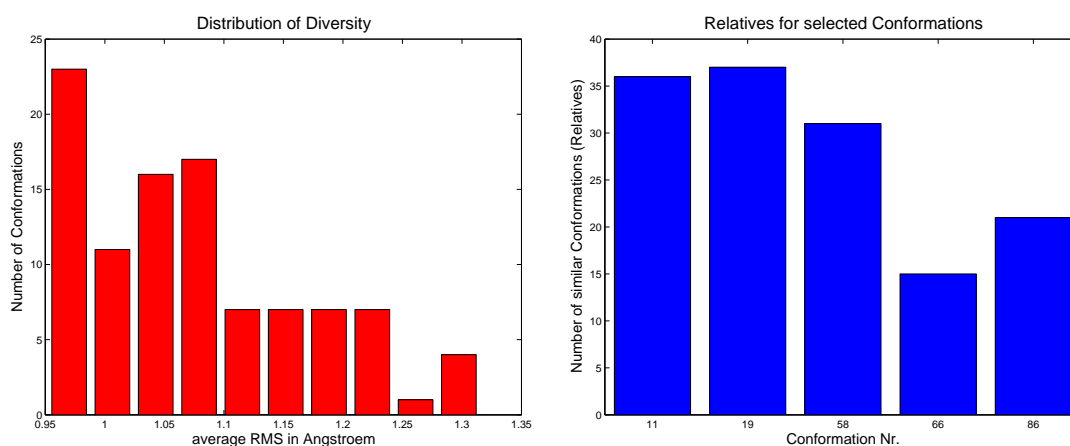


Abbildung C.15: Diversitätsverteilung und ausgewählte Repräsentanten von SH 3

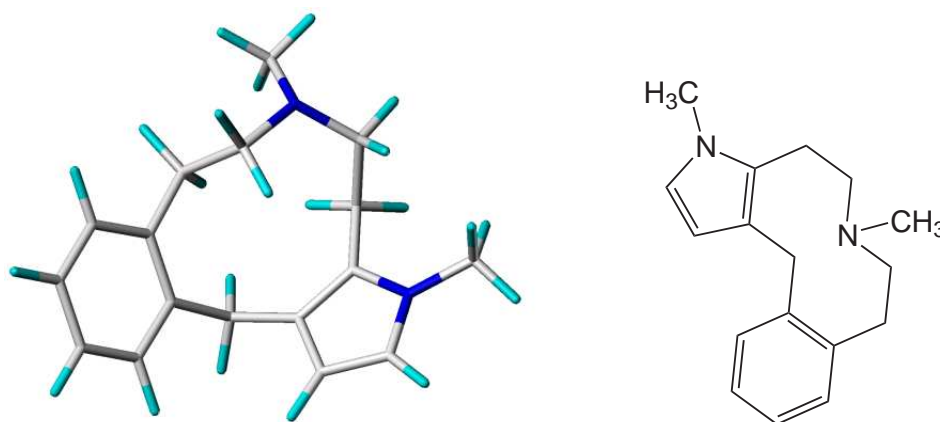


Abbildung C.16: Globale Energieminimumkonformation von SH 3

LE 403				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
31	R360000	-43.409	48	1.098
53	R560000	-43.319	40	1.151
39	R430000	-40.949	19	1.291
66	R680000	-39.940	27	1.360
91	R900000	-33.514	19	1.426

Tabelle C.11: Die repräsentativen Konformationen von LE 403 (Schwelle = 1.1 Å)

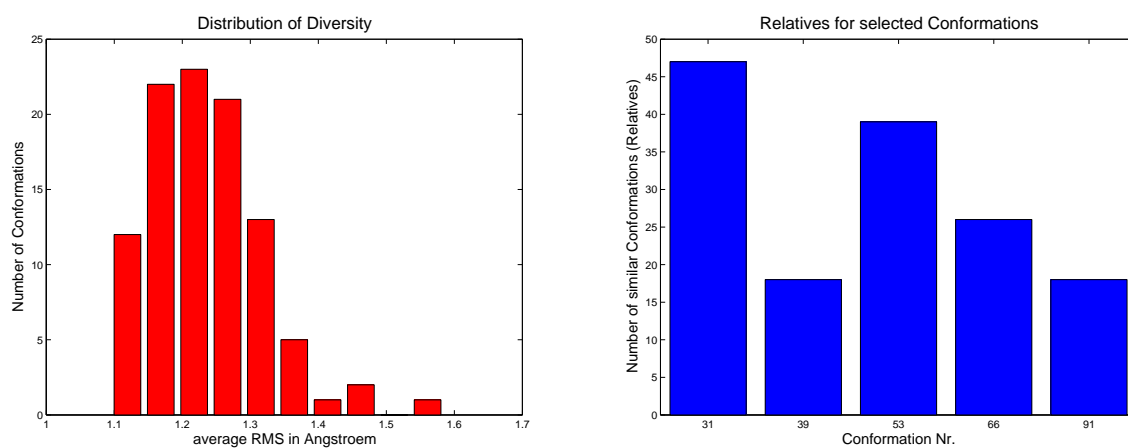


Abbildung C.17: Diversitätsverteilung und ausgewählte Repräsentanten von LE 403

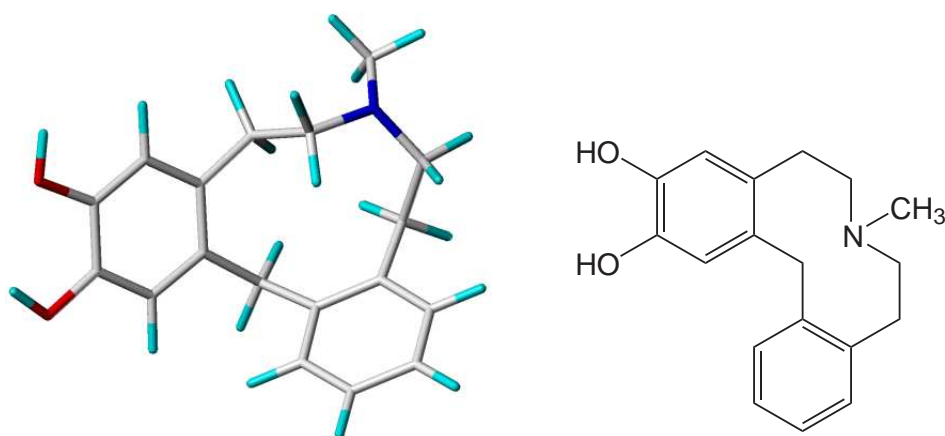


Abbildung C.18: Globale Energieminimumkonformation von LE 403

LE 400				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
35	R40000	-27.925	41	1.278
47	R500000	-27.696	51	1.183
86	R860000	-23.959	32	1.398
43	R470000	-21.647	21	1.496
79	R80000	-20.155	3	1.644
48	R510000	-18.777	15	1.635

Tabelle C.12: Die repräsentativen Konformationen von LE 400 (Schwelle = 1.2 Å)

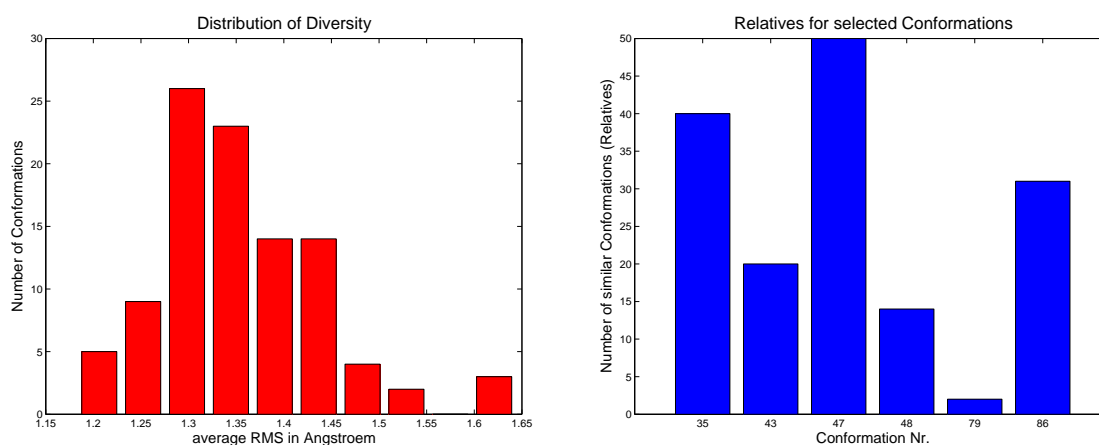


Abbildung C.19: Diversitätsverteilung und ausgewählte Repräsentanten von LE 400

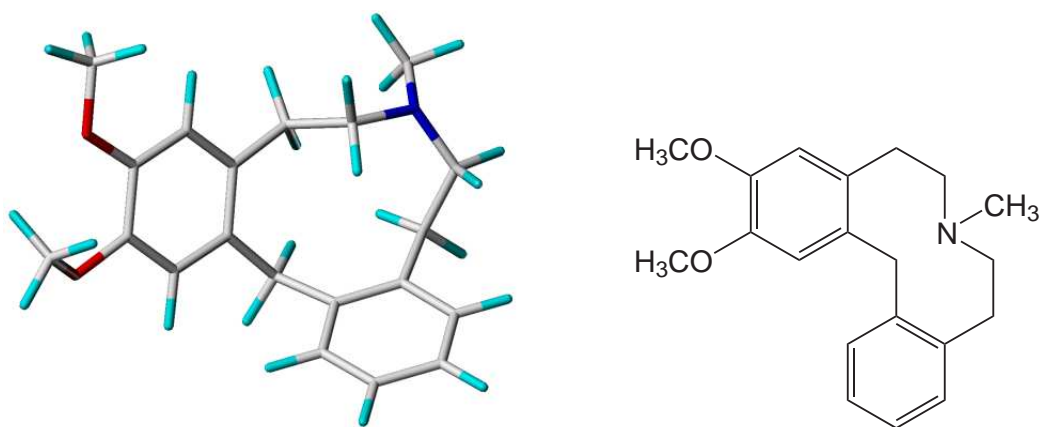


Abbildung C.20: Globale Energieminimumkonformation von LE 400



(-)2a SCH 39166 (SS)				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
181	R810000	-2.733	110	1.004
77	R1680000	-0.027	90	1.130

Tabelle C.13: Die repräsentativen Konformationen von (-)2a SCH 39166 (Schwelle = 1.0 Å)

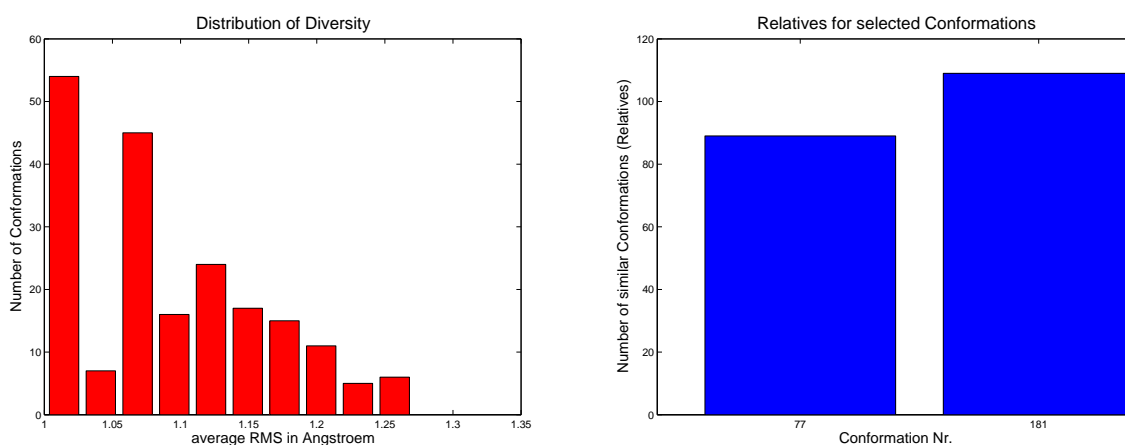


Abbildung C.21: Diversitätsverteilung und ausgewählte Repräsentanten von (-)2a SCH 39166

