

**Entwicklung, Validierung und Anwendung einer
interpretierbaren und alignment-freien
4D-QSAR-Methodik**



Dissertation

zur Erlangung des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von

Dipl.-Biol.

Josef Heinrich Scheiber

aus Tirschenreuth

Würzburg 2006

**Entwicklung, Validierung und Anwendung einer
interpretierbaren und alignment-freien
4D-QSAR-Methodik**



Dissertation

zur Erlangung des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von

Dipl.-Biol.

Josef Heinrich Scheiber

aus Tirschenreuth

Würzburg 2006

Eingereicht am:

bei der Fakultät für Chemie und Pharmazie

1. Gutachter:

2. Gutachter:

der Dissertation

1. Prüfer:

2. Prüfer:

3. Prüfer:

des Öffentlichen Promotionskolloquiums

Tag des Öffentlichen Promotionskolloquiums:

Doktorurkunde ausgehändigt am:

Teile der vorliegenden Dissertation wurden bereits an folgenden Stellen veröffentlicht:

Originalpublikationen:

- (1) Gronwald W., Brunner K., Kirchhöfer R., Nasser A., Trenner J., Ganslmeier B., Riepl H., Ried A., Scheiber J., Elsner R., Neidig K.-P., Kalbitzer H.R.
AUREMOL, a New Program for the Automated Structure Elucidation of Biological Macromolecules. *Bruker Reports 154/155* (2004) 11-14
- (2) Kahle C., Deubner R., Schollmayer C., Scheiber J., Baumann K., Holzgrabe U.
NMR spectroscopic and molecular modelling studies on cyclodextrin-dipeptide inclusion complexes. *Eur. J. Org. Chem.* (2005), 1578-1589.
- (3) Scheiber J., Stiefl N., Baumann K.
xMaP: A novel 4D-QSAR technique based on molecular surface properties and conformer ensembles. In: *QSAR & Molecular Modelling in Rational Design of Bioactive Molecules; Eds. Aki (Sener) E., Yalcin I.*; (2006), 147-149
- (4a) Schmuck C., Heil M., Scheiber J., Baumann K.
Charge interactions do the job: A combined statistical and combinatorial approach to finding artificial receptors for binding tetrapeptides in water. *Angew. Chem. Int. Ed.* 44 (2005) 7208-7212
- (4b) Schmuck C., Heil M., Scheiber J., Baumann K.
Ladungswechselwirkungen machen es möglich: ein kombinierter statistischer und kombinatorischer Ansatz zur Auffindung künstlicher Rezeptoren für die Bindung von Tetrapeptiden in Wasser. *Angewandte Chemie* 117, 44 (2005) 7374-7379
- (5) Vicik R., Busemann M., Gelhaus C., Stiefl N., Scheiber J., Schmitz W., Schulz F., Mladenovic M., Engels B., Leippe M., Baumann K., Schirmeister T.
Selective cathepsin L inhibitors targeting both primed and non-primed substrate binding sites. *ChemMedChem* 1, 10 (2006) 1126-1141
- (6) Holzgrabe U., Kapkova P., Alptüzün V., Scheiber J., Kugelmann E.
Targeting Acetylcholinesterase to treat neurodegeneration. *Expert Opinion on Therapeutic Targets* (Review, eingereicht)
- (7) Degel B., Staib P., Scheiber J., Martina E., Baumann K., Morschhäuser J., Schirmeister T.
Cis-configured epoxides and aziridines as new inhibitors of *Candida albicans* SAP2 (eingereicht)
- (8) Scheiber J., Stiefl N., Baumann K.
xMaP: A novel interpretable alignment-free 4D-QSAR technique based on molecular surface properties and conformer ensembles (eingereicht)
- (9) Waibel B., Scheiber J., Meier C., Hammitzsch M., Baumann K., Scriba G., Holzgrabe U.
Comparison of cyclodextrine-dipeptide inclusion complexes in absence and presence of urea by means of capillary electrophoresis, nuclear magnetic resonance and molecular modeling (eingereicht)

Vorträge:

- (1) Scheiber J., Stiefl N., Baumann K.
„Robustness of the new xMaP 4D-QSAR technique“
Doktorandentagung der Deutschen Pharmazeutischen Gesellschaft (DPhG)
Universität Leipzig, Februar **2005**
- (2) Scheiber J., Stiefl N., Baumann K.
“Conformational flexibility in QSAR analyses: The new alignment-free 4D-QSAR technique xMaP“
19. Darmstädter Molecular Modelling Workshop (MGMS-DS)
Universität Erlangen, Mai **2005**
- (3) Scheiber J., Baumann K.
“The alignment-free 4D-QSAR technique xMaP - Combining ligand- and target-based data”
Unilever Center of Molecular Informatics;
Universität Cambridge, Großbritannien, April **2006**
- (4) Scheiber J., Baumann K.
“The alignment-free 4D-QSAR technique xMaP – Applications and extensions”
20. Darmstädter Molecular Modelling Workshop (MGMS-DS)
Universität Erlangen, Mai **2006**
- (5) Scheiber J., Baumann K.
“Alignment-free 4D-QSAR with xMaP: Designing dopamine antagonists”
Institut für Pharmazie, Universität Jena, Juli **2006**
- (6) Scheiber J., Baumann K.
“Alignment-free 4D-QSAR using xMaP: Enhancing model interpretation through target structure data”
Doktorandentagung der Deutschen Pharmazeutischen Gesellschaft (DPhG)
Nürnberg-Heroldsberg, September **2006**
- (7) Scheiber J., Enzensperger Ch., Lehmann J., Baumann K.
“4D-QSAR: xMaP in real world applications – A valuable tool in designing novel dopamine-receptor antagonists”
Summer School Medicinal Chemistry (Kurzvortrag)
Regensburg, September **2006**

Posterpräsentationen (Auswahl von insgesamt 29 Posterbeiträgen):

- (1) Scheiber J., Stiefl N., Baumann K.
xMaP: A novel 4D-QSAR technique based on molecular surface properties and conformer ensembles; EURO-QSAR 2004, Istanbul, Türkei, September **2004**
- (2) Scheiber J., Stiefl N., Baumann K.
xMaP – Extended validation of a 4D-QSAR technique
DPhG-Jahrestagung, Universität Regensburg, Oktober **2004**
- (3) Scheiber J., Stiefl N., Baumann K.
Conformational flexibility in QSAR analyses – The alignment-independent 4D-QSAR technique xMaP
19. Darmstädter Molecular Modelling Workshop, Universität Erlangen, Mai **2005**
- (3) Scheiber J., Stiefl N., Baumann K.
Incorporating conformational flexibility into QSAR – Validation of a novel alignment-independent 4D-QSAR technique
International Conference Chemical Structures, Noordwijkerhout, Niederlande, Juni **2005**
- (4) Scheiber J., Baumann K.
Employing target structure information in 4D-QSAR analysis
MGMS Annual Meeting „From prediction to practice“, Trinity College, Dublin, Irland, September **2005**
- (5) Scheiber J., Baumann K.
A novel protein-structure based 4D-QSAR approach
Summer School Medicinal Chemistry, Shanghai Institute of Organic Chemistry, Shanghai, China, September/Okttober **2005**
- (6) Scheiber J., Bäumert J., Baumann K.
Alignment-free 4D-QSAR using xMaP: Optimization of dopamine antagonists
Spring Meeting of the UK QSAR Society, Cambridge, Großbritannien, April **2006**
- (7) Scheiber J., Baumann K.
Alignment-free 4D-QSAR: Applying the xMaP technique in prospective analyses
Workshop Cheminformatics in Europe, Obernai, Frankreich, Juni **2006**
- (8) Scheiber J., Enzensperger Ch., Lehmann J., Baumann K.
4D-QSAR: xMaP in real world applications
Summer School Medicinal Chemistry, Regensburg, Oktober **2006**

*Lebensklugheit bedeutet:
Alle Dinge möglichst wichtig,
aber keines völlig ernst nehmen.*

Arthur Schnitzler (1862-1931)
österreichischer Erzähler und Dramatiker

Danke!

Diese Stelle will ich dazu nutzen, all denjenigen zu danken die mich direkt und indirekt während der Entstehung dieser Arbeit unterstützt haben.

An allererster Stelle steht das größte Danke und das geht an meinen Chef & Betreuer Prof. Dr. Knut Baumann. Die vielen fruchtbaren Diskussionen mit ihm und umfangreiche Erklärungen seinerseits ermöglichten es mir, immer weiter in das Gebiet des computergestützten Wirkstoffdesigns und der Chemometrie einzusteigen und neben den Grundlagen auch tief gehende Zusammenhänge problemlos zu verstehen. Deshalb konnte ich viele eigene Ideen finden, weiterentwickeln, verbessern und mit anderen Sachen verknüpfen. Ganz toll war auch die Möglichkeit an so vielen interessanten Tagungen teilzunehmen (es waren über 30000 Flugkilometer!); Besonders erwähnenswert ist auch sein Einsatz in der Endphase dieser Arbeit! Lieber Knut, vielen Dank für alles!

Ein ganz großes Danke geht an meine Kollegen im AK Baumann: Die Zusammenarbeit mit euch hat wirklich viel Spaß gemacht! Bei Sebastian „Basti“ Rohrer bedanke ich mich für die alltägliche Zusammenarbeit in den verschiedensten Projekten sowie für das Korrekturlesen dieser Arbeit. Ulrike „Schmuli“ Schmid danke ich für die Beantwortung vieler Fragen zur Klassifizierung und ganz besonders für die netten Kinoabende. Markus Kossner danke ich für die gute Zusammenarbeit in seinem Praktischen Jahr und seit dem Beginn seiner Promotion. Euch Allen viel Spaß & Erfolg in Braunschweig! Auch die Ex-Mitglieder des AKs – Matthias Busemann, Patrick Dötsch und Nik Stiefl – will ich an dieser Stelle nicht vergessen.

Meinen Wahlpflichtfach-Praktikanten bin ich zu Dank verpflichtet: Johannes Bäumert hat beim D₁-Datensatz ein Super-Engagement gezeigt und viele interessante Daten zu Tage gefördert, sowie bei der Interpretation viele gute eigene Ideen einfließen lassen. Erst damit wurde das Mechanismusmodell möglich! Sonja Rietschel war bei der Erstellung des Hand-Alignments für den GK-Datensatz eine sehr wertvolle Hilfe und hat mir viel Zeit erspart.

Dann will ich Prof. Dr. Ulrike Holzgrabe sehr für die gute Zusammenarbeit bei allen möglichen Projekten danken, das hat mir einen tiefen Einblick in die medizinische Chemie ermöglicht. Aus ihrem AK habe ich speziell mit Claudia Meier und Benny Waibel im Rahmen des Cyclodextrin-Projektes sowie mit Eva Kugelmann beim AChE-Projekt gut zusammenarbeitet. Vielen Dank dafür! Bei Ebi Heller bedanke ich mich für die Beantwortung

vieler chemischer Fragen, sowie die Organisation und Durchführung vieler Weinproben und anderer Festivitäten. Außerdem danke ich noch Dani Brinz für mehr als eine Tasse Kaffee ;-)
Zu gewaltigem Dank bin ich Dr. Bernd Reyer verpflichtet, der für jegliche Art von Computerproblemen immer ein offenes Ohr hatte und schnell innovative Lösungen fand. Dr. Curd Schollmayer danke ich für seinen immerwährenden Beistand unter lauter Fußball-Banausen!

Bei Prof. Dr. Tanja Schirmeister bedanke ich mich herzlich für die gute Zusammenarbeit bei den verschiedenen hochinteressanten Projekten, mit dem Ziel, Proteasen zu inhibieren. Besonders bei Björn Degel, Radim Vicik und Christian Büchold möchte ich mich für die angenehme Kooperation bedanken!

Ein riesengroßes Danke geht an Prof. Dr. Jochen Lehmann und Christoph Enzensperger von der Universität Jena: Der D₁-Datensatz war ein tolles Spielfeld für xMaP. Herzlichen Dank für die Zurverfügungstellung! Besonders Christoph möchte ich für die gute und extrem interessante Kooperation danken. Danke auch für die Vortragseinladung nach Jena!

Ein ganz besonderes Dankeschön geht an Prof. Dr. Bernd Engels für die hervorragende und sehr angenehme Zusammenarbeit bei der Organisation des SFB-Symposiums. Dabei konnte ich sehr viel lernen!

Prof. Dr. Gerhard Bringmann danke ich sehr herzlich einerseits für die effektive Zusammenarbeit beim SFB-Symposium sowie andererseits für die hochinteressante wissenschaftliche Kooperation. Aus seinem Arbeitskreis möchte ich mich auch bei Tanja Gulder und Christian Winter bedanken.

Prof. Dr. Carsten Schmuck sei für die gute Zusammenarbeit beim Tripeptid-Projekt gedankt, woraus sich ja auch eine schöne Publikation ergeben hat.

Anagnostis „Noti“ Valotis und Prof. Dr. Petra Högger danke ich für den Glukokortikoid-Datensatz und die vielfältigen Erklärungen dazu.

Besonders in der Anfangszeit meiner Arbeit hat mir Katalin Nadassy von Accelrys in Cambridge geduldig die dümsten Fragen zu Catalyst beantwortet. Danke!

Prof. Dr. Gisbert Schneider von der Universität Frankfurt danke ich für die gute Kommunikation seit Dublin und die Chance, bereits im Rahmen meiner Doktorarbeit als Gutachter für Veröffentlichungen tätig werden zu können. Außerdem danke ich aus seinem Arbeitskreis Uli Fechner ganz besonders für die freundliche Aufnahme in Frankfurt!

Ganz toll wurde ich in Cambridge von Dr. John Mitchell und Dr. Noel O'Boyle aufgenommen. Vielen Dank, für die Chance, diesen Vortrag zu halten, sowie die sehr interessante und anekdotenreiche Stadtführung. Dr. Andreas Bender danke ich für die Vermittlung dieses Vortrags.

Prof. Dr. Tony Hopfinger – dem Entwickler der ersten 4D-QSAR-Technik – danke ich für die interessanten Diskussionen bei seinem Besuch in Würzburg. Außerdem herzlichen Dank für die Flasche Wein, das war eine ganz besondere Ehre!

Folgenden Leuten bin ich wegen der Bereitstellung von Datensätzen zu Dank verpflichtet: Prof. Dr. Mitchell Avery von der University of Mississippi für den Artemisinin-Datensatz. Dr. Andreas Göller von Bayer Healthcare aus Wuppertal für seine Daten, sowie Prof. Dr. Irmgard Merfort aus Freiburg für ihre NF- κ B-Daten. Das hat bei der Evaluierung von xMaP enorm geholfen!

Der deutschen Sektion der Molecular Graphics and Modelling Society danke ich für zwei Reisestipendien die die beiden Tagungsteilnahmen in Erlangen 2005 und 2006 erst ermöglicht haben. Dem ASIA LINK Medicinal Chemistry der Universität Regensburg danke ich sehr herzlich für den Trip nach Shanghai.

Für das Erträglich- und sogar Spaßiggestalten der Erstsemesterbetreuung bedanke ich mich sehr herzlich bei meinen Mitassistenten Radim, Björn, Thomas, Birgit, Tänscha, Eva und Moni. Außerdem danke ich den Erstis dafür, dass sie immer sehr kooperativ waren. Mein schlimmster Alptraum als zu betreuender Student bin nach wie vor ich selbst ;-). Da konnten auch vier Jahrgänge Pharmazie-Erstsemester nichts dran ändern ...

Bei meinen Regensburger Studienkollegen bedanke ich mich herzlichst für die eine oder andere Feier in diversen Städten sowie den regen Austausch von Papern quer durch die Republik ;-)

Meinen Eltern danke ich für die immerwährende Unterstützung seit ich denken kann. Ohne das würde ich heute nicht da stehen wo ich jetzt bin.

Meiner Freundin Steffi danke ich sehr herzlich dafür, dass sie mich in den letzten Wochen bei all dem Theater um das Fertigwerden tapfer ertragen hat :-) Danke für Alles!

Zu guter Letzt danke ich allen, die ich hier in dieser Aufzählung vergessen habe und die eigentlich auch hier stehen müssten! Ich habe euch nicht absichtlich vergessen!

Inhaltsverzeichnis

1.	Einführung und Zielsetzung	1
2.	Theoretische Grundlagen	6
2.1.	Moleküldeskriptoren	6
2.1.1	„Klassische“ Deskriptoren	6
2.1.2	Zweidimensionale Deskriptoren	7
2.1.3	Dreidimensionale Deskriptoren.....	8
2.1.3.1	Translations- und Rotationsvariante Deskriptoren.....	9
2.1.3.2	Translations- und Rotationsinvariante Deskriptoren.....	12
2.1.4	Die vierte Dimension in der QSAR.....	26
2.1.4.1	Die Methode von Hopfinger	26
2.1.4.2	Die Methode von Dobler und Vedani	28
2.1.4.3	Weitere Ansätze	29
2.1.5	Fazit zu bisherigen Deskriptorentwicklungen.....	29
2.2.	Mathematische Modellierung.....	31
2.2.1	Verwendete Notation und grundlegende Operationen	31
2.2.2	Wichtige Schritte zur Datenvorbehandlung	32
2.2.2.1	Zentrierung	33
2.2.3	Regressionstechniken	33
2.2.3.1	Einfache Lineare Regression.....	33
2.2.3.2	Multiple lineare Regression(MLR)	33
2.2.3.3	Singulärwertzerlegung (SVD).....	35
2.2.3.4	Hauptkomponentenregression (PCR).....	37
2.2.3.5	Partial Least Squares Regression (PLS).....	37
2.2.4	Datenmodellierung und -validierung	37
2.2.4.1	Kreuzvalidierung.....	39
2.2.4.2	Gütekriterien.....	41
2.2.4.3	Ensemble averaging	42
2.3.	Variablenselektion.....	44
2.3.1	Der Suchalgorithmus.....	45
2.3.2	Die Gütefunktion.....	46
3.	Ergebnisse und Diskussion.....	49
3.1.	xMaP – eine interpretierbare alignment-freie 4D-QSAR-Methodik.....	49
3.2.	Verwandtschaft zu den Vorgängertechniken	49
3.3.	Die Berechnung des Deskriptors	51
3.3.1	Berechnung der Konformerensembles	52
3.3.2	Oberflächenberechnung für die Konformere	53
3.3.3	Verschmelzung von Oberflächenbereichen mit identischen Eigenschaften	55
3.3.4	Charakterisierung der Oberflächeneigenschaften durch Deskriptoren	55
3.3.5	Umsetzung der Konformerendaten zu Moleküldeskriptoren	58
3.4.	Erstellung und Validierung der Modelle	59
3.5.	Interpretation und Visualisierung der Modelle	60
3.5.1	Dreidimensionale Rückprojektion der xMaP-Daten	61
3.5.2	Zweidimensionale Rückprojektion der xMaP-Daten	62
3.5.3	Fazit zur Interpretierbarkeit.....	63
3.6.	Einbindung von Informationen über die Targetstruktur	63
3.7.	Untersuchte Datensätze	65
3.7.1.	Inhibitoren der Acetylcholinesterase (Abkürzung: AZT)	65
3.7.2.	Prostaglandin F _{2α} -Analoge (PGF _{2α})	67
3.7.3.	Modulatoren des muskarinischen M ₂ -Rezeptors (M ₂).....	68