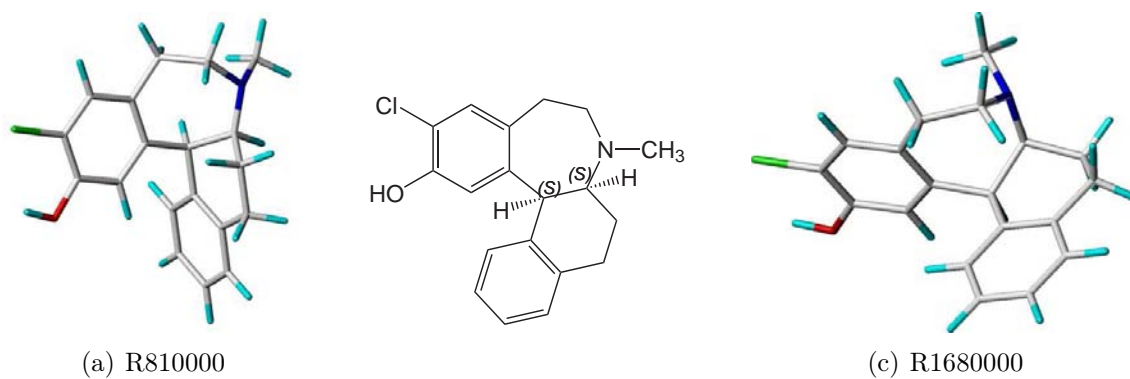


Abbildung C.22: Gefundene repräsentative Konformationen von (-)2a SCH 39166

(-)2b SCH 39166 (RS)				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
76	R1670000	-3.715	134	0.670
102	R1900000	-0.938	62	1.005
133	R380000	12.098	12	1.010

Tabelle C.14: Die repräsentativen Konformationen von (-)2b SCH 39166 (Schwelle = 0.6 Å)

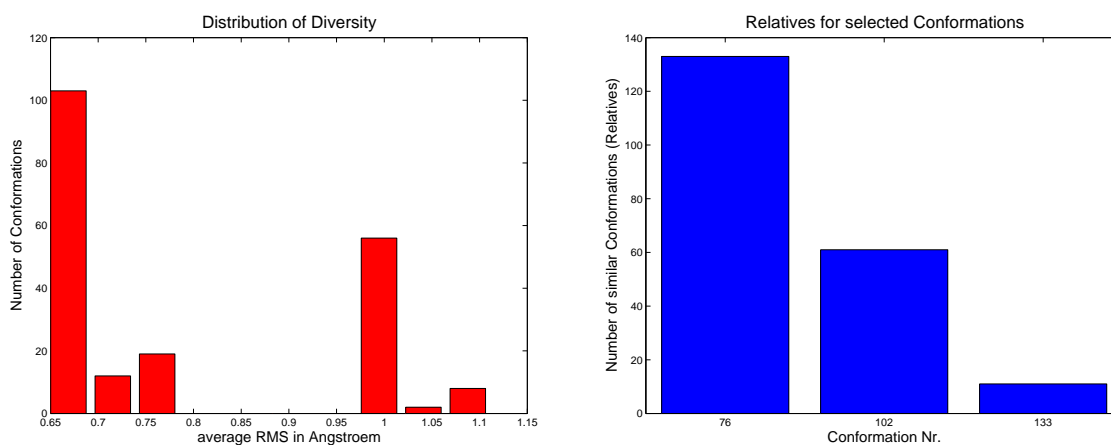


Abbildung C.23: Diversitätsverteilung und ausgewählte Repräsentanten von (-)2b SCH 39166

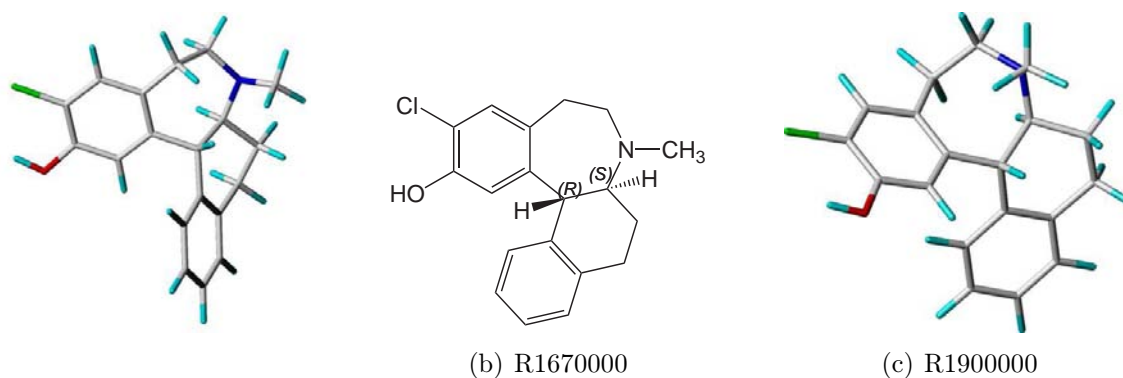


Abbildung C.24: Die häufigsten Konformationen von (-)2b SCH 39166 – Globale Energieminimumkonformation links.

(+)2a SCH 39166 (RR)				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
4	R1010000	-2.733	109	1.009
160	R620000	-0.027	91	1.127

Tabelle C.15: Die repräsentativen Konformationen von (+)2a SCH 39166 (Schwelle = 1.0 Å)

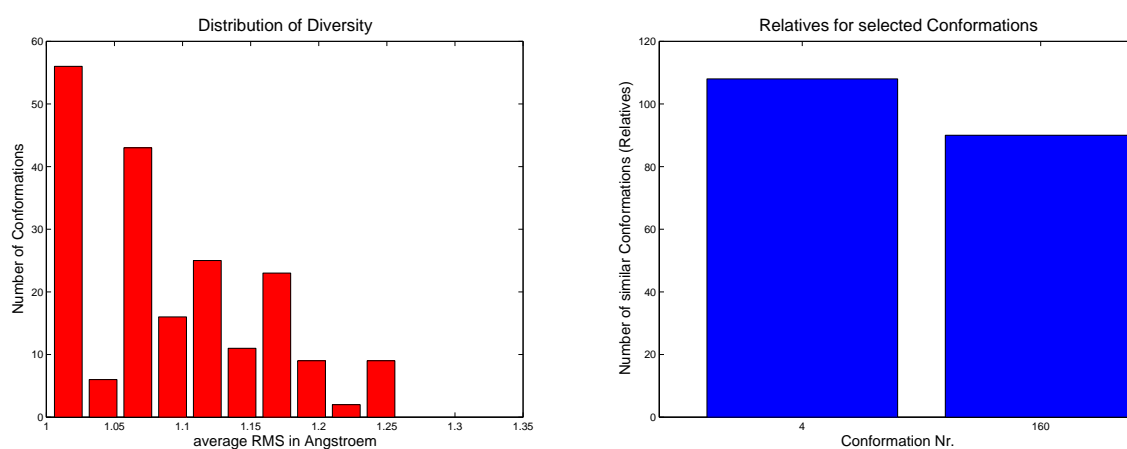


Abbildung C.25: Diversitätsverteilung und ausgewählte Repräsentanten von (+)2a SCH 39166

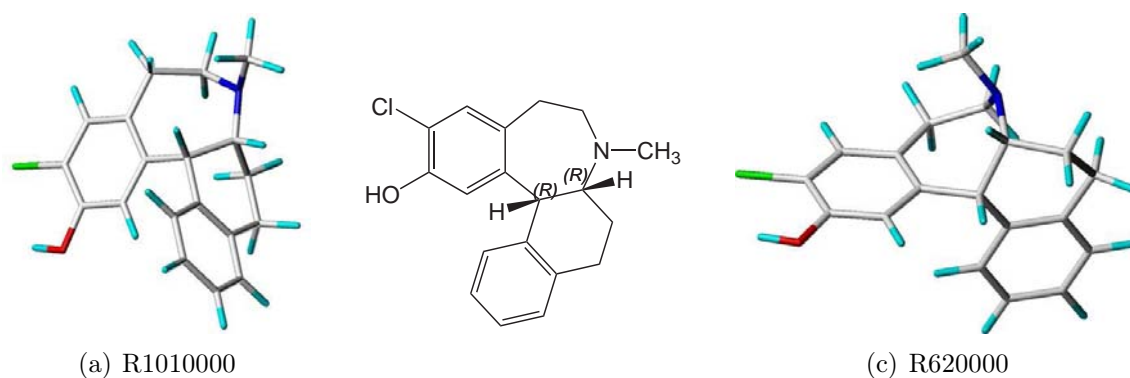


Abbildung C.26: Gefundene repräsentative Konformationen von (+)2a SCH 39166 – Globale Energieminimumkonformation links.

(+)2b SCH 39166 (SR)				
Nr.	Konformation	HoF (kcal)	Häufigkeit	Ø RMS
55	R1480000	-3.715	147	0.579
9	R1060000	-0.938	50	1.096
29	R1240000	10.242	1	1.001
71	R1620000	12.098	11	1.059

Tabelle C.16: Die repräsentativen Konformationen von (+)2b SCH 39166 (Schwelle = 0.6 Å)

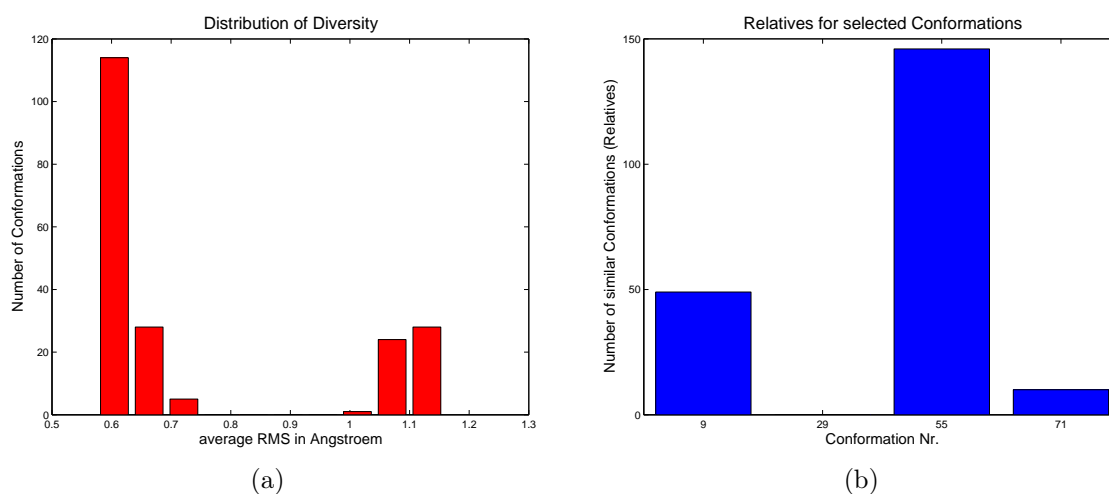


Abbildung C.27: Diversitätsverteilung und ausgewählte Repräsentanten von (+)2b SCH 39166

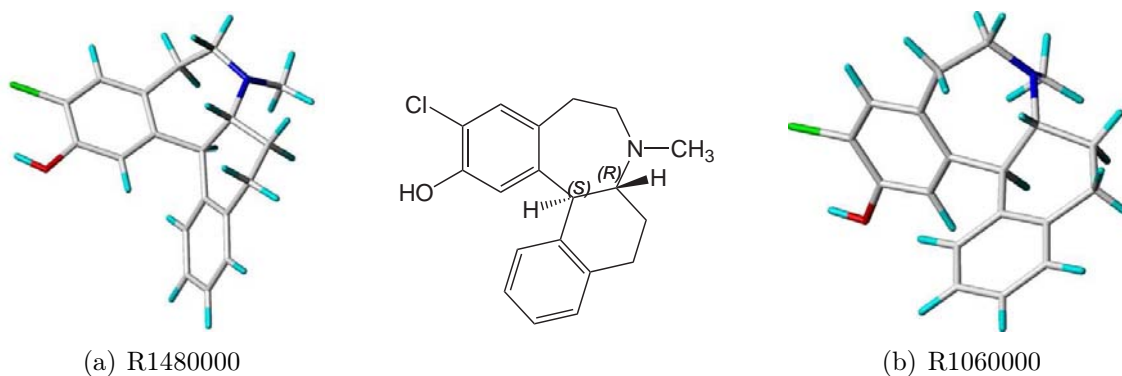


Abbildung C.28: Die häufigsten Konformationen von (+)2b SCH 39166 – Globale Energieminimumkonformation links.

# D. Anhang Programme und Skripte

## D.1 Das Programm PLS-Toolbox

Die bei einer PLS-Regression zu beachtenden Parameter sind sehr zahlreich. Es gibt z. B. verschiedene Möglichkeiten der Datenvorbehandlung wie die Zentrierung um den Nullpunkt oder die Skalierung auf eine gemeinsame Standardabweichung. Durch diese Vorbehandlung kann das Ergebnis der Regression beeinflusst werden. A. Höskuldsson diskutiert in [119] die Notwendigkeit und den Einfluss der Skalierung auf Regressionsmethoden wie der PLS. Ein weiterer Einflussfaktor auf das Ergebnis der PLS-Regression entsteht bei Einsatz der Kreuzvalidierung (CV). Man kann z. B. die X- und Y-Werte während der Kreuzvalidierung neu skalieren. Ob der Mittelwert der X- bzw. Y-Werte während der Kreuzvalidierung konstant bleibt oder jedes Mal neu berechnet wird, beeinflusst ebenfalls den  $q^2$ -Wert. Der Einfluss der Skalierung während der Kreuzvalidierung wird in [120] anhand eines Beispiels beschrieben.