3.7.4. Inhibitoren der HIV-1 Reversen Transkriptase (HEPT)	69
3.7.5. Dopamin-Antagonisten (D ₁)	70
3.7.6. Glukokortikoide (GK)	72
3.7.7. Naphtylisochinolin-Alkaloide (NIQ)	74
3.7.8. Weitere Datensätze	75
3.8. Standardparameter und Parametervariationen	75
3.8.1. Einfluss der Konformerenberechnung	76
3.8.2. Einfluss des „Energiefensters“	78
3.8.3. Einfluss der Konformerenwichtung	82
3.8.4. Einfluss der strukturbasierten Konformerenberechnung.....	84
3.8.5. Einfluss der Regressionstechnik.....	86
3.8.6. Einfluss der Variablenselektion	87
3.8.7. Ausgewählte Standardparameter.....	88
3.9. Ergebnisse der Modellbildung	90
3.9.1. Inhibitoren der Acetylcholinesterase (AZT)	91
3.9.2. Prostaglandin F _{2α} -Analoga (PGF _{2α})	104
3.9.3. Modulatoren des muskarinischen M ₂ -Rezeptors (M ₂).....	110
3.9.4. Inhibitoren der HIV-1 Reversen Transkriptase (HEPT)	116
3.9.5. Dopamin-Antagonisten (D ₁)	122
3.9.5.1 Ein erstes QSAR-Modell.....	122
3.9.5.2 Vorhersagen bereits synthetisierter Moleküle.....	127
3.9.5.3 Strukturvorschläge	129
3.9.5.4 Aktualisierung des bestehenden Modells.....	131
3.9.5.5 Docking in Homologiemodell und strukturbasierte Interpretation	132
3.9.5.6 xMaP-Modelle für Antagonisten weiterer Dopaminrezeptoren.....	136
3.9.5.6 Fazit für den Dopamin-Datensatz	137
3.9.6. Glukokortikoide (GK)	138
3.9.6.1 Der Einfluss der Lipophilie.....	138
3.9.6.2 Das xMaP-Modell für den GK-Datensatz.....	139
3.9.6.3 CoMFA- und CoMSIA-Modelle für den GK-Datensatz.....	143
3.9.7. Naphtylisochinolin-Alkaloide (NIQ)	146
3.9.8. Weitere Datensätze.....	148
4. Zusammenfassung und Ausblick	151
Summary and Outlook	155
Anhang I. Wichtige Abkürzungen und Symbole	159
Anhang II. Strukturen und biologische Aktivitäten der untersuchten Datensätze	161
II.1. Inhibitoren der Acetylcholinesterase (AZT).....	161
II.2. Prostaglandin F _{2α} -Analoga (PGF _{2α}).....	165
II.3. Modulatoren des muskarinischen M ₂ -Rezeptors (M ₂)	169
II.4. Inhibitoren der HIV-1 Reversen Transkriptase (HEPT).....	174
II.5. Dopamin-Antagonisten (D ₁).....	176
II.6. Glukokortikoide (GK)	182
II.7. Naphtylisochinolin-Alkaloide (NIQ).....	185
Anhang III. Der Ein- und Dreibuchstabencode für Aminosäuren	188
Literaturverzeichnis.....	189
Lebenslauf.....	205

1. Einführung und Zielsetzung

„Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glucosid wie Schloss und Schlüssel zueinander passen müssen, um eine chemische Wirkung aufeinander ausüben zu können.“

Diese vor nunmehr 112 Jahren vom Chemiker Emil Fischer geprägte Metapher beschreibt das so genannte „Schlüssel-Schloss-Prinzip“, die Grundlage der modernen medizinischen Chemie [1]: Die Suche nach immer neuen und besseren „Schlüsseln“, die es ermöglichen, verschiedenste Krankheiten behandelbar zu machen. Mit dem Ziel, diesen Vorgang zu beschleunigen, sind in den letzten Jahren die computerbasierten Methoden zur Wirkstoffentwicklung immer mehr in den Fokus der Aufmerksamkeit gerückt.

Die Wirkstoffsuche am Computer wird per Definition in zwei unterschiedliche Bereiche untergliedert: Einerseits die ligandbasierten Techniken, was nach Fischers Bild bedeutet, dass über das Schloss keinerlei Informationen bekannt sind. Es sind jedoch verschiedene mehr oder minder gute Schlüssel bekannt, aus denen mit verschiedenen Methoden wichtige Informationen extrahiert werden können.

Andererseits gibt es die strukturbasierten Techniken, bei denen die genaue Struktur des Schlosses, also eines Rezeptors oder Enzyms, bekannt ist. Mit dieser Grundlage können dann passende Schlüssel entwickelt werden.

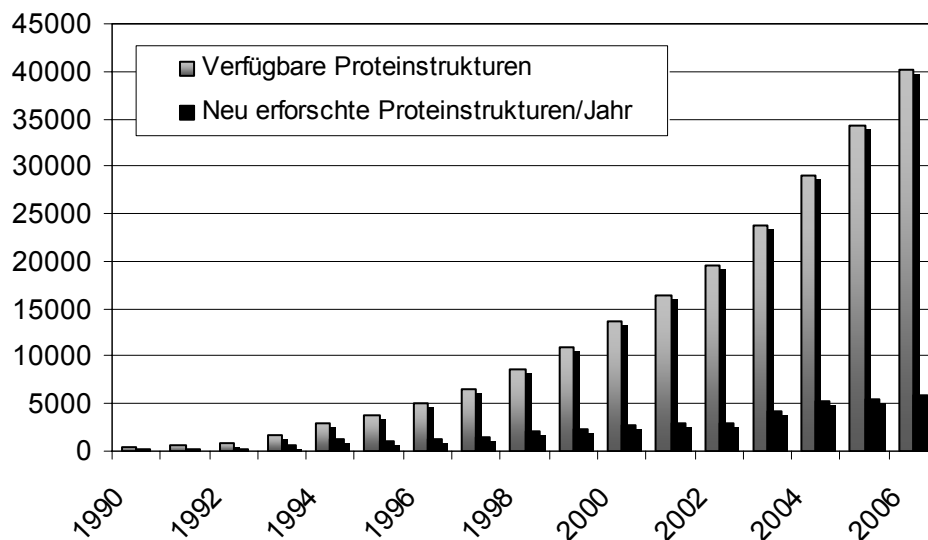


Abbildung 1: Die Entwicklung der Anzahl bekannter Proteinstrukturen in der Protein-Datenbank (PDB) der Forschungsgemeinschaft für Bioinformatik (RCSB) seit 1990 (Stand: November 2006)

Die Anzahl der aufgeklärten Proteinstrukturen steigt seit einigen Jahren rasant an (siehe Abbildung 1), von etwa 40000 Proteinen ist bisher die Struktur bestimmt worden [2,3]. Dennoch sind die rein ligandbasierten Methoden weiterhin eines der zentralen Hilfsmittel der modernen computerbasierten Wirkstoffentwicklung: Die Sequenzdatenbank UniProt/TrEMBL enthält derzeit (Oktober 2006) rund 3.3 Millionen Einträge zu sequenzierten Proteinen [4-6]. Eine genaue dreidimensionale Struktur ist demzufolge erst für gut ein Prozent der bisher bekannten Proteine bestimmt worden. Für eine Vielzahl der pharmazeutisch relevanten Zielstrukturen ist es daher nicht möglich, Informationen über ihre dreidimensionale Struktur bei der Wirkstoffentwicklung zu Rate zu ziehen. Daher werden die ligandbasierten Methoden weiterhin eines der zentralen Hilfsmittel der computerbasierten Wirkstoffentwicklung sein.

Die wichtigste Methode aus diesem Bereich ist die QSAR (*engl.*: Quantitative Structure-Activity Relationships; Quantitative Struktur-Wirkungsbeziehungen). Ziel hierbei ist es, durch Verwendung einer mathematischen Funktion die Verbindung zwischen der biologischen Aktivität eines Moleküls und seinen chemischen sowie physikalischen Eigenschaften zu finden. Ein derartiger Zusammenhang zwischen Struktur und Eigenschaften wurde erstmals im Jahre 1844 von Kopp hergestellt [7]. Er erforschte den Zusammenhang zwischen dem Siedepunkt von Alkanen und ihren Kettenlängen. 24 Jahre später, also 1868, postulierten Crum-Brown und Fraser erstmals die Verbindung zwischen physiologischer Aktivität Φ und chemischer Konstitution C mit einer mathematischen Funktion [8]:

$$\Phi = f(C)$$

Damit ist das zentrale Paradigma einer QSAR-Analyse eigentlich schon vollständig erklärt: Die zu untersuchenden Moleküle müssen in einer mathematisch fassbaren Art und Weise beschrieben werden. Das geschieht mit den so genannten Deskriptoren, die im Folgenden noch genauer definiert werden. Zwischen den Deskriptoren und der bekannten biologischen Aktivität kann der funktionelle Zusammenhang mittels einer mathematischen Funktion hergestellt werden. Diese Funktion kann wiederum dazu genutzt werden, die Aktivität noch nicht synthetisierter oder noch nicht am gewünschten Zielprotein getesteter Moleküle über deren Deskriptoren vorherzusagen. Die wichtigste Grundvoraussetzung hierbei ist eine eindeutig bestimmte und damit quantifizierbare biologische Aktivität der zu untersuchenden Moleküle. Damit ist sehr kostengünstig eine gezielte Priorisierung der zu testenden Moleküle möglich. Deshalb wird die QSAR auf breiter Basis sowohl in der Industrie als auch im

akademischen Umfeld immer dann eingesetzt, wenn Wirkstoffe optimiert werden sollen. Dieser Kostenfaktor wird in den nächsten Jahren zu einem immer stärkeren Einsatz von Computermethoden führen. Vor allem die Industrie steht sehr stark unter dem Druck, den Preis für die Entwicklung eines neuen Arzneimittels zumindest nicht noch weiter ansteigen zu lassen.

Es überrascht daher nicht, dass dem Gebiet der QSAR in den letzten Jahren sehr viel Aufmerksamkeit geschenkt wurde. Das hat sich insbesondere in der Entwicklung von neuen Deskriptoren niedergeschlagen. Einen umfassenden Überblick zu dieser Thematik bietet das Standardwerk von Todeschini und Consonni [9]. Für viele bei der Optimierung von Wirkstoffen auftretende Probleme gibt es Ansätze, diese mit Hilfe verschiedener QSAR-Methoden zu lösen. Deshalb sind viele unterschiedliche Deskriptoren zur Beschreibung der Vielzahl von Molekülparametern erforderlich. Erst dadurch können die QSAR-Ansätze an das entsprechende Problem angepasst werden.

Ein in den letzten Jahren auffallender Trend ist die Entwicklung von Deskriptoren, die ohne großen Benutzereinfluss zu berechnen und sehr leicht zu interpretieren sind [10-14]. Diese Interpretierbar- und somit auch Kommunizierbarkeit der Ergebnisse ist einer der zentralen Punkte bei der Entwicklung von neuartigen Deskriptoren. Genau diese vereinfacht nämlich die Interaktion von Computerchemikern mit denjenigen Chemikern, die neue Moleküle anschließend auch synthetisieren sollen. Ist diese Zusammenarbeit erfolgreich, kann der Prozess der Arzneistoffentwicklung beschleunigt werden. Die erwünschte Reduktion des Nutzereinflusses liegt darin begründet, dass im Normalfall immer das gleiche Ergebnis am Ende stehen soll, unabhängig davon, welche Person das Experiment durchführt.

Parallel zur Reduktion des Nutzereinflusses, aber bisher nicht in Kombination damit wurde in den letzten Jahren ein weiteres Problem in Angriff genommen. Die Standardmethoden der QSAR berücksichtigen immer nur ein einziges Konformer. Von diesem wird angenommen, dass es das biologisch aktive der zu untersuchenden Moleküle ist. Diese Annahme stellt natürlich eine grobe Vereinfachung dar, da es die Flexibilität der untersuchten Moleküle nicht widerspiegelt. Es gibt in den letzten Jahren sehr viele interessante Ansätze zur Reduzierung und Eliminierung dieses Konformer-Auswahl-Problems [15-18].

Die angesprochenen Punkte werden in den folgenden Kapiteln genauer erläutert. Die Intention dieser Aufzählung ist folgende: Es ist unmöglich, einen Strukturdeskriptor zu entwickeln, der für alle Herausforderungen in der Optimierung von Molekülen gleich gut geeignet ist. Nichtsdestotrotz ist es möglich, einen Anforderungskatalog für aktuelle, viel versprechende Deskriptoren zu definieren:

- leichte Interpretierbarkeit
- Visualisierung der Ergebnisse
- Reduzierung bzw. komplette Entfernung der potenziellen Ergebnisverzerrung (Überoptimierung) durch den Benutzer:
 - (a) Unabhängigkeit von der räumlichen Ausrichtung der Moleküle
 - (b) Eliminierung des Konformerenauswahl-Problems
- Schnelle Berechenbarkeit
- Breite Anwendbarkeit und Optimierbarkeit auf verschiedene Problemstellungen

Ziel dieser Arbeit war die Entwicklung eines neuen Strukturdeskriptors, der den genannten Anforderungen bestmöglich gerecht wird und weitere Rahmenbedingungen erfüllt. Wie seine Vorgängertechnik [14,19] sollte auch der neu zu entwickelnde Deskriptor die Moleküle über deren Eigenschaftsverteilung auf ihrer Oberfläche charakterisieren. Dies ist erstrebenswert, da dadurch genau der für die Wechselwirkungen mit dem Zielmolekül entscheidende Bereich charakterisiert wird. Es wird quasi die Außenseite des Schlüssels und nicht sein Innenleben beschrieben. Darüber hinaus sollte der Deskriptor unabhängig von der Position der Moleküle im Raum sein, was über eine Beschreibung über ein internes Koordinatensystem erfolgen sollte. Im Gegensatz zu fast allen bisher entwickelten Techniken sollte der Deskriptor nicht auf einem einzelnen Konformer basieren. Stattdessen sollte die komplette Flexibilität der zu untersuchenden Moleküle mitmodelliert werden.

Die Kombination dieser beiden Anforderungen führt dazu, dass der Einfluss eines Anwenders faktisch eliminiert wird. Es gab bisher keinen ernsthaften Versuch, beide Voraussetzungen in einer Technik zu kombinieren. Von diesem Standpunkt aus betrachtet ist die in dieser Arbeit beschriebene Technik also einzigartig. Analog zur Vorgängertechnik sollte die Oberfläche durch Radialverteilungsfunktionen beschrieben werden. Diese sind extrem gut interpretierbar, was einen Vorsprung gegenüber möglichen Konkurrenztechniken verschafft. Alle genannten Punkte konnten im Rahmen dieser Arbeit erfolgreich umgesetzt werden.

Um die Identifizierung relevanter Informationen zu ermöglichen, sollte der neue Deskriptor auch in eine völlig neue 4D-QSAR-Methodik integriert werden. Dies sollte unter Verwendung von Variablenselektion, Visualisierung der selektierten Variablen und einer sehr anspruchsvollen Validierung der erhaltenen Modelle geschehen.

Überdies sollte die entwickelte Methode sowohl an Benchmarkdatensätzen aus verschiedensten Bereichen der pharmazeutischen Forschung als auch an Datensätzen der Fakultät für Chemie und Pharmazie eingesetzt werden. Speziell im Rahmen von Kooperationen innerhalb des Sonderforschungsbereiches 630 sollte die Technik eingesetzt werden. Außerdem sollten, wenn möglich, interessante externe Datensätze untersucht werden.

2. Theoretische Grundlagen

Der folgende Teil lässt sich im Wesentlichen in zwei Bereiche untergliedern. Dies geschieht analog zu den im Rahmen dieser Arbeit eingesetzten Techniken. Im ersten Teil werden verschiedene zum neuen Deskriptor verwandte und grundlegende Techniken detaillierter erläutert und klassifiziert. Dies dient der Veranschaulichung im Vergleich zur kompletten Deskriptorenfamilie.

Der zweite Teil beschreibt die mathematische Modellierung der Daten und die Methoden zur Validierung der mathematischen Modelle. Dafür werden alle eingesetzten Methoden zusammen mit ihren Vorläufern erläutert, insofern solche existieren. Dies geschieht ebenfalls mit dem Ziel, die angewendeten Methoden einzugruppieren und abzugrenzen.

2.1. Moleküldeskriptoren

Der folgende Teil bietet einen kurzen, umfassenden Überblick über viele Deskriptoren, die früher häufig eingesetzt wurden oder derzeit eingesetzt werden. Ein besonderes Augenmaß liegt dabei auf der Beschreibung von Techniken, die der hier entwickelten Technik entweder als Vorläufer zugrunde liegen oder aber sie in verschiedenen Punkten inspiriert haben. In der QSAR wichtige Methoden, die keinen direkten Einfluss auf diese Arbeit hatten, werden nur kurz erwähnt.

2.1.1 „Klassische“ Deskriptoren

Als klassische Deskriptoren werden alle diejenigen bezeichnet, die nur einen einzigen speziellen Molekülparameter charakterisieren. Dazu zählen auch im Labor bestimmte Messwerte. Diese Deskriptoren werden häufig auch als eindimensional bezeichnet. Ihr Einsatz lässt sich insbesondere auf die bahnbrechende Publikation von Hansch, Maloney und Fujita im Jahre 1962 zurückführen [20]. Dabei werden eindeutig charakterisierte Parameter eines Moleküls, wie beispielsweise der Oktanol-Wasser-Verteilungskoeffizient, mit bestimmten, zu optimierenden molekularen Eigenschaften korreliert. Daraus können entsprechende Schlüsse zur Weiterentwicklung gezogen werden. Weitere Beispiele für derartige Deskriptoren sind die molekulare Masse, das Volumen oder beispielsweise die Anzahl der drehbaren Bindungen im untersuchten Molekül. Diese Parameter können sehr schnell und einfach bestimmt werden. Daher werden sie häufig zum Filtern großer

Datenbanken eingesetzt, prominentestes Beispiel hierfür ist Lipinski's „*Rule of Five*“[21]. Andererseits müssen viele derartige Parameter im Labor bestimmt werden. Der gravierendste Nachteil dieser im Labor bestimmten Werte ist, dass sie nicht für Vorhersagen von nicht synthetisierten Molekülen eingesetzt werden können.

2.1.2 Zweidimensionale Deskriptoren

Im Gegensatz zu den eindimensionalen Deskriptoren betrachten zweidimensionale Deskriptoren Moleküle nicht als Ganzes. Die Moleküle werden nicht mehr direkt charakterisiert, sondern ihre Zusammensetzung über verschiedene Maßzahlen beschrieben. Zweidimensionale Deskriptoren charakterisieren den Aufbau eines Moleküls aus dessen Atomen und den Bindungen zwischen den Atomen. Das kommt der Darstellung eines Moleküls aus der Perspektive eines Chemikers am Nächsten. Bei Analysen mit zweidimensionalen Deskriptoren wird die Information über das Vorhandensein bestimmter Atomgruppen kodiert. Wichtige Vertreter sind die topologischen Indizes [22-25], die Fragmentvektoren [26-30] und die Atompaaire [31].

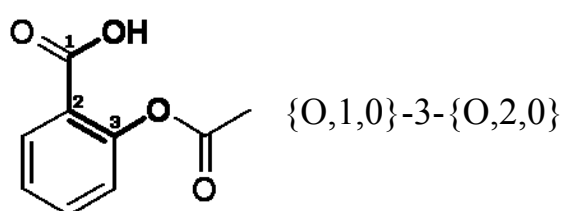
Die topologischen Indizes beruhen darauf, ein Molekül als mathematischen Graph zu beschreiben. Dieser Graph beschreibt neben den Atomtypen die Verknüpfungen untereinander. Diese Information wird in der Regel als so genannte Adjazenzliste im Rechner verarbeitet[32]. Ein solcher Index ist eine eindeutige Maßzahl für die Beschaffenheit des mathematischen Graphs und damit auch direkt für die chemische Zusammensetzung des Moleküls. Derartige Indizes sind extrem einfach und schnell zu berechnen und wurden häufig erfolgreich eingesetzt. Unglücklicherweise können die Ergebnisse nicht interpretiert werden.

Die Fragmentvektoren beschreiben die Moleküle durch die Häufigkeit verschiedener Substrukturelemente. Damit ist eine sehr einfache Darstellung von Molekülen möglich und des Weiteren sind darauf beruhende QSAR-Modelle meist leicht interpretierbar. Zu Problemen in der Anwendung der Fragmentvektoren führen bei Vorhersagen Substrukturen, die nicht in den Trainingsdaten vorhanden waren. Trotzdem ist die Art der Molekülbeschreibung über Fragmente sehr erfolgreich. Sie hat jedoch keinen Einfluss auf die hier entwickelte Technik genommen und soll deshalb nicht weiter erläutert werden. Für einen genaueren Einblick wird auf die zitierten Referenzen verwiesen [26-30].