Diese Möglichkeit der Beeinflussung der Kreuzvalidierung ist jedoch bei kaum einem Statistikprogramm vorhanden. Je nach verwendetem Algorithmus können diese Programme so zu unterschiedlichen Ergebnissen gelangen.

Um den größtmöglichen Einfluss auf solche Parameter bei routinemäßigen PLS-Analysen zu behalten, wurde die PLS-Toolbox entwickelt. Diese vereint unter einer

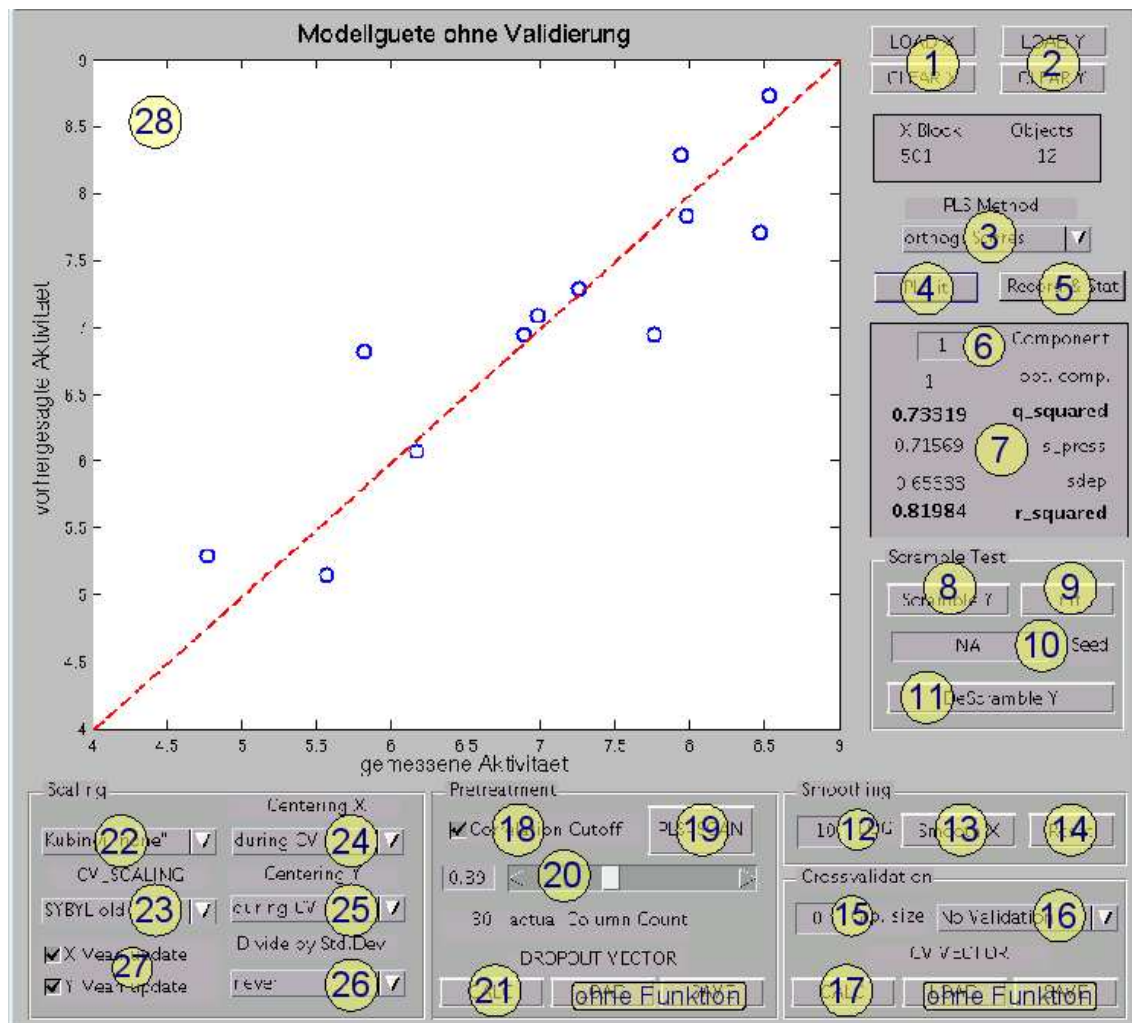


Abbildung D.1: Programmfenster der PLS-Toolbox; die markierten Programmfunktionen werden im Text erläutert

intuitiv zu bedienenden graphischen Oberfläche die wichtigsten Möglichkeiten der Datenmanipulation vor und während der PLS-Regression mit Kreuzvalidierung. Die wesentlichen Programmfähigkeiten werden im folgenden Text anhand der in Abbildung D.1 markierten Elemente erläutert.

#### 1 LOADX/CLEARX

- öffnet den Dateidialog
- durch Doppelklick wird die Datei mit den (unabhängigen) X-Variablen geladen
- CLEARX löscht geladene X-Variablen

#### 2 LOADY/CLEARY

- siehe **1**

### **3** PLS-Method

- Auswahl der PLS-Methode
- momentan sind die Methoden „orthogonal Scores“ und „NIPALS“ implementiert

### **4** PLS it

- Durchführung der PLS-Analyse mit allen aktuellen Einstellungen

### **5** Record & Stat

- von aufeinanderfolgenden PLS-Analysen (**PLS it**) oder Korrelationsschwellenwertanalysen (**PLS-SCAN**) werden Mittelwert, Median und Standardabweichung im Konsolenfenster ausgedruckt

### **6** Component

- Eingabe der gewünschten Komponentenzahl
- wirkt sich auf die dargestellten Ergebnisse einer aktuellen oder zukünftigen PLS-Einzelanalyse aus (nicht auf Korrelationsschwellenwertanalysen)

### **7** Statistische Kennzahlen der aktuellen PLS-Analyse oder der besten PLS einer Korrelationsschwellenwertanalyse

**opt. comp.** optimale Anzahl der PLS-Komponenten (Kriterium ist der minimale SPRESS-Wert)

**q\_squared**  $q^2$ -Wert (kreuzvalidierter Regressionskoeffizient)

**s\_press** SPRESS Fehlerwert

**sdep** SDEP Fehlerwert

**r\_squared**  $r^2$ -Wert: nicht validierter Regressionskoeffizient (Anzeige nur bei Durchführung einer PLS-Analyse mit Einstellung „No Validation“ bei **16**)

### **8** Scramble Y

- bringt die (abhängigen) Y-Daten in eine zufällige Reihenfolge

#### 9 Init

- initialisiert den Zufallsgenerator mit einem aus der aktuellen Uhrzeit abgeleiteten Wert (Seed)
- dieser Wert wird im Eingabekasten Seed angezeigt

#### 10 Seed

- Anzeige des für die Initialisierung verwendeten Wertes
- um vorangegangene Resultate zu reproduzieren, kann ein selbstbestimmter Wert eingegeben werden

#### 11 DeScramble Y

- bringt die Y-Werte wieder in die ursprüngliche (der geladenen Datei entsprechenden) Reihenfolge

#### 12 Rundungsgenauigkeit

- Eingabe des dekadischen Logarithmus der Rundungsgenauigkeit

#### 13 Smooth X

- runden der X-Daten auf die bei **12** eingegebene Zahl von Nachkommastellen

#### 14 Reset

- stellt die ursprünglichen (der geladenen Datei entsprechenden) X-Werte wieder her

#### 15 Grp. size

- Eingabe der Gruppengröße für eine Leave-many-out-Kreuzvalidierung
- erfordert die anschließende Betätigung der Schaltfläche CALC **17**

#### 16 Einstellung des Validierungsmodus



**Leave-one-out** bei Drücken der Schaltfläche PLS it 4 wird eine PLS-Analyse mit LOO-Kreuzvalidierung durchgeführt

**Leave-many-out** Einstellung für LMO-Kreuzvalidierung

**No Validation** es wird keine Validierung durchgeführt (vorher muss bereits einmal eine Validierung durchgeführt worden sein)

## 17 CALC

- stellt die Kreuzvalidierungsgruppen zusammen
- bei LMO-Kreuzvalidierung ist das Ergebnis zufällig, aber bei Setzen eines definierten Random Seeds ( 10 ) reproduzierbar
- ist die letzte Gruppe nicht vollständig (weil die Objektzahl nicht glatt durch die Gruppengröße teilbar ist), wird sie mit den ersten Mitgliedern der ersten Gruppe aufgefüllt

## 18 Correlation Cutoff

- schaltet die Verwendung eines Korrelationsschwellenwertes an bzw. aus

## 19 PLS-SCAN

- führt eine Korrelationsschwellenwertabtastung durch, d. h. für jeden Schwellenwert von 0,0-0,99 (Schrittweite 0,01) wird eine PLS-Analyse durchgeführt
- Die statistischen Kennzahlen der besten dieser Analysen werden in der Konsole sowie in 7 angezeigt, zusätzlich wird der Mittelwert mit Standardabweichung und der Median der Kennzahlen berechnet
- ist Record & Stat 5 aktiviert, wird ab der dritten Wiederholung die Abtastung insgesamt für die Mittelwert- und Medianbildung herangezogen

## 20 Korrelationsschwellenwert

- durch Verstellung des Schiebereglers oder direkte Eingabe kann der Korrelationsschwellenwert eingestellt werden
- nur Variablen, die mit einem  $r^2 > \text{Schwellenwert}$  mit Y korrelieren, werden bei einer PLS-Analyse benutzt, wenn 18 angeschaltet wurde

## 21 CALC

- berechnet den Vektor der von der PLS-Analyse auszuschließenden Variablen (deren  $r^2$  kleiner ist, als der Korrelationsschwellenwert)
- die tatsächliche Anzahl der in der PLS-Analyse verwendeten Variablen wird bei „actual Column Count“ angezeigt

**22** Einstellung der Scaling-Variante — noch nicht voll funktional

**23** Einstellung des Scaling während der Kreuzvalidierung — noch nicht voll funktional

**24** Centering X

Einstellung, ob und wann die X-Daten um ihren Mittelwert zentriert werden sollen:

**during CV** während der Kreuzvalidierung

**once befor CV** einmal vor der Kreuzvalidierung

**never** keine Zentrierung

**25** Centering Y

- siehe Centering X **24**

**26** Divide by Std.Dev

- Einstellung, ob und wann die Daten durch die Standardabweichung dividiert werden sollen
- siehe auch Centering X **24**

**27** Mittelwertneuberechnung

**X Mean update** Einstellung, ob die Mittelwerte der X-Werte während der Kreuzvalidierung neu berechnet werden

**Y Mean update** Einstellung, ob der Mittelwert der Y-Werte während der Kreuzvalidierung neu berechnet wird

## D.2 Das SYBYL-Skript match\_all

---

```

uims define macro mad_match_all sybylbasic yes

setvar old_timeout $cgq_timeout
set cgq_timeout 0

setvar outputstr %prompt("string" "match_res" \
  "Outputfile (will be overwritten !)")
setvar flx_line %prompt("string" "*-<h>" \
  "Atomexpressions to be matched")

setvar atoms_to_match $flx_line
setvar outputfile %cat($outputstr ".dat")
if %dir($outputfile)
  setvar schrott %file_delete(%files(file $outputfile))
endif

setvar mol_contents %mols(*)
setvar num_mol %count($mol_contents)

for i in %range(1 %math($num_mol-1))
  setvar i_mol %arg($i $mol_contents)
  for j in %range(%math($i+1) $num_mol)
    setvar j_mol %arg($j $mol_contents)
    match $i_mol($atoms_to_match) $j_mol($atoms_to_match)
    echo $i_mol " " $j_mol " " $match_rms
    setvar props %cat($i " " %mol_info($i_mol name) " " $j " \
      " %mol_info($j_mol name) " " $MATCH_RMS)
    setvar schrott %system("echo $props >> $outputfile")
  endfor
endfor

setvar CGQ_timeout $old_timeout
.
```