Relativ ähnlich zu den Fragmentvektoren ist der so genannte Atompaaire-Deskriptor. Hierbei werden ausgehend von einem Startatom alle Atome, die sich in einem bestimmten Abstand befinden, mittels des zu Grunde liegenden Atomtyps als Molekülfragmente kodiert:

$$\langle \text{Atomtyp}_1 \rangle - \langle \text{Abstand} \rangle - \langle \text{Atomtyp}_2 \rangle$$

Hierbei ist der Abstand zwischen den beiden relevanten Atomen definiert als die Anzahl alternierender Sequenzen von Bindungen und Atomen, die auf dem kürzesten Pfad zwischen diesen beiden Atomen zu finden sind. Die Atomtypdefinition schließt hierbei die Anzahl der Bindungen zu „schweren“ Atomen (alle Nicht-Wasserstoffe) und die Anzahl der π -Elektronen mit ein: {Atomart, Anzahl Nachbarn, Anzahl π -Elektronen}. Ein Fragment zwischen dem einfach gebundenen Sauerstoff der Carboxyl-Gruppe und dem der Ester-Gruppe in der Acetylsalicylsäure würde wie folgt bezeichnet werden:



Damit erhält man topologische potenzielle Zweipunkt-Pharmakophore, die sehr gut interpretiert werden können. Durch die Kombination von mehreren oder gar mittels einzelner dieser potenziellen Pharmakophore kann oft die biologische Aktivität ganzer Molekülreihen erklärt werden. Diese Art der Kodierung des Atompaares-Deskriptors ist eine der essenziellen Grundlagen des im Rahmen dieser Arbeit neu entwickelten Deskriptors. Zusammen mit einer detaillierten Beschreibung der Vorläufertechniken wird diese Tatsache noch ausführlich erläutert.

2.1.3 Dreidimensionale Deskriptoren

Prinzipiell ist der gerade beschriebene Atompaares-Deskriptor eine hervorragende Methode zur Beschreibung von Molekülen. Trotzdem krankt er an einem entscheidenden Punkt: Die dreidimensionale Raumstruktur der untersuchten Moleküle wird völlig außer Acht gelassen. Der geometrische Aufbau der Moleküle wird ignoriert, obschon die Anzahl an Bindungen zwischen zwei Atomen gut mit deren räumlichem Abstand korreliert.

Aus diesem Grund gibt es seit etwa zwanzig Jahren Weiterentwicklungen der ein- und zweidimensionalen Moleküldeskriptoren zur dritten Dimension. Die Grundidee wurde schon deutlich früher geprägt [33]. Trotzdem erlaubten erst neuere mathematische Modellierungstechniken kombiniert mit der besser werdenden maschinellen Rechenleistung die Entwicklung und Nutzung von dreidimensionalen Moleküldeskriptoren. Im Hinblick auf die Beschreibung der neu entwickelten Technik werden die 3D-Deskriptoren wie folgt

gegliedert: Zunächst werden Techniken, die eine gleichsinnige Ausrichtung der Moleküle im Raum (sog. *Alignment*) benötigen vorgestellt. Danach werden die translations- und rotationsinvarianten (TRI-) Deskriptoren beschrieben. Diese erfordern eine solche gleichsinnige räumliche Ausrichtung nicht mehr.

Bereits an dieser Stelle sei angemerkt, dass auch die 3D-Deskriptoren auf einer Annahme beruhen, die das Ergebnis potenziell stark beeinflussen kann: Für jedes Molekül kann nur ein Konformer in der Analyse verwendet werden. Von diesem Konformer wird angenommen, dass es dem biologisch aktiven sehr ähnlich ist. Mögliche molekulare Flexibilität wird hierbei völlig außer Acht gelassen, was in Abbildung 2 zur Verdeutlichung gezeigt ist. Dieses Problem wurde erst im Rahmen der 4D-QSAR gelöst.

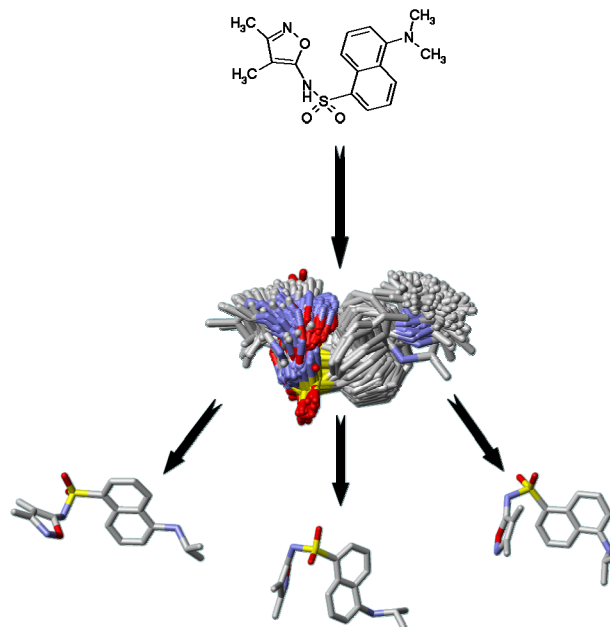


Abbildung 2: Das zentrale Problem aller 3D-Deskriptoren: Obwohl ein Molekül aufgrund seiner Flexibilität eine Reihe von verschiedenen Konformeren einnehmen kann, so wird doch immer nur genau eines für die Analyse ausgewählt.

2.1.3.1 Translations- und Rotationsvariante Deskriptoren

Diese Deskriptoren waren die ersten, die die dreidimensionale Molekülstruktur erfassen können. Die neueren Vertreter sind fast ausschließlich feldbasiert. Dies bedeutet, dass die Moleküle zuerst in einer Gitterbox platziert werden. Anschließend wird ihre Umgebung mit diversen Sonden abgetastet. Dies ist in Abbildung 3 gezeigt. Der bekannteste feldbasierte 3D-Deskriptor ist zweifellos die „Comparative Molecular Field Analysis“ (CoMFA) [34]. An Hand dieser Methode werden die Grundlagen der 3D-Deskriptoren detaillierter beschrieben. Die dahinter stehende Idee ist es, die Affinität eines Wirkstoffes zu seiner biologischen Zielstruktur durch potenzielle nicht-kovalente Wechselwirkungen zu beschreiben. Diese

Grundidee wurde auch in veränderter Form in dieser Arbeit verwendet. In der CoMFA werden die Interaktionskräfte durch die Beschreibung der sterischen und elektrostatischen Moleküleigenschaften kodiert. Um diese zu verarbeiten, benutzt CoMFA (sowie viele weitere Vertreter dieser Deskriptorklasse) ein Sondenatom, mit dem die räumliche Umgebung eines Moleküls innerhalb der Gitterbox abgetastet wird. Diese Gitterbox ist von ihrer Größe her so angelegt, dass sie eine etwas größere Ausdehnung als das größte Molekül des Datensatzes hat (ca. 4 Å). Als Auflösung wird normalerweise 1 oder 2 Å gewählt. An allen Gitterpunkten wird die Wechselwirkungsenergie des Moleküls mit dem Sondenatom errechnet. Bei CoMFA ist dies standardmäßig ein einfach positiv geladenes sp^3 -hybridisiertes Kohlenstoffatom. Die hierbei erhaltenen molekularen Interaktionsfelder (MIF) beschreiben, wie das Molekül an der entsprechenden Position mit dem Sondenatom in Wechselwirkung treten kann. Sterische Interaktionen werden mit dem Lennard-Jones-Potenzial (E_S), elektrostatische mit dem Coulomb-Potenzial (E_C) beschrieben. Die hierbei berechneten Interaktionsenergien werden in einer eindeutig definierten Reihenfolge ausgelesen und in einem Zeilenvektor für jedes einzelne Molekül abgespeichert. Dadurch entsteht eine Datenmatrix (sog. **X**-Matrix), die als Startpunkt für weitergehende mathematische Analysen benutzt wird. Durch diese Kodierung der molekularen Eigenschaften ist es möglich, bestimmte Bereiche mit stark positivem oder stark negativem Einfluss auf die biologische Aktivität mittels mathematischer Verfahren zu identifizieren. Diese Information kann bei der ausführlichen Interpretation der Unterschiede im Datensatz und zur Entwicklung neuer Moleküle benutzt werden. Zur Visualisierung werden die entsprechenden Produkte aus den Regressionskoeffizienten der QSAR-Gleichung und der Standardabweichung der Deskriptoren bei einzelnen Sondenpunkten dargestellt. Über farbliche Unterscheidungen in den entsprechenden Regionen können entscheidende chemische Gruppen identifiziert werden. Aufgrund der Überlagerung von Molekülstrukturen mit den entsprechenden Feldern können die Ergebnisse sehr gut interpretiert werden. Die bisher getroffenen Aussagen treffen im Wesentlichen auch für verwandte feldbasierte Techniken wie CoMSIA [35], GRID [36] und HASL [37] zu. Diese bieten Lösungen für einige Probleme der CoMFA. Zum Teil werden beispielsweise andere Methoden zur Potenzialberechnung eingesetzt. Die direkte Oberflächeninteraktion wird durch eine Kombination weiterer Potenziale beschrieben. Diese Unterschiede wurden bereits häufig in der Literatur diskutiert, daher wird auf entsprechende Referenzen verwiesen [19,38].

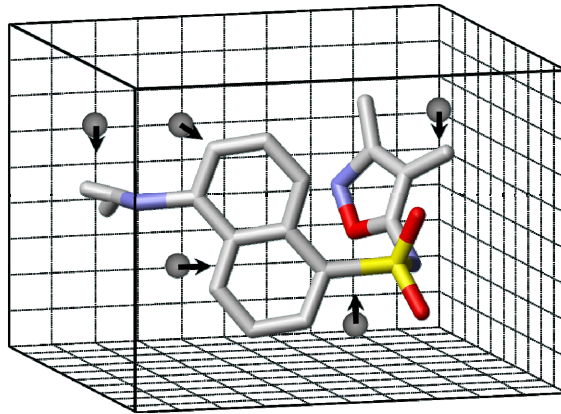


Abbildung 3: Schematische Darstellung der Grundlage aller feldbasierten, translations- und rotationsvarianten QSAR-Techniken. An jedem Gitterpunkt wird die Interaktionsenergie zwischen Sonde (die grauen Punkte repräsentieren Kohlenstoff-Atome) und Molekül mit Hilfe verschiedener Potentiale berechnet. Jegliche Änderung der Position in der Gitterbox führt zu Veränderungen des Deskriptors.

So gut die Interpretier- und Anwendbarkeit dieser Techniken ist, so vereinen sie auch ein schwerwiegendes Problem, das immer entscheidenden Einfluss auf das gefundene Ergebnis hat: Die QSAR vergleicht die unterschiedlichen zuvor errechneten Felder anhand der Raumpositionen der verschiedenen Interaktionsenergien. Dafür ist es nötig, dass die Moleküle vor der Berechnung der Felder eine Position in der Gitterbox einnehmen, die der im Rezeptor/Enzym gleicht. Dies kann aber nur gewährleistet werden, wenn die Moleküle im Raum so überlagert werden, dass diese Bedingung erfüllt ist. Dieser so genannte Alignment-Schritt hat massiven Einfluss auf das Ergebnis. Schon kleine Verschiebungen im Raum führen zu völlig anderen Deskriptoren. Zur Illustrierung dieses Einflusses dient Abbildung 4.