## D.3 Das Programm conf\_elecT

```

%%
%% by Mad 02/2002 version 112004
%%
%% function that gets a list of RMS-values from the SYBYL match-
%% algorithm (produced by mad_match_all.spl) and eliminates the ones
%% which are similar to each other.
%% The purpose is to remain the most diverse conformations from e.g.
%% simulated annealing
%%
function [Elected] = conf_elecT(Match_rms_list,Treshold,Erg_list,Relative); 10

Subtract_electrostatic = 0;
if nargin < 2
    Treshold = 0.4;
    Modop = 1;
    %% no energy modus !
elseif nargin == 4
    if Relative == 0
        Modop = 2;
        Subtract_electrostatic = 0; 20
    else
        Modop = 3;
        %% family modus !
    end
elseif nargin == 3
    Modop = 2;
    %% energy modus !
elseif nargin == 2
    Modop = 1;
    %% no energy modus ! 30
end
Elected = [];
% Load Match-RMS-List
if exist(Match_rms_list,'file') == 2
    Rms_file = fopen(Match_rms_list,'r');
    Rms_list = textscan(Rms_file,'%f %s %f %s %f');
else
    disp('MATCH_RMS-List not found!');
    Elected = strvcats([Match_rms_list ' not found ! ']);
    return 40
end
% Get the number of conformations
Num_conf = Rms_list{3}(size(Rms_list{3},1));
Rms_mat = zeros(Num_conf);
% Build the upper triangular matrix (half of the symmetric matrix)
Help_count = 1;
for i = 1:Num_conf
    Rms_mat(i,i+1:Num_conf) = Rms_list{5}(Help_count:Help_count+Num_conf-i-1);
    if i < Num_conf
        Conformation_names{i} = Rms_list{2}(Help_count); 50
    else
        Conformation_names{i} = Rms_list{4}(size(Rms_list{2},1));
    end
    Help_count = Help_count + Num_conf - i;
end
% Build the lower part of the symmetric matrix
Rms_mat = Rms_mat + Rms_mat';
if Modop == 2
    %% energy modus !
    if exist(Erg_list,'file') == 2 60
        Energy_list = load(Erg_list,'-ascii');
        if Num_conf ~= size(Energy_list,1)
            disp('Energy-List not ok !');
            Elected = strvcats([Erg_list ' not ok ! ']);
        end
    end
end

```

```

    if size(Energy_list,2) > 1
        % we have extra electrostatic energy values and we want subtract them
        % from the total energy
    if Subtract_electrostatic == 1
        Energy_list_total = Energy_list(:,1);
        Energy_list = Energy_list(:,1) - Energy_list(:,2) - Energy_list(:,3);
    else
        Energy_list_total = Energy_list(:,1) - Energy_list(:,2) - Energy_list(:,3);
        Energy_list = Energy_list(:,1);
    end
    Electrostatic_subtracted = 1;
else Electrostatic_subtracted = 0;
end
else
    disp('Energie-Liste nicht gefunden !');
    Elected = strvcats([Erg_list ' not found ! ']);
    return
end
[List_sort, List_idx] = sort(Energy_list);
end
%% Sum up the RMS-Values for each Conformation
Rms_sum = sum(Rms_mat);
%% Normalize by making the average
Rms_sum_norm = Rms_sum./Num_conf;
%% plot the histogram for distribution of diversity
figure(1);
[Div_freq, Div_loc] = hist(Rms_sum_norm);
bar(Div_loc, Div_freq)
h = findobj(gca, 'Type', 'patch');
set(h, 'FaceColor', 'r', 'EdgeColor', 'k')
title('Distribution of Diversity', 'FontSize', 16)
xlabel('average RMS in Angstroem', 'FontSize', 14)
ylabel('Number of Conformations', 'FontSize', 14)
Conf_idx = [1:Num_conf]';
Sim = [];
Sim_num_all = zeros(1, Num_conf);
for i = 1:Num_conf
    Sim = find(Rms_mat(i,:) < Treshold);
    Sim_conf_all{i} = Sim;
    Sim_num_all(i) = size(Sim,2);
end
if Modop == 1
    %% this seems to be a better representation of the similarity for each
    %% selected conformation because the last one isn't the looser
    [List_sort, List_idx] = sort(Sim_num_all);
    List_idx = flipud(List_idx);
end
if Modop == 3
    fprintf(1, 'Number of Relatives for %g : %g \n', [Relative Sim_num_all(Relative)]);
    fprintf(1, 'Relatives for: %3g \n', Relative);
    fprintf(1, '%g ', Sim_conf_all{Relative});
end
i = 1;
Sim = [];
while i <= Num_conf
    Sim = find(Rms_mat(find(Conf_idx == List_idx(i)), :) < Treshold);
    if size(Sim,2) == 1
        Sim = [];
    else
        Diagpos = find(Sim == find(Conf_idx == List_idx(i)));
        Sim(Diagpos) = [];
    end
    if ~ isempty(Sim)
        Rms_mat(:, Sim) = [];
        Rms_mat(Sim, :) = [];
        for j = 1:size(Sim,2)
            List_idx(find(List_idx == Conf_idx(Sim(j)))) = [];

```

```

        end
        Conf_idx(Sim) = [];
        Num_conf = size(Rms_mat,1);
    end
    Sim=[];
    i = i+1;
end
%% plot the histogram of the Distribution of Similarity
figure(2);
[Sim_freq, Sim_loc]=hist(Sim_num_all);
bar(Sim_loc,Sim_freq)
h = findobj(gca,'Type','patch');
set(h,'FaceColor','r','EdgeColor','k')
title('Distribution of Similarity','FontSize',16)
xlabel('Number of Relatives','FontSize',14)
ylabel('Number of Conformations','FontSize',14)
%% plot the number of relatives for all conformations
figure(3);
bar(1:size(Sim_num_all,2),Sim_num_all);
h = findobj(gca,'Type','patch');
set(h,'FaceColor','b','EdgeColor','w')
axis tight
title('Relatives for all Conformations','FontSize',16)
xlabel('Conformation Nr.','FontSize',14)
ylabel('Number of similar Conformations (Relatives)','FontSize',14)
%% plot the number of relatives for the selected conformations
figure(4);
bar(Sim_num_all(Conf_idx));
set(gca,'Xtick',1:size(Conf_idx,1))
set(gca,'Xticklabel',Conf_idx)
set(gca,'XticklabelMode','manual')
h = findobj(gca,'Type','patch');
set(h,'FaceColor','b','EdgeColor','k')
title('Relatives for selected Conformations','FontSize',16)
xlabel('Conformation Nr.','FontSize',14)
ylabel('Number of similar Conformations (Relatives)','FontSize',14)
Sum_conf=Rms_sum_norm(Conf_idx);
%% take the average RMS from the previously calculated vector (only for
%% selected conformations)
Num_conf=1:size(Rms_mat,2);
if Modop == 2
    if Electrostatic_subtracted == 1
        Summen = [Num_conf' Sum_conf' Conf_idx Energy_list(Conf_idx) ...
            Energy_list_total(Conf_idx) Sim_num_all(Conf_idx)'];
    else
        Summen = [Num_conf' Sum_conf' Conf_idx Energy_list(Conf_idx) ...
            Sim_num_all(Conf_idx)'];
    end
    disp('Sortiert nach Nummer: ');
    fprintf(1,'\n');
    for i = 1:size(Num_conf,2)
        fprintf(1,'%s ',char(Conformation_names{Conf_idx(i)}));
        if Electrostatic_subtracted == 1
            fprintf(1,'%3g %.3f %3g %.3f %.3f %3g\n', Summen(i,:));
        else
            fprintf(1,'%3g %.3f %3g %.3f %3g\n', Summen(i,:));
        end
    end
    fprintf(1,'\n');
    [Sum_sort,Sum_sort_idx]=sort(Sum_conf);
    Sum_sort=flipud(Sum_sort');
    Sum_sort_idx=flipud(Sum_sort_idx');
    if Electrostatic_subtracted == 0
        Elected = [Sum_sort Conf_idx(Sum_sort_idx) Energy_list(Conf_idx...
            (Sum_sort_idx)) Sim_num_all(Conf_idx(Sum_sort_idx,1))' ];
    else

```

```

        Elected = [Sum_sort Conf_idx(Sum_sort_idx) Energy_list(Conf_idx...
        (Sum_sort_idx)) Energy_list_total(Conf_idx...
        (Sum_sort_idx)) Sim_num_all(Conf_idx(Sum_sort_idx,1))' ];
    end
    disp('Sortiert nach durchschnittl. RMS: ');
    fprintf(1,'\n');
    for i = 1:size(Num_conf,2)
        fprintf(1,'%s ',char(Conformation_names{Conf_idx(Sum_sort_idx(i))}));
        if Electrostatic_subtracted == 1
            fprintf(1,'%3g %.3f %3g %.3f %.3f %3g\n', [Num_conf(i) Elected(i,:)]);
        else
            fprintf(1,'%3g %.3f %3g %.3f %3g\n', [Num_conf(i) Elected(i,:)]);
        end
    end
    [Erg_sort,Erg_sort_idx]=sort(Energy_list(Conf_idx));
    Elected = sortrows(Elected,3);
    fprintf(1,'\n');
    disp('Sortiert nach Energie: ');
    fprintf(1,'\n');
    for i = 1:size(Num_conf,2)
        fprintf(1,'%s ',char(Conformation_names{Conf_idx(Erg_sort_idx(i))}));
        if Electrostatic_subtracted == 1
            fprintf(1,'%3g %.3f %3g %.3f %.3f %3g\n', [Num_conf(i) Elected(i,:)]);
        else
            fprintf(1,'%3g %.3f %3g %.3f %3g\n', [Num_conf(i) Elected(i,:)]);
        end
    end
    fprintf(1,'\n');
    disp('Distribution of Diversity')
    fprintf(1,'\n');
    fprintf(1,'%3g \n',[Div_loc' Div_freq']');
    fprintf(1,'\n');
    disp('Distribution of Similarity')
    fprintf(1,'\n');
    fprintf(1,'%3g \n',[Sim_loc' Sim_freq']');
elseif Modop == 1
    Summen = [Num_conf' Sum_conf' Conf_idx Sim_num_all(Conf_idx)'];
    disp('Sortiert nach Nummer: ');
    fprintf(1,'\n');
    for i = 1:size(Num_conf,2)
        fprintf(1,'%s ',char(Conformation_names{Conf_idx(i)}));
        fprintf(1,'%3g %.3f %3g %3g\n', Summen');
    end
    fprintf(1,'\n');
    [Sum_sort,Sum_sort_idx]=sort(Sum_conf);
    Sum_sort=flipud(Sum_sort');
    Sum_sort_idx=flipud(Sum_sort_idx');
    Elected = [Sum_sort Conf_idx(Sum_sort_idx) ...
        Sim_num_all(Conf_idx(Sum_sort_idx,1))'];
    disp('Sortiert nach durchschnittl. RMS: ');
    fprintf(1,'\n');
    for i = 1:size(Num_conf,2)
        fprintf(1,'%s ',char(Conformation_names{Conf_idx(Sum_sort_idx(i))}));
        fprintf(1,'%3g %.3f %3g %3g\n', [Num_conf(i) Elected(i,:)]);
    end
    fprintf(1,'\n');
    disp('Distribution of Diversity')
    fprintf(1,'\n');
    fprintf(1,'%3g \n',[Div_loc' Div_freq']');
    fprintf(1,'\n');
    disp('Distribution of Similarity')
    fprintf(1,'\n');
    fprintf(1,'%3g \n',[Sim_loc' Sim_freq']');
end

```

## D.4 Das SYBYL-Skript auto\_pls

```

# (c) Mad 2004 Ver 060104
# PLS Macro that automatical loops through conformations and does
# PLS Analyses
# Before starting it select all rows that should not be permuted
# (Hint: Create a set out of it) and have the required ".dat" -Files
# for permutation lying in the current default directory
# the following arguments (in order) are needed:
# Filename of the permutation file (without extension .dat)
# Minimum Q_SQUARED as a threshold for time consuming PLS
# Column Filtering Sigma for real PLS (not applicable for SAMPLS)
# Maximum Number of Components (for SAMPLS)
# Number of CV Groups (= Number of compounds for LOO)

uims define macro auto_pls SybylBasic
setvar old_timeout $cgq_timeout
setvar cgq_timeout 0

# this switches off the dialog box
setvar FIRST_SAMPLS done
setvar filename_perm %promptif("$1" "filename" "autopls" "Permutation file \"
    "Filename with permuting conformations and without extension .dat")
setvar MIN_Q_SQUARED %promptif("$2" "positive_real" "0.01" "Minimum Q_SQUARED")
setvar MINIMUM_SIGMA %promptif("$3" "positive_real" "1" "Column Filtering sigma")
echo Minimum Q_SQUARED $MIN_Q_SQUARED
echo Minimum MINIMUM_SIGMA $MINIMUM_SIGMA

setvar permutation_file %cat($filename_perm ".dat")

#open the outputfile for writing
setvar outputres_fh %open(%cat($filename_perm ".out") "w")

if %dir($permutation_file)
    setvar pf_fh %open($permutation_file "r")
else
    echo File $permutation_file not found !
    break
endif
setvar pf_line %read($pf_fh)
setvar max_cl 1
while 1
    setvar tttt %read($pf_fh)
    setvar tempdepcol %arg(1 $tttt)
    if %eof($pf_fh)
        setvar max_cl %math($max_cl -1)
        break
    endif
    if %syb_int($tempdepcol)
        setvar col_line[$max_cl] $tttt
        setvar max_cl %math($max_cl +1)
    endif
endwhile
setvar schrott %close($pf_fh)
setvar max_groups %count($pf_line)
for i in %range(1 $max_groups)
    setvar group_names[$i] %arg($i $pf_line)
    if %dir(%cat($group_names[$i] ".dat"))
        setvar pl_fh %open(%cat($group_names[$i] ".dat") "r")
    else
        echo File %cat($group_names[$i] ".dat") not found !
        break
    endif
    setvar pl_line %read($pl_fh)
    setvar name_line %read($pl_fh)
    setvar schrott %close($pl_fh)
    setvar permut_groups[$i] $pl_line
    setvar element_names[$i] $name_line

```



```

    setvar group_sizes[$i] %count($pl_line)
    setvar element_counter[$i] 1
endfor
70

setvar permutation_count 1
setvar ready 0

while %EQ($ready 0)
    for permuter in %range(1 $max_groups)
        setvar helpset[$permuter] %arg($element_counter[$permuter] $permut_groups[$permuter])
    endfor
    setvar permutations[$permutation_count] $helpset
    setvar permutations_index[$permutation_count] $element_counter
    setvar permutation_count %math($permutation_count+1)
    setvar uebertrag 1
    setvar permuter 1
    while %EQ($uebertrag 1)
        setvar element_counter[$permuter] %math($element_counter[$permuter] + 1)
        if %GT($element_counter[$permuter] $group_sizes[$permuter])
            setvar element_counter[$permuter] 1
            setvar uebertrag 1
            setvar permuter %math($permuter + 1)
            if %GT($permuter $max_groups)
                setvar ready 1
                setvar uebertrag 0
            endif
        else
            setvar uebertrag 0
        endif
    endwhile
endwhile
setvar permutation_count %math($permutation_count - 1)
setvar temp_name %cat($filename_perm ".tmp")
if %dir($temp_name)
    setvar schrott %file_delete(%files(file $temp_name))
endif
100
TABLE SUBSET CREATE "zyxxxtemp" ROW {selected()}
capture $temp_name TABLE SUBSET LIST MEMBERS zyxxxtemp
setvar num_rows_sel %system("grep Number $temp_name | cut -d: -f2")
setvar num_rows_sel %math($num_rows_sel + $max_groups)

setvar CROSSVALIDATION %promptif("$5" "INT" $num_rows_sel "Number of CV Groups")

echo CROSSVALIDATION $CROSSVALIDATION
110

setvar COMPONENTS %promptif("$4" "INT" %math($CROSSVALIDATION-2) \
    "Maximum Number of Components")
if %gt($COMPONENTS %math($CROSSVALIDATION-2))
    setvar COMPONENTS %math($CROSSVALIDATION-2)
endif
echo COMPONENTS $COMPONENTS

for a in %range(1 $max_cl)
    setvar tempdir $filename_perm
    setvar k 1
    for i in $col_line[$a]
        setvar temp_name %cat($filename_perm ".tmp")
        if %dir($temp_name)
            setvar schrott %file_delete(%files(file $temp_name))
        endif
        capture $temp_name TABLE LIST DESCRIPTION COLUMN $i
        setvar temp1 %system("grep COLUMN $temp_name")
        setvar schrott %file_delete(%files(file $temp_name))
        setvar col_name[$i] %arg(3 $temp1)
        setvar dir_name %cat($tempdir "_" $col_name[$i])
        setvar tempdir $dir_name
        if %gt($k 1)
            setvar columns_to_use %cat($columns_to_use ", " $i)
        else

```

```

        setvar dependent_col $i
        setvar columns_to_use $i
    endif
    setvar k %math($k+1)
endfor
140
if %FILE_ISDIR($dir_name)
    setvar system_str %cat("rm -r " $dir_name)
    setvar schrott %system("$system_str")
endif
setvar schrott %file_make_dir($dir_name)
for i in %range(1 $permutation_count)
    tailor set qsar minimum_sigma $MINIMUM_SIGMA ||
    tailor set pls CROSSVALIDATION $CROSSVALIDATION ||
    tailor set pls SCALING_METHOD COMFA_STD ||
    TABLE SELECT ROW {zyxxxtemp}
    150
    for j in $permutations[$i]
        setvar rowselection %cat($j "+{selected()}")
        TABLE SELECT ROW $rowselection
    endfor
    setvar k 1
    setvar tempname $filename_perm
    for j in $permutations_index[$i]
        setvar xname $element_names[$k]
        setvar filename_lis %cat($tempname "_" $group_names[$k] %arg($j $xname))
        160
        setvar k %math($k + 1)
        setvar tempname $filename_lis
    endfor
    # first we do SAMPLS
    echo "Doing SAMPLS"
    QSAR ANALYSIS SAMPLS {SELECTED()} $columns_to_use $dependent_col \
        $COMPONENTS COMFA_STD SPL
    parse_sampls SAMPLS.output $COMPONENTS SDEP
    # writing the SAMPLS-Results to file
    setvar outputstring %cat($filename_lis " " $QSAR_CROSSVALIDATED_R_SQUARED \
        " " $Standard_error_op " " $QSAR_OPTIMAL_COMPONENTS)
    170
    %write($outputres_fh $outputstring)
    # then we do ordinary PLS with optimized components
    # only if SAMPLS Q_SQUARED is good enough
    if %gt($QSAR_CROSSVALIDATED_R_SQUARED $MIN_Q_SQUARED)
        tailor set pls COMPONENTS $QSAR_OPTIMAL_COMPONENTS ||
        echo $filename_lis
        setvar filename_pls %cat($filename_perm "_" %irand() %irand() "_" $i)
        if %dir(%cat($filename_pls ".pls"))
            setvar schrott %file_delete(%files(file %cat($filename_pls ".pls")))
            180
        endif
        QSAR ANALYSIS DO INTERACTIVE {SELECTED()} $columns_to_use PLS \
            $dependent_col | $filename_pls
        if %dir(%cat($filename_lis ".lis"))
            setvar schrott %file_delete(%files(file %cat($filename_lis ".lis")))
        endif
        QSAR ANALYSIS LIST ASCII_FILE $filename_lis ALL
        if %dir(%cat($filename_lis ".lis"))
            QSAR ANALYSIS DELETE PLS $filename_pls
            setvar system_str %cat("mv " %cat($filename_lis ".lis") " " $dir_name)
            190
            setvar schrott %system("$system_str")
        endif
    endif
endfor
endfor
TABLE SELECT ROW {zyxxxtemp}
TABLE SUBSET DELETE zyxxxtemp
echo %close($outputres_fh)
setvar CGQ_timeout $old_timeout
200

```

## D.5 Das Programm plsreport

```

%%
%% by Mad 01/2004 version 140305
%%
%% function analyzes a number of PLS-Reports (SYBYL-lis-Files)
%% and reports mean Residual-Values
%%
%% mögliche Verbesserungen:
%% - Wichtung des Fehlers mit der Häufigkeit
%% - Vorschlag für Selektion für Validierungs PLS (externe Vorhersage o.ä.)
10

function [Mean_res] = plsreport(Filename_all,Q_squared_threshold);

Mean_res = 1;
if exist(Filename_all,'file') == 2
    Sel_file = fopen(Filename_all,'r');
else
    disp('File not found !');
    Mean_res = strvcats([Filename_all ' not found ! ']);
    return
end
20
if nargin < 2
    Q_squared_threshold = 0;
end
Line_count = 1;
Index = {};
Conf_names = {};
Res_sum{1} = [];
Q_squared = [];
while feof(Sel_file) == 0
    File_line{Line_count} = fgetl(Sel_file);
    30
    if exist(File_line{Line_count},'file') == 2
        fprintf(1,'\nReading SYBYL report file %s',File_line{Line_count});
        lis_file = fopen(File_line{Line_count},'r');
        Test_line = [];
        for occ = 1:2
            while isempty(findstr('R squared',Test_line))== 1
                Test_line = fgetl(lis_file);
            end
            if occ==1 ; Test_line = []; end
        end
        40
        Line_decode = sscanf(Test_line,'%s %s %f');
        Q_squared_val = Line_decode(size(Line_decode,1));
        Q_squared = [Q_squared Q_squared_val];
        if Q_squared_val > Q_squared_threshold
            frewind(lis_file);
            Test_line = [];
            while isempty(findstr('Residual Values',Test_line))== 1
                Test_line = fgetl(lis_file);
            end
            50
            for i = 1:2
                schrott = fgetl(lis_file);
            end
            Data_count = 1;
            Data_line{Data_count} = fgetl(lis_file);
            while isempty(findstr('Standard Error of Predictions (Crossvalidated)',...
                Data_line{Data_count})) == 1
                Data_count = Data_count + 1;
                Data_line{Data_count} = fgetl(lis_file);
            end
            60
            i = 1;
            while i < Data_count
                if isempty(findstr('# ',Data_line{i})) == 1 & isempty(findstr('----',...
                    Data_line{i})) == 1 & isempty(findstr(char(12),Data_line{i})) == 1
                    Line_decode = sscanf(Data_line{i},'%i %s %f');
                    Index = Line_decode(1);
                end
                i = i + 1;
            end
        end
    end
end

```

---

```

        if i == 1 & Line_count == 1
            Indices = Index;
            Conf_names{Index} = char(Line_decode(2:size(Line_decode,1)-1));
            Res_sum{Index} = abs(Line_decode(size(Line_decode,1)));
            Res_count{Index} = 1;
        elseif ismember(Index, Indices) == 0
            Indices = [Indices Index];
            Conf_names{Index} = char(Line_decode(2:size(Line_decode,1)-1));
            Res_sum{Index} = abs(Line_decode(size(Line_decode,1)));
            Res_count{Index} = 1;
        else
            Res_sum{Index} = Res_sum{Index} + abs(Line_decode(size...
                (Line_decode,1)));
            Res_count{Index} = Res_count{Index} + 1;
        end
        i = i+1;
    else
        i = i+1;
    end
end
Line_count = Line_count + 1;
end
else
    disp('File not found !');
    Mean_res = strvcats([File_line{Line_count} ' not found ! ']);
end
fclose(lis_file);
end
fclose(Sel_file);
Num_conf = size(Indices,2);
Mean_res = zeros(Num_conf,1);
Indices = sort(Indices);
fprintf(1,'\n\n');
disp('Average Residuals for each occurring Conformation: ');
fprintf(1,'\n');
disp('Line Name          Count MeanRes');
for i = 1:Num_conf
    Mean_res(i) = Res_sum{Indices(i)}/Res_count{Indices(i)};
    String = strvcats(['%g \t ' Conf_names{Indices(i)} ' \t %g \t %f ']);
    fprintf(1,String,[Indices(i) Res_count{Indices(i)} Mean_res(i) ]);

end
figure(1);
bar(Mean_res);
set(gca,'Xtick',1:Num_conf)
set(gca,'Xticklabel',Indices)
%set(gca,'Xticklabel',{Conf_names{Indices}})
set(gca,'XticklabelMode','manual')
h = findobj(gca,'Type','patch');
set(h,'FaceColor','b','EdgeColor','k')
title('Average Residual Values','FontSize',16)
xlabel('Line Nr. in SYBYL MSS','FontSize',14)
ylabel('Mean Residual Value','FontSize',14)
axis tight;
figure(2);
hist(Q_squared);
title('Distribution of q^2','FontSize',16)
xlabel('q^2','FontSize',14)
ylabel('count','FontSize',14)
axis tight;