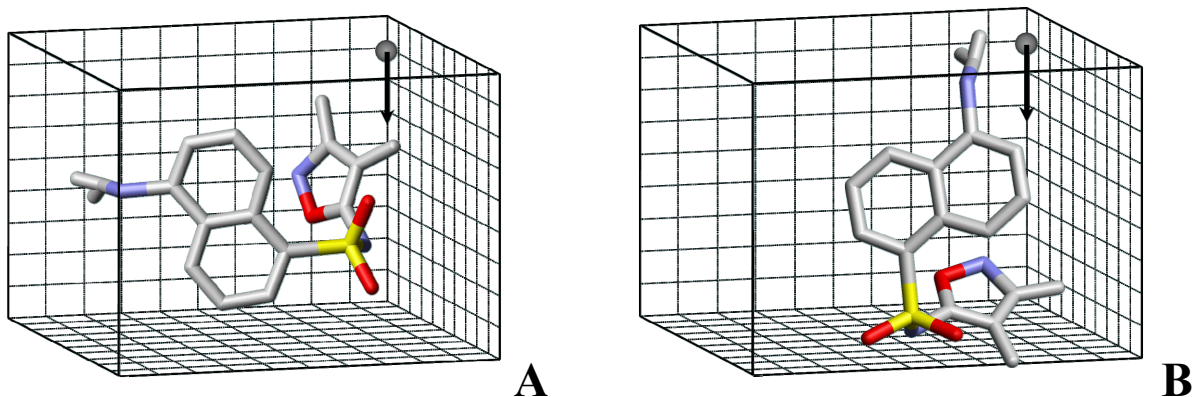


Abbildung 4: Das zentrale Problem aller translations- und rotationsvarianten, feldbasierten Techniken: Eine Änderung der Molekülposition im Raum führt zu völlig anderen Deskriptoren. Deswegen müssen, um eine Vergleichbarkeit gewährleisten zu können, alle Moleküle gleichsinnig ausgerichtet werden. Dies ist als **Alignment-Problem** bekannt geworden.

Bei falscher Überlagerung ist es unmöglich, anschließend eine vergleichende Analyse mit sinnvollem Ergebnis durchzuführen. Es kann keine chemisch sinnvolle Aussage getroffen werden. Betrachtet man das in Abbildung 4 gezeigte Sondenatom, würde im in Abb. 4A gezeigten Fall die Interaktion mit der Methylgruppe bestimmt. Wird das Molekül so wie in Abb. 4B rotiert würde dagegen eine Interaktion mit dem Aromaten gemessen, obwohl sich das Sondenatom an identischer Stelle befindet. Diese gemeinsame Ausrichtung (engl. *Alignment*) im Raum ist die entscheidende Grundlage für die Anwendung feldbasierter Techniken.

Dieser Schritt der Ausrichtung im Raum ist normalerweise extrem zeitaufwändig. Er kann außerdem durch die subjektive Einschätzung des jeweiligen Benutzers stark beeinflusst und das Ergebnis somit verzerrt werden. Es gibt eine Reihe von Ansätzen zur Automatisierung dieser Überlagerung [39-44]. Trotzdem ist der Alignment-Schritt immer noch der schwierigste und aufwändigste bei der Anwendung feldbasierter QSAR-Techniken.

Die genannten Ansätze sind bei heterogenen Datensätzen meistens von vornherein zum Scheitern verurteilt. Solche Datensätze setzen sich aus Molekülen mit unterschiedlichen Grundstrukturen zusammen oder aus Molekülen mit einer hohen Anzahl potenziell pharmakophorer Gruppen zusammen. In Folge dessen kann im Alignment kein sinnvolles Ergebnis gefunden werden.

Deshalb wurde in letzter Zeit ein starker Fokus auf die Entwicklung von Deskriptoren gelegt, die diese Überlagerung nicht mehr benötigen. Diese translations- und rotationsinvarianten (TRI-)Deskriptoren wurden in den letzten Jahren stark beforscht. Die in dieser Arbeit neu entwickelte Technik weist auch die Eigenschaft der Translations- und Rotationsinvarianz auf. Im Folgenden wird Wert auf eine detaillierte Beschreibung dieser Techniken gelegt. Gut dokumentierte TRI-Techniken gibt es derzeit nur auf dem Gebiet der 3D-QSAR.

2.1.3.2 Translations- und Rotationsinvariante Deskriptoren

Grundlagen

Der erste 3D-TRI-Deskriptor, der in der QSAR eingesetzt wurde, ist die „Molecular Transform“ [33]. Dabei wird eine vereinfachte theoretische Streufunktion aus der Elektronendiffraktion verwendet. Durch verschiedene Atomeigenschaften beschreibt der Deskriptor die jeweiligen Moleküleigenschaften. Das eigentliche Fundament für alle weiteren Entwicklungen jedoch stellt die Publikation von Moreau und Broto dar [45]: Erstmals verwendeten sie die Autokorrelationsfunktion im Bereich der Molekülbeschreibung:

$$ACF(l) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_A} \delta_{i,j} (P_i \cdot P_j)$$

Gleichung 1

mit $\delta_{i,j}=1$ wenn $d_l \leq r_{i,j} < d_{l+1}$ sonst $\delta_{i,j}=0$

$\delta_{i,j}$ (Kronecker's δ) ist hier gleich 1, wenn die Distanz zwischen den Atomen ($r_{i,j}$) im Bereich d_l und d_{l+1} liegt. Der Index l steht für die jeweilige Distanzkategorie, N_A ist die Gesamtatomzahl des Moleküls, P_i und P_j sind Eigenschaften des i ten und j ten Atoms. Bei der Autokorrelationsfunktion gibt es also für jede Distanzkategorie ein Element in einem Vektor \mathbf{c} . In jedem Element wird die Information über das Produkt der Eigenschaftswerte aller Atompaaire einer definierten Kategorie gesammelt. Ist eine Kombination doppelt vorhanden, so wird sie zu der bereits vorhandenen addiert. Der Vektor \mathbf{c} beschreibt folglich die Verteilung der Atompaaire im zu beschreibenden Molekül. Nachdem die Autokorrelation zunächst als topologischer zweidimensionaler Deskriptor genutzt wurde, kam es bald zum Einsatz Euklidischer Distanzen[46]. Damit wurde eine dreidimensionale Molekülbeschreibung erreicht. Die Einträge des Vektors \mathbf{c} sind in Standardanwendungen die Kategorien $\{0\text{-res}, 1\text{-res}, 2\text{-res}, \dots, d_{max}\}$, wobei d_{max} die maximale Distanz im gesamten Datensatz oder eine vom Benutzer vorgegebene maximale Distanz darstellt. Die Größe der Auflösung (res) dagegen hängt von der angewendeten Methode ab. Im Falle einer zweidimensionalen Beschreibung ist $res = 1$ [Einheit: topologische Bindung]. Diese ist wie folgt definiert: Jedes Atom auf dem kürzesten Pfad zwischen dem Start- und Zielatom stellt einen Knoten im mathematischen Graphen dar. Ausgehend vom Startatom zählt der Weg über eine Bindung zum nächsten Atom als eine topologische Bindung. Im nächsten Schritt wird dieses nächste Atom als Startatom verwendet. Dieser Vorgang wird wiederholt bis das Endatom erreicht ist und alle Bindungen gezählt wurden. Es wird die alternierende Sequenz aus Atomen und Bindungen beschrieben. Dagegen kann bei geometrischen Distanzen res eine vom Benutzer frei wählbare (auch nicht äquidistante, d.h. ungleich große) räumliche Distanz sein. Die Einheit räumlicher Distanzen ist [\AA]. Diese Grundidee wird in ähnlicher Form bei fast allen neueren TRI-Deskriptoren, so auch bei der hier neu entwickelten Technik, eingesetzt.

Im Folgenden wird häufig der Begriff „Radialverteilungsfunktion“ verwendet: Dieser beschreibt einen Vektor, der die Verteilung von Atomeigenschaften im Molekül relativ zu ihren Distanzen kodiert. In jedem Element des Vektors wird die Information über die Häufigkeit einer Eigenschafts-Eigenschafts-Kombination in einem definierten Distanzbereich gesammelt. Dabei werden sowohl Kombinationen gleicher Eigenschaften, die

Autokorrelationen, als auch Kombinationen verschiedener Eigenschaften, die Kreuzkorrelationen, berücksichtigt. Eine Radialverteilungsfunktion ist daher eine Kombination mehrerer Histogramme, die die Häufigkeit verschiedener Eigenschaftskombinationen im Bereich 0 bis d_{max} beschreiben. Für einen hypothetischen Fall, bei dem 3 verschiedene Eigenschaften A, B und C bis zu einer Maximaldistanz von 3 Å bei einer Auflösung von 1 Å kodiert werden sollen, würde der Vektor insgesamt 18 Elemente beinhalten. Dieser Vektor ist als Beispiel in Abbildung 5 gezeigt. Bei den vorgestellten Techniken ist das Prinzip genau das Gleiche. Es werden mehr Eigenschaften und größere Distanzen eingesetzt, der resultierende Vektor ist dementsprechend größer. Eine Radialverteilungsfunktion beschreibt ein Molekül unbeeinflusst von der Position im Raum. Das Ergebnis ist translations- und rotationsinvariant und somit alignment-unabhängig.