```

---

## D.6 Das SYBYL-Skript random\_groups\_pls

```

# (c) Mad 2004 Ver 131004
# PLS Macro that automatically performs random groups PLS for model validation
#
# Before starting it select all rows that should be considered
# (Hint: Create a set out of it)

# the following arguments (in order) are needed:
# Filename of the output file (without extension .out)
# Maximum Number of Components
# Column Filtering Sigma for PLS
# Count for loop of repetition
# Number of CV Groups
# Columns to use
# Dependent Column
# example: random_groups_pls d3_rnd5 4 2.0 20 3 "1,4" 1
# will perform it for: 4 Components
#                      2.0 Column filtering
#                      20 Rounds
#                      3 CV Groups
#                      Cols 1 and 4 to use
#                      Column 1 as dependent Column
# output file would be: d3_rnd5
10

uims define macro random_groups_pls SybylBasic
setvar old_timeout $cgq_timeout
setvar cgq_timeout 0

setvar filename_out %promptif("$1" "filename" "output_res" "Output file \
    " "Filename for output of q-squared and without extension ")
setvar COMPONENTS %promptif("$2" "INT" "5" "Minimum Q_SQUARED")
setvar MINIMUM_SIGMA %promptif("$3" "positive_real" "1" "Column Filtering sigma")
setvar REPEAT_COUNT %promptif("$4" "INT" "5" "How many runs")
30

#open the outputfile for writing
setvar outputres_fh %open(%cat($filename_out ".out") "w")

setvar CROSSVALIDATION %promptif("$5" "INT" "3" "Number of CV Groups")
setvar columns_to_use %promptif("$6" "col_exp" "1 4" "Columns to use")
setvar dependent_col %promptif("$7" "col_sel" "1 4" "dependent Column")
40

setvar dir_name $filename_out
if %FILE_ISDIR($dir_name)
    setvar system_str %cat("rm -r " $dir_name)
    setvar schrott %system("$system_str")
endif
setvar schrott %file_make_dir($dir_name)
tailor set qsar minimum_sigma $MINIMUM_SIGMA ||
tailor set pls CROSSVALIDATION $CROSSVALIDATION ||
tailor set pls SCALING_METHOD COMFA_STD ||
tailor set pls COMPONENTS $COMPONENTS ||
50

for counter in %range(1 $REPEAT_COUNT)
setvar filename_lis %cat("rnd_" $CROSSVALIDATION "groups_" $counter)
echo $filename_lis
setvar filename_pls $filename_lis
if %dir(%cat($filename_pls ".pls"))
    setvar schrott %file_delete(%files(file %cat($filename_pls ".pls")))
endif
QSAR ANALYSIS DO INTERACTIVE {SELECTED()} $columns_to_use PLS \
    $dependent_col | $filename_pls
if %dir(%cat($filename_lis ".lis"))
    setvar schrott %file_delete(%files(file %cat($filename_lis ".lis")))
endif
60
QSAR ANALYSIS LIST ASCII_FILE $filename_lis ALL
if %dir(%cat($filename_lis ".lis"))
    QSAR ANALYSIS DELETE PLS $filename_pls

```

```
        setvar system_str %cat("mv " %cat($filename_lis ".lis") " " $dir_name)
        setvar schrott %system("$system_str")
    endif
    setvar results[$counter] $QSAR_CROSSVALIDATED_R_SQUARED
    setvar outputstring %cat($filename_lis " " $QSAR_CROSSVALIDATED_R_SQUARED \
        " " $QSAR_OPTIMAL_COMPONENTS)
    %write($outputres_fh $outputstring)
endfor
echo
echo Basic Statistics:
echo
echo Mean Q_SQUARED %stats($results mean)
echo Standard " " Dev. %stats($results sd)
echo %close($outputres_fh)
setvar CGQ_timeout $old_timeout
.
```

---

## D.7 Das Programm cosmo\_anA

```

%%
%% by Mad 09/2004 version 13092005
%%
%% function analyzes a cosmo-file
%% and computes and graphs the electrostatic distribution
%% it uses the parzen_window function to calculate the density estimation

function [Result] = cosmo_pw_anA(Filename_all,X_vec,Window_width,Legend)

%% Column in the cosmo-file which holds the sigma-Values
Sigma_col = 8;
%% Column in the cosmo-file which holds the area-Values
Area_col = 7;
%% Column in the cosmo-file which hold the coordinates
Coordinate_cols = 3:5;
%% Column in the cosmo-file which holds the charge-values
Charge_col = 6;

if nargin < 3
    if nargin < 2
        if nargin < 1
            disp('You must specify at least Filename and X Vector [min:intercept:max] ');
        else
            disp('You must specify an X Vector [min:intercept:max] ');
        end
        return
    else
        % no Window_width given
        % using only default KernelBandwidth
        Default_width=true;
        disp('No window width for density estimation given');
        disp('Using window width calculated from stand. dev. for all files');
    end
else Default_width=false;
end
if exist(Filename_all,'file') == 2
    if isempty(strfind(Filename_all,'.lst'))
        Listfilemod = 0;
    else
        Listfilemod = 1;
        List_file = fopen(Filename_all,'r');
        File_count = 0;
        while feof(List_file) == 0
            File_count = File_count + 1;
            Cosmo_files{File_count} = fgetl(List_file);
        end
        fclose(List_file);
    end
else
    disp('File not found !');
    Result = strvcats([Filename_all ' not found ! ']);
    return
end
if Listfilemod == 0
    File_count=1;
    Cosmo_files{File_count}= Filename_all;
else
    if ~Default_width && length(Window_width) < File_count
        disp('Not enough numbers in Window_width vector given for individual scaling');
        fprintf(1,'\nUsing given first window width %f for all files\n',Window_width(1));
        Window_width=ones(1,File_count)*Window_width(1);
    elseif Default_width
        Window_width=zeros(1, File_count);
    end
end
end
Avg_sigma=cell(1,File_count);

```

```

Density_refined=cell(1,File_count);
Result=cell(1,File_count);
F_name=cell(1,File_count);
for i = 1:File_count
    if exist(Cosmo_files{i},'file') == 2
        Cosmo_file = fopen(Cosmo_files{i},'r');
        fprintf(1,'\nReading COSMO output file %s \n',Cosmo_files{i});
        Test_line = [];
        [F_pathstr,F_name{i}] = fileparts(Cosmo_files{i});
        while isempty(findstr('$segment_information',Test_line))== 1
            Test_line = fgetl(Cosmo_file);
        end
        Test_line = fgetl(Cosmo_file);
        while isempty(findstr('#',Test_line))== 0
            Test_line = fgetl(Cosmo_file);
        end
        fseek(Cosmo_file, -size(Test_line,2), 'cof');
        tempdat = textscan(Cosmo_file, '%f');
        Cosmo_mat=reshape(tempdat{:}',9,[]);
        fclose(Cosmo_file);
        Min_sig=min(Cosmo_mat(:,Sigma_col));
        Max_sig=max(Cosmo_mat(:,Sigma_col));
        Mean_segment_r = sqrt(mean(Cosmo_mat(:,Area_col))/pi);
        fprintf(1,'Minimum sigma potential: %f \n',Min_sig);
        fprintf(1,'Maximum sigma potential: %f \n',Max_sig);
        fprintf(1,'Mean Segment radius : %f \n',Mean_segment_r);
        Avg_sigma{i} = sigma_avg(Cosmo_mat(:,[Coordinate_cols Charge_col Area_col Sigma_col]));
        if Default_width
            [Dns,W_width] = parzen_window([Cosmo_mat(:,Area_col) Avg_sigma{i}],X_vec);
            Window_width(i) = W_width;
        else
            [Dns] = parzen_window([Cosmo_mat(:,Area_col) Avg_sigma{i}],X_vec,Window_width(i));
        end
        fprintf(1,'Using Bandwidth : %f \n', Window_width(i));
        Density_refined{i} = Dns;
        Result{i}=[X_vec' Density_refined{i}];
    else
        fprintf(1,'\n Cosmo-File %s nicht gefunden \n',Cosmo_files{i})
        return
    end
end
hold off
for i = 1:File_count
    hold all
    plot(X_vec,Density_refined{i},'LineWidth',2);
    if i == File_count
        title('COSMO Sigma Profile ', 'Interpreter','none','FontSize',16)
        xlabel('\sigma[e/A^2]','FontSize',14)
        ylabel('Density estimation \rho^x(\sigma)','FontSize',14)
        if nargin < 4
            legend(char(F_name))
        else
            legend(char(Legend))
        end
    end
end
end
hold off

%% this subfunction does the averaging of the sigma values for a given
%% radius R_av as described in Klamt et al. J.Phys.Chem.A 1998, 102,
%% 5074-5085

function Sigma_avg = sigma_avg(Cosmo_out)
%% Cosmo_out is a matrix consisting of:
%% [X Y Z coordinates of segments (in a.u.)]
%% [charge] [area (in A^2)] [charge/area]

R_av = 0.5; %this is the optimized Value according to A.Klamt

```



---

```

Num_segments = size(Cosmo_out,1);
%% first we should convert the coordinates to Angstroem
%% 1 Bohr is 0.529177249 Angstroem

Segment_position = Cosmo_out(:,1:3)*0.529177249;
R2_mu = Cosmo_out(:,5)./pi;
R2_av = R_av^2;
Sigma_avg = zeros(Num_segments,1);
for nu = 1:Num_segments
    Avg_term = zeros(Num_segments,1);
    for mu = 1:Num_segments
        D2_munu = (norm(Segment_position(mu,:) - Segment_position(nu,:)))^2;
        Avg_term(mu) = (R2_mu(mu) * R2_av)/(R2_mu(mu) + R2_av) * ...
            exp(-D2_munu/(R2_mu(mu) + R2_av));
    end
    Sigma_avg(nu) = sum(Avg_term.*Cosmo_out(:,6)) / sum(Avg_term);
end
function [Density_est, Window_width] = parzen_window(Raw_data, Vector, Window_width)
Data_size = size(Raw_data,1);
Vector_size = size(Vector,2);
if nargin < 3
    %% no Window_width given
    %% Calculate the Window_width from the standard deviation (just like
    %% ksdensity does)
    Window_width = std(Raw_data(:,2))*((4/(3*Data_size))^(1/5));
end
%% initialize the Density_est vector with <Vector_size> zeros
Density_est = zeros(Vector_size,1);
%% initialize the Kernel_densities vector with <Data_size> zeros
for i = 1:Data_size
    for j = 1:Vector_size
        Density_est(j) = Density_est(j) + normpdf(Vector(j),Raw_data(i,2),Window_width);
    end
end
Density_est = Density_est / Data_size;

```

---



# Literatur

- [1] S. Borman, New QSAR techniques eyed for environmental assessments, *Chem. Eng. News* **1990**, 68, 20–23.
- [2] C. E. Overton, Studien zur Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie, Gustav Fischer Verlag, Jena, **1901**.
- [3] R. L. Lipnick, Charles Ernest Overton: narcosis studies and a contribution to general pharmacology, *Trends Pharmacol. Sci.* **1986**, 7, 161–164.
- [4] R. L. Lipnick, Hans Horst Meyer and the lipoid theory of narcosis, *Trends Pharmacol. Sci.* **1989**, 10, 265–269.
- [5] L. P. Hammett, Some relations between reaction rates and equilibrium constants, *Chem. Rev. (Washington, DC, U. S.)* **1935**, 17, 125–136.
- [6] C. Hansch und T. Fujita,  $\rho - \sigma - \pi$  analysis. A method for the correlation of biological activity and chemical structure, *J. Am. Chem. Soc.* **1964**, 86, 1616–1626.
- [7] S. M. Free und J. W. Wilson, A mathematical contribution to structure-activity studies, *J. Med. Chem.* **1964**, 7, 395–399.
- [8] P. Zbinden, M. Dobler, G. Folkers und A. Vedani, PrGen: Pseudoreceptor modeling using receptor-mediated ligand alignment and pharmacophore equilibration, *Quant. Struct-Act. Relat.* **1998**, 17, 122–130.
- [9] M. Weigt, Entwicklung eines neuen Verfahrens zur Erzeugung von Pseudorezeptormodellen, Diplomarbeit, Martin-Luther-Universität Halle-Wittenberg **2000**.
- [10] A. J. Hopfinger, S. Wang, J. S. Tokarski, B. Jin, M. Albuquerque, P. J. Madhav und C. Duraiswami, Construction of 3D-QSAR models using the 4D-QSAR analysis formalism, *J. Am. Chem. Soc.* **1997**, 119, 10509–10524.
- [11] A. Vedani und M. Dobler, 5D-QSAR: the key for simulating induced fit?, *J. Med. Chem.* **2002**, 45, 2139–2149.

- [12] F. Glover, Tabu Search, Part I, *ORSA J. Comput.* **1989**, 1, 190–206.
- [13] F. Glover, Tabu Search, Part II, *ORSA J. Comput.* **1990**, 2, 4–32.
- [14] K. Baumann, H. Albert und M. von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations, *J. Chemom.* **2002**, 16, 339–350.
- [15] K. Baumann, H. Albert und M. von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations, *J. Chemom.* **2002**, 16, 351–360.
- [16] S. Wold, H. Wold und W. J. Dunn, The collinearity problem in linear regression: The partial least squares approach to generalized inverses, *SIAM J. Sci. Stat. Comput.* **1984**, 5, 753–743.
- [17] C. Bystroff, S. J. Oatley und J. Kraut, Crystal structures of Escherichia coli dihydrofolate reductase: the NADP<sup>+</sup> holoenzyme and the folate-NADP<sup>+</sup> ternary complex. Substrate binding and a model for the transition state, *Biochemistry* **1990**, 29, 3263–3277.
- [18] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. L. Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto und M. Miyano, Crystal structure of rhodopsin: A G protein-coupled receptor, *Science* **2000**, 289, 739–745.
- [19] M. A. Schramm, Zur Synthese und biologischen Aktivität von Azecino- und Azonino[5,4-b]indolen, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn **1998**.
- [20] P. M. Schweikert, Dibenzo[d,g]- und Benzo[d]thieno[3,2-g]azecine als potentielle Arzneistoffe zur Behandlung der Schizophrenie, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn **1999**.
- [21] M. Decker, Synthese und pharmakologische Evaluierung strukturell neuartiger Dopamin-Rezeptor-Liganden vom Azecin- und Azepin-Typ, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn **2001**.
- [22] S. Lankow, Neuartige Dopamin-Rezeptor-Liganden vom Benz[d]indolo[2,3-g]azecin-Typ Synthese, Struktur und biologische Aktivität, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn **2001**.

- [23] B. Hoefgen, Etablierung eines funktionellen Calcium-Assays und seine Anwendung zum Screening potentieller Liganden an humanen, klonierten Dopamin-Rezeptoren, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn **2002**.
- [24] G. J. Macdonald, C. L. Branch, M. S. Hadley, C. N. Johnson, D. J. Nash, A. B. Smith, G. Stemp, K. M. Thewlis, A. K. K. Vong, N. E. Austin, P. Jeffrey, K. Y. Winborn, I. Boyfield, J. J. Hagan, D. N. Middlemiss, C. Reavill, G. J. Riley, J. M. Watson, M. Wood, S. G. Parker und C. R. A. Jr, Design and synthesis of trans-3-(2-(4-((3-(3-(5-methyl-1,2,4-oxadiazolyl))-phenyl)carboxamido)cyclohexyl)ethyl)-7-methylsulfonyl-2,3,4,5-tetrahydro-1H-3-benzazepine (SB-414796): a potent and selective dopamine D3 receptor antagonist, *J. Med. Chem.* **2003**, 46, 4952–4964.
- [25] B. L. Roth, E. Lopez, S. Patel und W. K. Kroeze, The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches?, *Neuroscientist* **2000**, 6, 252–262.
- [26] Y. Cheng und W. Prusoff, Relationship between the inhibition constant (K<sub>i</sub>) and the concentration of inhibitor which causes 50 per cent inhibition (I<sub>50</sub>) of an enzymatic reaction, *Biochem. Pharmacol.* **1973**, 22, 3099–30108.
- [27] M. Berridge, P. Lipp und M. Bootman, The versatility and universality of calcium signalling, *Nat. Rev. Mol. Cell Biol.* **2000**, 1, 11–21.
- [28] R. Sunahara, H. Guan, B. O'Dowd, P. Seeman, L. Laurier, G. Ng, S. George, J. Torchia, H. V. Tol und H. Niznik, Cloning of the gene for a human dopamine D5 receptor with higher affinity for dopamine than D1, *Nature* **1991**, 350, 614–619.
- [29] K. D. Burris, T. F. Molski, C. Xu, E. Ryan, K. Tottori, T. Kikuchi, F. D. Yocca und P. B. Molinoff, Aripiprazole, a novel antipsychotic, is a high-affinity partial agonist at human dopamine D2 receptors, *J. Pharmacol. Exp. Ther.* **2002**, 302, 381–389.
- [30] V. Butterweck, A. Nahrstedt, J. Evans, S. Hufeisen, L. Rauser, J. Savage, B. Popadak, P. Ernsberger und B. L. Roth, In vitro receptor screening of pure constituents of St. John's wort reveals novel interactions with a number of GPCRs, *Psychopharmacology (Berl)* **2002**, 162, 193–202.
- [31] B. Capuano, I. Crosby und E. Lloyd, Schizophrenia: genesis, receptorology and current therapeutics, *Curr. Med. Chem.* **2002**, 9, 521–548.

- [32] D. Cussac, A. Newman-Tancredi, L. Sezgin und M. Millan, [3H]S33084: a novel, selective and potent radioligand at cloned, human dopamine D3 receptors, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2000**, 361, 569–572.
- [33] A. Hameg, F. Bayle, P. Nuss, P. Dupuis, R. P. Garay und M. Dib, Affinity of cyamemazine, an anxiolytic antipsychotic drug, for human recombinant dopamine vs. serotonin receptor subtypes, *Biochem. Pharmacol.* **2003**, 65, 435–440.
- [34] M. Jarvis, H. Yu, K. Kohlhaas, K. Alexander, C. Lee, M. Jiang, S. Bhagwat, M. Williams und E. Kowaluk, ABT-702 (4-amino-5-(3-bromophenyl)-7-(6-morpholinopyridin-3-yl)pyrido[2, 3-d]pyrimidine), a novel orally effective adenosine kinase inhibitor with analgesic and anti-inflammatory properties: I. In vitro characterization and acute antinociceptive effects in the mouse, *J. Pharmacol. Exp. Ther.* **2000**, 295, 1156–1164.
- [35] L. Johansson, D. Sohn, S. Thorberg, D. Jackson, D. Kelder, L. Larsson, L. Rényi, S. Ross, C. Wallsten, H. Eriksson, P. Hu, E. Jerning, N. Mohell und A. Westlind-Danielsson, The pharmacological characterization of a novel selective 5-hydroxytryptamine<sub>1A</sub> receptor antagonist, NAD-299, *J. Pharmacol. Exp. Ther.* **1997**, 283, 216–225.
- [36] E. Kowaluk, J. Mikusa, C. Wismer, C. Zhu, E. Schweitzer, J. Lynch, C. Lee, M. Jiang, S. Bhagwat, A. Gomtsyan, J. McKie, B. Cox, J. Polakowski, G. Reinhart, M. Williams und M. Jarvis, ABT-702 (4-amino-5-(3-bromophenyl)-7-(6-morpholino-pyridin- 3-yl)pyrido[2,3-d]pyrimidine), a novel orally effective adenosine kinase inhibitor with analgesic and anti-inflammatory properties. II. In vivo characterization in the rat, *J. Pharmacol. Exp. Ther.* **2000**, 295, 1165–1174.
- [37] C. Lawler, C. Prioleau, M. Lewis, C. Mak, D. Jiang, J. Schetz, A. Gonzalez, D. Sibley und R. Mailman, Interactions of the novel antipsychotic aripiprazole (OPC-14597) with dopamine and serotonin receptor subtypes, *Neuropsychopharmacology* **1999**, 20, 612–627.
- [38] Y. Liao, B. Venhuis, N. Rodenhuis, W. Timmerman, H. Wikström, E. Meier, G. Bartoszyk, H. Böttcher, C. Seyfried und S. Sundell, New (sulfonyloxy)piperazinyldibenzazepines as potential atypical antipsychotics: chemistry and pharmacological evaluation, *J. Med. Chem.* **1999**, 42, 2235–2244.
- [39] D. Marona-Lewicka und D. E. Nichols, Aripiprazole (OPC-14597) fully substitutes for the 5-HT<sub>1A</sub> receptor agonist LY293284 in the drug discrimination assay in rats, *Psychopharmacology (Berl)* **2004**, 172, 415–421.

- [40] M. Millan, A. Dekeyne, J. Rivet, T. Dubuffet, G. Lavielle und M. Brocco, S33084, a novel, potent, selective, and competitive antagonist at dopamine D(3)-receptors: II. Functional and behavioral profile compared with GR218,231 and L741,626, *J. Pharmacol. Exp. Ther.* **2000**, 293, 1063–1073.
- [41] M. Millan, A. Gobert, A. Newman-Tancredi, F. Lejeune, D. Cussac, J. Rivet, V. Audinot, T. Dubuffet und G. Lavielle, S33084, a novel, potent, selective, and competitive antagonist at dopamine D(3)-receptors: I. Receptorial, electrophysiological and neurochemical profile compared with GR218,231 and L741,626, *J. Pharmacol. Exp. Ther.* **2000**, 293, 1048–1062.
- [42] M. J. Millan, L. Maiofiss, D. Cussac, V. Audinot, J.-A. Boutin und A. Newman-Tancredi, Differential actions of antiparkinson agents at multiple classes of monoaminergic receptor. I. A multivariate analysis of the binding profiles of 14 drugs at 21 native and cloned human receptor subtypes, *J. Pharmacol. Exp. Ther.* **2002**, 303, 791–804.
- [43] M. J. Millan, M. Brocco, M. Papp, F. Serres, C. D. L. Rochelle, T. Sharp, J.-L. Peglion und A. Dekeyne, S32504, a novel naphthoxazine agonist at dopamine D3/D2 receptors: III. Actions in models of potential antidepressive and anxiolytic activity in comparison with ropinirole, *J. Pharmacol. Exp. Ther.* **2004**, 309, 936–950.
- [44] M. J. Millan, B. D. Cara, M. Hill, M. Jackson, J. N. Joyce, J. Brotchie, S. McGuire, A. Crossman, L. Smith, P. Jenner, A. Gobert, J.-L. Peglion und M. Brocco, S32504, a novel naphthoxazine agonist at dopamine D3/D2 receptors: II. Actions in rodent, primate, and cellular models of antiparkinsonian activity in comparison to ropinirole, *J. Pharmacol. Exp. Ther.* **2004**, 309, 921–935.
- [45] M. J. Millan, D. Cussac, A. Gobert, F. Lejeune, J.-M. Rivet, C. M. L. Cour, A. Newman-Tancredi und J.-L. Peglion, S32504, a novel naphthoxazine agonist at dopamine D3/D2 receptors: I. Cellular, electrophysiological, and neurochemical profile in comparison with ropinirole, *J. Pharmacol. Exp. Ther.* **2004**, 309, 903–920.
- [46] A. Newman-Tancredi, D. Cussac, V. Audinot, J.-P. Nicolas, F. D. Ceuninck, J.-A. Boutin und M. J. Millan, Differential actions of antiparkinson agents at multiple classes of monoaminergic receptor. II. Agonist and antagonist properties at subtypes of dopamine D(2)-like receptor and alpha(1)/alpha(2)-adrenoceptor, *J. Pharmacol. Exp. Ther.* **2002**, 303, 805–814.
- [47] A. Newman-Tancredi, D. Cussac, Y. Quentric, M. Touzard, L. Verri  le, N. Carpentier und M. J. Millan, Differential actions of antiparkinson agents at multi-

- ple classes of monoaminergic receptor. III. Agonist and antagonist properties at serotonin, 5-HT(1) and 5-HT(2), receptor subtypes, *J. Pharmacol. Exp. Ther.* **2002**, 303, 815–822.
- [48] D. E. Nichols, S. Frescas, D. Marona-Lewicka und D. M. Kurrasch-Orbaugh, Lysergamides of isomeric 2,4-dimethylazetidines map the binding orientation of the diethylamide moiety in the potent hallucinogenic agent N,N-diethyllysergamide (LSD), *J. Med. Chem.* **2002**, 45, 4344–4349.
- [49] L. Phebus, K. Johnson, J. Zgombick, P. Gilbert, K. V. Belle, V. Mancuso, D. Nelson, D. Calligaro, A. Kiefer, T. Branchek und M. Flaugh, Characterization of LY344864 as a pharmacological tool to study 5-HT<sub>1F</sub> receptors: binding affinities, brain penetration and activity in the neurogenic dural inflammation model of migraine, *Life Sci.* **1997**, 61, 2117–2126.
- [50] H. Schoemaker, Y. Claustre, D. Fage, L. Rouquier, K. Chergui, O. Curet, A. Obilin, F. Gonon, C. Carter, J. Benavides und B. Scatton, Neurochemical characteristics of amisulpride, an atypical dopamine D<sub>2</sub>/D<sub>3</sub> receptor antagonist with both presynaptic and limbic selectivity, *J. Pharmacol. Exp. Ther.* **1997**, 280, 83–97.
- [51] D. A. Shapiro, S. Renock, E. Arrington, L. A. Chiodo, L.-X. Liu, D. R. Sibley, B. L. Roth und R. Mailman, Aripiprazole, a novel atypical antipsychotic drug with a unique and robust pharmacology, *Neuropsychopharmacology* **2003**, 28, 1400–1411.
- [52] A. Tang, S. Franklin, C. Himes, M. Smith und R. Tenbrink, PNU-96415E, a potential antipsychotic agent with clozapine-like pharmacological properties, *J. Pharmacol. Exp. Ther.* **1997**, 281, 440–447.
- [53] B. J. Venhuis, D. Dijkstra, D. Wustrow, L. T. Meltzer, L. D. Wise, S. J. Johnson und H. V. Wikström, Orally active oxime derivatives of the dopaminergic prodrug 6-(N,N-di-n-propylamino)-3,4,5,6,7,8-hexahydro-2H-naphthalen-1-one. Synthesis and pharmacological activity, *J. Med. Chem.* **2003**, 46, 4136–4140.
- [54] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins und J. D. Thompson, Multiple sequence alignment with the Clustal series of programs., *Nucleic Acids Res.* **2003**, 31, 3497–3500.
- [55] P. G. Strange, Dissociation constants of neuroleptic drugs at dopamine receptors, *Neuropsychopharmacology* **1997**, 16, 116–122.