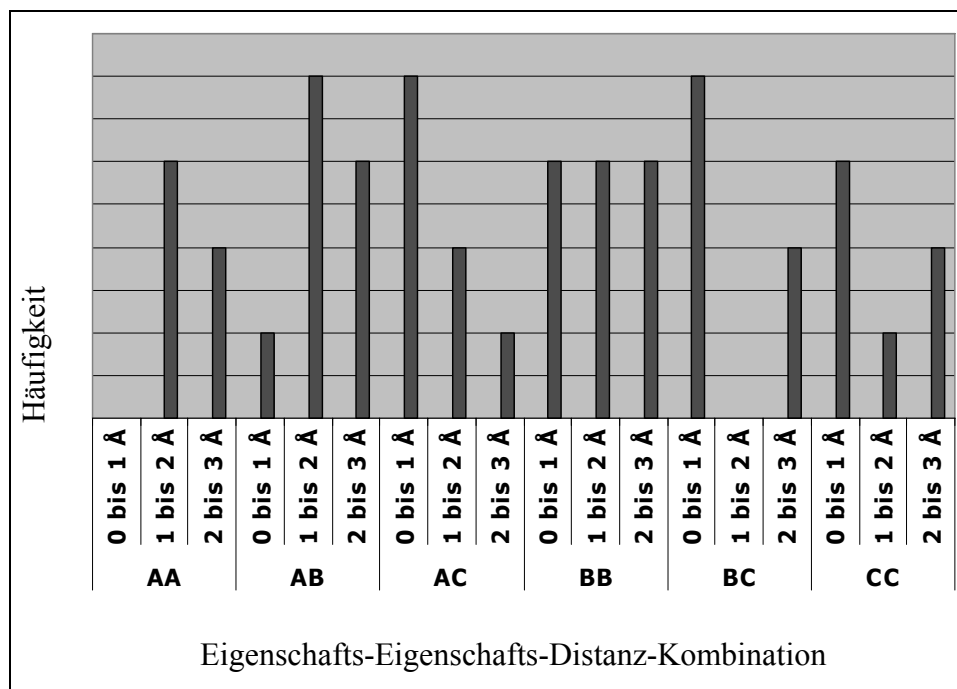


Abbildung 5: Eine Radialverteilungsfunktion, die die Häufigkeit der Kombinationen dreier Eigenschaften A, B und C bis zu einer Maximaldistanz von 3 Å beschreibt.

Es existieren zwei bekannte Deskriptorfamilien, die die Translations- und Rotationsinvarianz auf andere Art und Weise erreichen: Zunächst sind hier die „*Weighted Holistic Invariant Molecular Descriptors*“ (WHIM) [47] zu nennen. Diese platzieren die Moleküle über eine definierte Translation und Rotation im Koordinatensystem. Die Koordinaten der so positionierten Moleküle werden in eindeutige Kenngrößen umgewandelt, um als Deskriptoren eingesetzt zu werden. Verschiedene Erweiterungen der WHIM-Technik verwenden nicht mehr nur Atome, sondern auch Oberflächenpunkte [48,49] oder GRID-Felder [50]. Die dabei berechneten Deskriptoren sind jedoch nicht interpretierbar. Dies gilt analog für den EVA-

Deskriptor [51,52], bei dem berechnete Molekülspektren als Deskriptoren eingesetzt werden. Dies können IR- oder auch NMR-Spektren sein. Sowohl WHIM als auch EVA sind nicht mit der hier entwickelten Technik verwandt und werden deswegen nicht näher erläutert.

Der Start-End-Shortest-Path-Deskriptor (SESP)

Der SESP-Deskriptor ist der Erste aller im Arbeitskreis Baumann entstandenen Deskriptoren [53]. Das macht ihn zum ersten direkten Vorläufer der in dieser Arbeit neu entwickelten Technik. Deswegen soll der SESP-Deskriptor detailliert beschrieben werden.

Der Vorläufer von SESP ist der Start-End-Vektor (SE-Vektor) [54,55]. Die gemeinsame Basis ist eine topologische Beschreibung der Moleküle als mathematischer Graph. Unterschied zwischen SESP und dem SE-Vektor ist, dass SESP nur topologische Distanzen, also nur kürzeste Pfade zwischen zwei Atomen verwendet. Der SE-Vektor verarbeitet dagegen alle möglichen Pfade. Besonders bei zyklischen Molekülen wird der Unterschied deutlich, da bei azyklischen nur kürzeste Pfade existieren. Aufgrund dieser Vernachlässigung anderer Pfade ist eine Erweiterung zu einer geometrischen Variante möglich.

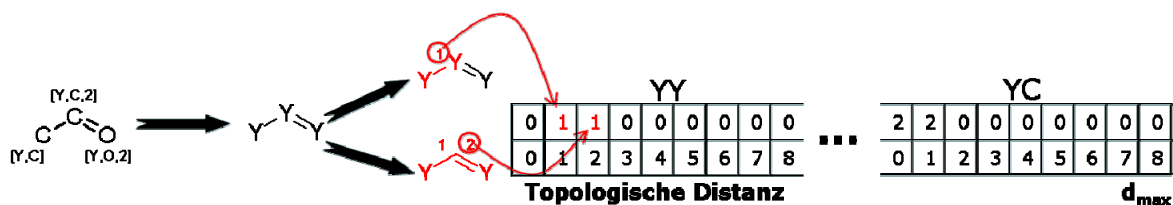


Abbildung 6: Die Berechnung des SESP-Deskriptors exemplarisch für den Atomtyp Y (Y=Heavy) am Beispiel Acetaldehyd.

Prinzipiell ist SESP eine Modifikation des vorher erwähnten Atompaares-Deskriptors [31]. Dabei kodiert SESP den Hybridisierungsgrad aller Atome im Datensatz direkt und beinhaltet darüber hinaus über einen generischen Atomtyp die Molekülgestalt. Damit wird die Molekülinformation deutlich kompakter kodiert.

Eine Übersicht über die Art der Berechnung zeigt Abbildung 6. Die Radialverteilungsfunktion, welche auch als p-Vektor bezeichnet wird, wird derartig aufgeteilt, dass für jede Eigenschafts-Eigenschafts-Distanz-Kombination (EED) genau ein Eintrag besteht. Diese selektive Distanz-Zählstatistik wird mit Nullen initialisiert. Eine EED beschreibt genau ein Element dieses Vektors.

Jedes einzelne Atom wird einmal als Startpunkt genommen und die EED zu allen anderen Atomen bestimmt. Kodiert man wie in Abbildung 6 das Acetaldehyd, so könnte beispielsweise der terminale Kohlenstoff (Typ [Y, C]) als erster Startpunkt genommen werden. Dabei beschreibt das „Y“ den generischen Atomtyp „Heavy“ (schweres Atom), den alle Nicht-Wasserstoffatome erhalten. Das „C“ steht für Kohlenstoff, also den chemischen Atomtyp. Es gibt genau zwei Pfade, die eine entsprechende EED beschreiben: Die EED von einem Kohlenstoff zum anderen (Typ [Y, C, 2]), bei der genau eine Bindung zwischen den Atomen liegt. Die zweite EED ist die vom ersten Kohlenstoff zum Sauerstoff (Typ [Y, O, 2]; O beschreibt den chemischen Atomtyp), hierbei liegen zwei Bindungen zwischen den Atomen. In beiden Fällen beschreibt die Zahl bei den Eigenschaften den Hybridisierungsgrad. Der Eintrag in den EEDs YY_1 und YY_2 würde jeweils um eins erhöht. Analog erfolgt das auch bei allen anderen Atomtypkombinationen, wie zum Beispiel YC oder CC . Zusätzlich gibt es noch EEDs die das Vorhandensein eines einzelnen Atoms kodieren, nämlich jeweils Distanz Null einer jeden EED, wie z.B. YY_0 oder YC_0 . Bei dieser Vorgehensweise wird für jede EED mit $d > 0$ der Eintrag doppelt, weshalb dieser am Schluss noch halbiert werden muss.

Im Endeffekt stellt also der SESP-Deskriptor eine Kombination distanzabhängiger Histogramme der topologischen Distanzmatrix dar, die in einer Radialverteilungsfunktion kombiniert werden. Dies wird analog bei den später entwickelten Methoden aus dem Arbeitskreis Baumann so durchgeführt. Diese mathematische Operation ist die Grundlage zur Erreichung der Invarianz gegenüber Translation und Rotation. Visualisiert man die entsprechende Radialverteilungsfunktion der Atom- und Bindungseigenschaften, so wird die Vorgehensweise der SESP-Kodierung schnell deutlich. Dies ist in Abbildung 7 veranschaulicht.

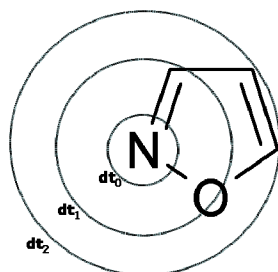


Abbildung 7: Die Kodierung des p-Vektors als Radialverteilungsfunktion relativ zum Stickstoff. Die topologische Distanz wird in dt_n ($n=0, \dots, d_{max}$) kodiert.

Der SESP-Deskriptor wurde zu einer geometrischen Variante erweitert. Dabei wird die geometrische mit der topologischen Distanz so gewichtet, dass auch Information über die dreidimensionale Struktur der Moleküle kodiert wird. So werden die Stereochemie sowie

unterschiedliche Konformationen der untersuchten Moleküle berücksichtigt. Der Deskriptor wird durch die Erweiterung sensitiv im Hinblick auf die Verwendung von verschiedenen Konformeren für ein Molekül. Bei der Berechnung wird die Euklidische Distanz zwischen zwei Atomen durch die Anzahl der dazwischen liegenden Bindungen dividiert. Dieser Schritt wird nur für EEDs durchgeführt, die Distanzen größer Null beschreiben. Hierbei werden konformell verschiedene Verbindungen wie Wanne/Sessel im Cyclohexan diskriminiert [56]. Das hat normalerweise keinen signifikanten Einfluss auf die Modellqualität, da topologische und Euklidische Distanz normalerweise hochkorreliert sind. Demzufolge ist SESP kein reiner 2D-Deskriptor, sondern muss eher als zweieinhalbdimensional betrachtet werden. Seine Leistungsfähigkeit hat er schon in vielerlei Anwendungen unter Beweis gestellt [56].

Auf GRID basierende 3D-TRI-Deskriptoren

In dieser Kategorie gibt es insgesamt drei verschiedene Deskriptorfamilien. Einerseits sind das die MOLPRINT-3D-Deskriptoren, die von Andreas Bender in der Gruppe von Robert Glen entwickelt wurden. Andererseits sind das VolSurf [11,12] und GRIND (GRid INdependent) [13], die aus der Gruppe von Gabriele Cruciani kommen und bei Manuel Pastor weiterentwickelt werden. Diese verschiedenen Techniken werden im Folgenden kurz erläutert, da sie als TRI-Deskriptoren sehr bedeutend sind. Der Bezug zu der im Rahmen dieser Arbeit entwickelten Technik ist dagegen gering.