- [56] P. G. Strange, Antipsychotic drugs: importance of dopamine receptors for mechanisms of therapeutic actions and side effects, *Pharmacol. Rev.* **2001**, 53, 119–133.
- [57] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola und J. R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* **1984**, 81, 3684–3690.
- [58] T. A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, *J. Comput. Chem.* **1996**, 17, 490–519.
- [59] T. A. Halgren, Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions, *J. Comput. Chem.* **1996**, 17, 520–552.
- [60] T. A. Halgren, Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94, *J. Comput. Chem.* **1996**, 17, 553–586.
- [61] T. A. Halgren, Merck molecular force field. IV. conformational energies and geometries for MMFF94, *J. Comput. Chem.* **1996**, 17, 587–615.
- [62] T. A. Halgren, Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules, *J. Comput. Chem.* **1996**, 17, 616–641.
- [63] T. A. Halgren, MMFF VI. MMFF94s Option for energy minimization studies, *J. Comput. Chem.* **1999**, 20, 720–729.
- [64] T. A. Halgren, MMFF VII. Characterization of MMFF94 MMFF94s, and other widely available force fields for conformations energies and for Intermolecular Interaction energies and geometries, *J. Comput. Chem.* **1999**, 20, 730–748.
- [65] R. D. Cramer und J. D. Bunce, The DYLOMMS method: Initial results from a comparative study of approaches to 3D QSAR, in D. Hadzi und B. Jerman-Blasiz (Hg.), *QSAR in Drug Design and Toxicology*, Elsevier, Amsterdam, **1987** 3–12.
- [66] R. D. Cramer, D. E. Patterson und J. D. Bunce, Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.* **1988**, 110, 5959–5967.
- [67] R. D. Cramer und S. B. Wold, Comparative Molecular Field Analysis (CoMFA), online <http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=5025388> **1991**.

- [68] Tripos Inc., SYBYL 7.0, Computer Program.
- [69] J. W. M. Nissink, M. L. Verdonk, J. Kroon, T. Mietzner und G. Klebe, Superposition of molecules: Electron density fitting by application of fourier transforms, *J. Comput. Chem.* **1997**, 18, 638–645.
- [70] R. D. Cramer, Partial Least Squares (PLS): Its strengths and limitations, in Perspectives in Drug Discovery and Design, ESCOM, **1993**, 269–278.
- [71] M. Clark und R. D. Cramer, The Probability of chance correlations using partial least-squares (PLS), *Quant. Struct-Act. Relat.* **1993**, 12, 137–154.
- [72] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi und S. Clementi, Generating Optimal Linear PLS Estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems, *Quant. Struct-Act. Relat.* **1993**, 12, 9–20.
- [73] S. J. Cho und A. Tropsha, Cross-validated  $r^2$ -guided region selection for Comparative Molecular Field Analysis. A simple method to achieve consistent results, *J. Med. Chem.* **1995**, 38, 1060–1066.
- [74] R. Wang, Y. Gao, L. Liu und L. Lai, All-Orientation Search and All-Placement Search in Comparative Molecular Field Analysis, *J. Mol. Model. (Online)* **1998**, 4, 276–283.
- [75] R. D. Cramer und J. D. Bunce, The developing practice of Comparative Molecular Field Analysis, in H. Kubinyi (Hg.), 3D QSAR in Drug Design: Theory Methods and Applications, ESCOM, Leiden, **1993** 443–485.
- [76] U. Norinder, Single and domain made variable selection in 3D QSAR applications, *J. Chemom.* **1996**, 10, 95–105.
- [77] A. Golbraikh und A. Tropsha, Beware of  $q^2$ , *J. Mol. Graphics Modell.* **2002**, 20, 269–276.
- [78] A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K. Lee und A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models, *J. Comput.-Aided Mol. Des.* **2003**, 17, 241–253.
- [79] B. Bush und R. Nachbar, Sample-distance Partial Least Squares: PLS optimized for many variables, with application to CoMFA, *J. Comput.-Aided Mol. Des.* **1993**, 7, 587–619.
- [80] J. J. P. Stewart und et al., MOPAC 7, Computer Program **1995**.

- [81] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez und J. A. Pople, Gaussian 03 Revision B.04, Computer Program **2003**.
- [82] R. E. Chipkin, L. C. Iorio, V. L. Coffin, R. D. McQuade, J. G. Berger und A. Barnett, Pharmacological profile of SCH39166: a dopamine D1 selective benzonaphthazepine with potential antipsychotic activity, *J. Pharmacol. Exp. Ther.* **1988**, 247, 1093–1102.
- [83] J. G. Berger, W. K. Chang, J. W. Clader, D. Hou, R. E. Chipkin und A. T. McPhail, Synthesis and receptor affinities of some conformationally restricted analogues of the dopamine D1 selective ligand (5R)-8-chloro-2,3,4,5-tetrahydro-3-methyl-5-phenyl-1H-3-benzazepin-7-ol, *J. Med. Chem.* **1989**, 32, 1913–1921.
- [84] M. A. Tice, T. Hashemi, L. A. Taylor, R. A. Duffy und R. D. McQuade, Characterization of the binding of SCH39166 to the five cloned dopamine receptor subtypes, *Pharmacol., Biochem. Behav.* **1994**, 49, 567–571.
- [85] Chemical Computing Group Inc., Molecular Operating Environment (MOE), Computer Program **2004**.
- [86] G. E. Kellogg, S. F. Semus und D. J. Abraham, HINT: a new method of empirical hydrophobic field calculation for CoMFA, *J. Comput.-Aided Mol. Des.* **1991**, 5, 545–552.
- [87] TALETE, DRAGON, Computer Program **2005**.
- [88] K. Baumann, Distance Profiles (DiP): A translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules, *Quant. Struct.-Act. Relat.* **2002**, 21, 507–519.

- [89] N. Stiefl, G. Bringmann, C. RummeY und K. Baumann, Evaluation of extended parameter sets for the 3D-QSAR technique MaP: implications for interpretability and model quality exemplified by antimalarially active naphthylisoquinoline alkaloids, *J. Comput.-Aided Mol. Des.* **2003**, 17, 347–365.
- [90] N. Stiefl und K. Baumann, Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure-activity relationship technique, *J. Med. Chem.* **2003**, 46, 1390–1407.
- [91] N. Stiefl und K. Baumann, Structure-based validation of the 3D-QSAR technique MaP, *ChemInform* **2005**, 36, 739–749.
- [92] A. Klamt und G. Schüürmann, COSMO: A new approach to dielectric screening in solvents with expressions for the screening energy and its gradient, *J. Chem. Soc., Perkin Trans. 2* **1993**, 5, 799–805.
- [93] A. Schäfer, A. Klamt, D. Sattel, J. C. W. Lohrenz und F. Eckert, COSMO implementation in TURBOMOLE: Extension of an efficient quantum chemical code towards liquid systems, *Phys. Chem. Chem. Phys.* **2000**, 2, 2187–2193.
- [94] A. Klamt und F. Eckert, COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, *Fluid Phase Equilib.* **2000**, 172, 43–72.
- [95] A. Klamt, Conductor-like Screening Model for Real Solvents: A new approach to the quantitative calculation of solvation phenomena, *J. Phys. Chem.* **1995**, 99, 2224–2235.
- [96] A. Klamt und F. Eckert, COSMO-RS: a novel way from quantum chemistry to free energy, solubility and general QSAR-descriptors for partitioning, in H. Hoeltje und W. Sippl (Hg.), Rational Approaches to Drug design, Prous Science, **2001** 195–205.
- [97] A. Klamt, F. Eckert und M. Hornig, COSMO-RS: a novel view to physiological solvation and partition questions, *J. Comput.-Aided Mol. Des.* **2001**, 15, 355–365.
- [98] A. M. Zissimos, M. H. Abraham, A. Klamt, F. Eckert und J. Wood, A comparison between the two general sets of linear free energy descriptors of Abraham and Klamt, *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1320–1331.
- [99] A. Klamt, F. Eckert, M. Hornig, M. E. Beck und T. Bürger, Prediction of aqueous solubility of drugs and pesticides with COSMO-RS, *J. Comput. Chem.* **2002**, 23, 275–281.

- [100] C. Mehler, A. Klamt und W. Peukert, Use of COSMO-RS for the prediction of adsorption equilibria, *AIChE J.* **2002**, 48, 1093–1099.
- [101] A. Klamt, F. Eckert und M. Diedenhofen, Prediction of soil sorption coefficients with a conductor-like screening model for real solvents, *Environ. Toxicol. Chem.* **2002**, 21, 2562–2566.
- [102] H. Ikeda, K. Chiba, A. Kanou und N. Hirayama, Prediction of solubility of drugs by conductor-like screening model for real solvents, *Chem. Pharm. Bull. (Tokyo)* **2005**, 53, 253–255.
- [103] S. Oleszek-Kudlak, M. Grabda, E. Shibata, F. Eckert und T. Nakamurat, Application of the conductor-like screening model for real solvents for prediction of the aqueous solubility of chlorobenzenes depending on temperature and salinity, *Environ. Toxicol. Chem.* **2005**, 24, 1368–1375.
- [104] R. J. Oldland, Predicting phase equilibria using COSMO-based thermodynamic models and the VT-2004 sigma-profile, Diplomarbeit, Virginia Polytechnic Institute and State University **2004**.
- [105] C. DeBoor, A practical guide to splines, Springer, **2001**.
- [106] E. Parzen, On estimation of a probability density function and mode, *Ann. Math. Stat.* **1962**, 33, 1065–1076.
- [107] M. I. Glavinovic, Comparison of parzen density and frequency histogram as estimators of probability density functions., *Pfluegers Arch.* **1996**, 433, 174–179.
- [108] R. Ahlrichs, M. Bar, M. Haser, H. Horn und C. Kolmel, Electronic structure calculations on workstation computers: The program system turbomole, *Chem. Phys. Lett.* **1989**, 162, 165–169.
- [109] K. Eichkorn, F. Weigend, O. Treutler und R. Ahlrichs, Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials, *Theor. Chem. Acc.* **1997**, 97, 119–124.
- [110] M. Haeser und R. Ahlrichs, Improvements on the direct SCF method, *J. Comput. Chem.* **1989**, 10, 104–111.
- [111] A. Schaefer, C. Huber und R. Ahlrichs, Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr, *J. Chem. Phys.* **1994**, 100, 5829–5835.

- [112] O. Treutler und R. Ahlrichs, Efficient molecular numerical integration schemes, *J. Chem. Phys.* **1995**, 102, 346–354.
- [113] A. Klamt, V. Jonas, T. Bürger und J. C. W. Lohrenz, Refinement and parametrization of COSMO-RS, *J. Phys. Chem. A* **1998**, 102, 5074–5085.
- [114] S. Klod und E. Kleinpeter, Ab initio calculation of the anisotropy effect of multiple bonds and the ring current effect of arenes-application in conformational and configurational analysis, *J. Chem. Soc., Perkin Trans. 2* **2001**, 1893–1899.
- [115] M. Kaupp, J. Autschbach, F. Ban, R. J. Boyd, D. Berthomieu, D. A. Case, Z. Chen, D. M. Chipman, I. Ciofini, J. E. D. Bene, M. Bühl und V. G. Malkin, Calculation of NMR and EPR Parameters, Wiley-VCH, **2004**.
- [116] M. Kaupp und V. G. Malkin, Special issue quantum chemical calculations of NMR and EPR parameters, *J. Comput. Chem.* **1999**, 20, 1199–1327.
- [117] A. Bagno, Complete prediction of the  $^1\text{H}$  NMR spectrum of organic molecules by DFT calculations of chemical shifts and spin-spin coupling constants., *Chemistry* **2001**, 7, 1652–1661.
- [118] H. Friebolin, Ein- und zweidimensionale NMR-Spektroskopie: Eine Einführung, VCH, **1988**.
- [119] A. Höskuldsson, Centring and scaling of data, online <http://www.acc.umu.se/tnkjtg/chemometrics/editorial/mar2004.pdf> **2004**.
- [120] H. Kubinyi und U. Abraham, Practical problems in PLS analyses, in H. Kubinyi (Hg.), 3D QSAR in Drug Design: Theory Methods and Applications, ESCOM, Leiden, **1993** 717–728.