Die MOLPRINT-3D-Technik [57] basiert auf einer 2D-Technik, den MOLPRINT-Deskriptoren [58,59]. Beide wurden bisher ausschließlich im Virtuellen Screening eingesetzt. Im Falle des 3D-Deskriptors wird zunächst mit dem MSMS-Algorithmus [60] die lösungsmittelzugängliche Oberfläche berechnet. Damit erhält man eine Beschreibung der molekularen Oberfläche durch Punkte. Diese Punkte sind nicht gleichverteilt. Die Auflösung wird je nach Anwendung von 0.5 bis 2 Å variiert. An diesem Punkt der Berechnung ist zu vermuten, dass das Ergebnis stark von der zufälligen Lage der Oberflächenpunkte beeinflusst werden kann. Das wird jedoch von den Autoren nicht explizit diskutiert. Anschließend wird mit verschiedenen Standard-GRID-Sonden [36,61] für jeden einzelnen Oberflächenpunkt eine Interaktionsenergie berechnet. Im nächsten Schritt der Berechnung der MOLPRINT-3D-Deskriptoren wird jeder einzelne Oberflächenpunkt als Zentrum mehrerer konzentrischer Kreise (normalerweise 2 oder 3) angenommen. Der Radius eines Kreises ist eine Distanzeinheit d_i . Eine Distanzeinheit entspricht genau dem Abstand zu den nächstliegenden Punkten um das Zentrum, ein Kreis beinhaltet alle nächsten Nachbarn des Zentrums. Für jeweils einen ausgewählten Bereich (z.B. $d_i=1$) wird die Interaktionsenergie in dieser Umgebung in einer Radialverteilungsfunktion abgespeichert.

Man erhält eine Radialverteilungsfunktion, die für jeden Oberflächenpunkt genau seine Umgebung mit dem Vorhandensein verschiedener Interaktionen beschreibt. Diese Umsetzung der Energien wird von den Autoren nur unvollständig beschrieben.

Die errechnete Radialverteilungsfunktion stellt den Ausgangspunkt für weitere Analysen im Bereich des Virtuellen Screenings dar. Leider können diese Deskriptoren nicht visualisiert werden, so dass eine Anwendung im Bereich der QSAR nicht vorteilhaft ist. Nichtsdestotrotz haben diese Deskriptoren unter Beweis gestellt, dass sie bei Ähnlichkeitssuchen sehr wertvoll sein können [57-59,62-65].

Ihr Haupteinsatzgebiet in der QSAR dagegen haben VolSurf [11,12] und GRIND [13]. Die Berechnung dieser Techniken zeigt Abbildung 8 und wird im Folgenden genauer erklärt.

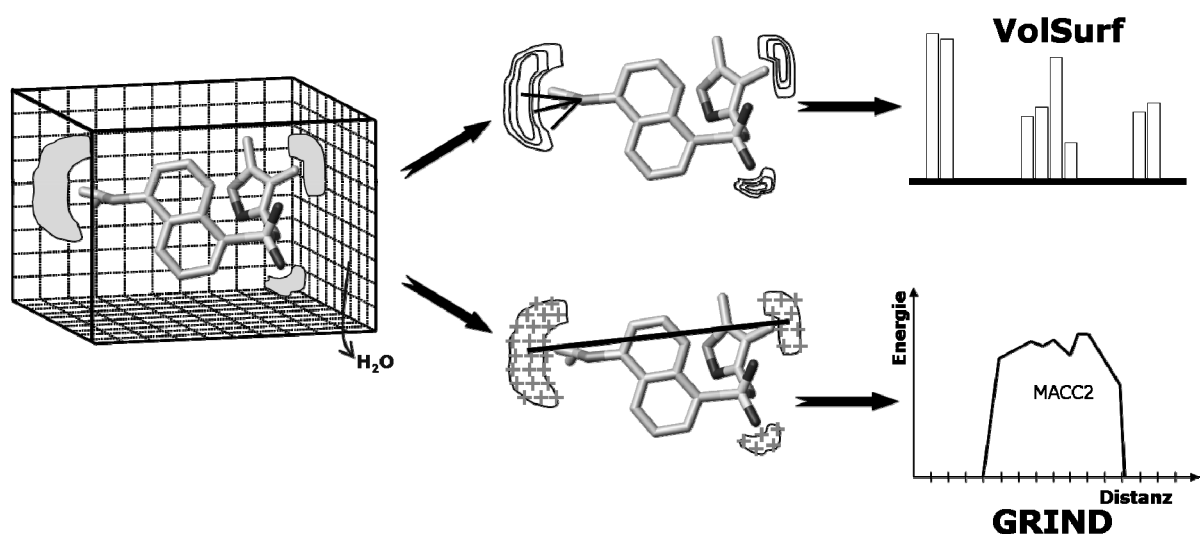


Abbildung 8: Die Berechnung der VolSurf- und GRIND-Deskriptoren. Nach Berechnung der Interaktionen mit dem GRID-Kraftfeld werden die Daten der molekularen Interaktionsfelder durch verschiedene mathematische Techniken in Deskriptoren umgewandelt (Darstellung verändert und ergänzt nach Fontaine[66,67]).

Die VolSurf-Deskriptoren dienen vorwiegend zur Beschreibung pharmakokinetischer Parameter. Sie werden hauptsächlich im Bereich der ADME-Vorhersage (Absorption, Distribution, Metabolismus, Exkretion) eingesetzt. Zunächst wird für jedes Molekül mit dem GRID-Kraftfeld ein molekulares Interaktionsfeld (MIF) berechnet [36,61]. Normalerweise werden hierfür die Wasser- und die hydrophobe (DRY-)Sonde benutzt. Alle anderen Sonden sind aber prinzipiell auch anwendbar [68]. Danach werden die Interaktionsenergien in mehrere Kategorien (normalerweise 8) unterteilt und über verschiedene mathematische Operationen in leicht interpretierbare Deskriptoren überführt. Diese Deskriptoren kodieren wichtige Moleküleigenschaften, wie beispielsweise die Fähigkeit zur Ausbildung von Wasserstoffbrückenbindungen und die Lipophilie des kodierten Moleküls. Zusätzlich wird

das Verhältnis der verschiedenen Eigenschaften im Deskriptor kodiert. Ein umfassender Überblick dazu findet sich bei Mannhold et al. [68]. Der große Vorteil dieser Deskriptoren ist, dass sie sehr gut rückprojiziert und interpretiert werden können. Dies geschieht mittels so genannter Koeffizientenplots, wobei sich vor allem große Interaktionsflächen einzelner Sonden gut darstellen lassen. VolSurf wurde erfolgreich in der Modellierung der Wasserlöslichkeit, von Oktanol-Wasser-Verteilungskoeffizienten, Verteilungsvolumina und metabolischer Stabilität eingesetzt [68-71].

Anhand der gleichen Datenbasis werden die GRIND-Deskriptoren berechnet [13]. Dies ist in Abbildung 8 illustriert. Das MIF wird hierbei bevorzugt mit den (Carbonyl-)Sauerstoff (O-), (Amid-)Stickstoff(N1-) und hydrophoben(DRY-) Sonden berechnet [13,72]. Dabei kommen sp^2 -hybridisierte O- und N-Atome als Stellvertreter für Wasserstoffbrücken-Donor- und -Akzeptor zum Einsatz. Das beruht auf der Annahme, dass so am Besten die Umgebung im Rezeptor beschrieben wird. Im nächsten Schritt werden die MIFs gefiltert, um die interessantesten Regionen zu identifizieren. Diese Filterung geschieht nach Einteilung der MIFs in negative Interaktionsenergien. Daraus werden durch einen von den Autoren entwickelten fedorov-artigen Optimierungsalgorithmus als wichtig angenommene Regionen extrahiert. Hierbei werden die einzelnen Regionen für die weiteren Berechnungen gewichtet, pro Kategorie wird nur der jeweils größte Eigenschaftswert gespeichert. Dieser Wert wird anschließend in den Bereich von 0 bis 1 skaliert. Sämtliche erwähnten Schritte sind von den Autoren nicht gut dokumentiert worden. Demzufolge kann die Beschreibung hier nicht umfassend sein. Die gefilterten MIFs, die die wichtigsten Regionen beschreiben werden in die GRIND-Deskriptoren überführt. Dafür wird eine Auto- und Kreuzkorrelationstransformation, die MACC-2-Transformation, genutzt. Hierfür wird zunächst das Produkt der Interaktionsenergien aller möglichen Eigenschaftspaare berechnet. Anhand der Distanzen zwischen den Knotenpunkten der einzelnen Bereiche wird das Produkt in definierten Kategorien abgespeichert. Ein Knotenpunkt ist dabei der Punkt im MIF mit der höchsten Interaktionsenergie. Jede Kategorie beschreibt einen kleinen Distanzbereich. Bei normalen Autokorrelationsanalysen wird die Information über verschiedene Bereiche mit gleichen Eigenschaften aufsummiert. Im Gegensatz dazu berücksichtigt GRIND nur das größte Produkt. Somit wird eine abschließende Interpretation der Daten möglich. Das geschieht über die Analyse von Korrelogrammen, die die Information über alle stärksten Produkte in sich tragen. Ausgehend von dieser Grundtechnik wurden verschiedene Erweiterungen publiziert. Diese versuchen beispielsweise, flexible Molekülbereiche besser zu modellieren [73], zusätzlich die Molekülgestalt mit zu berücksichtigen [66,67] oder einen festen Startanker für

die MACC-2-Transformation zu setzen [66,74]. Die letztgenannte Erweiterung gibt damit zum Teil die Translations- und Rotationsinvarianz wieder auf. Trotz der Vielzahl bekannter Probleme [66,72] ist GRIND die TRI-3D-Technik mit den meisten publizierten Anwendungen [72,75-79]. Dies liegt primär an der kommerziellen Verfügbarkeit. Das Potenzial für weitere Entwicklungen ist hier gegeben.

Distance Profile (DiP)

DiP ist die Erweiterung des SESP-Deskriptors hin zu einer reinen 3D-Technik, die direkt die geometrische Anordnung der einzelnen Atome im Molekül berücksichtigt [10,38]. Die grundsätzliche Berechnung ist in Abbildung 9 gezeigt. An Stelle der bei SESP verwendeten topologischen Distanz wird hier die Euklidische Distanz zwischen zwei Atomen bestimmt. Diese Information wird in einer Radialverteilungsfunktion abgespeichert, in der jedes Element eine definierte Atom-Eigenschafts-Distanz-Atom-Eigenschafts-Kombination beschreibt.

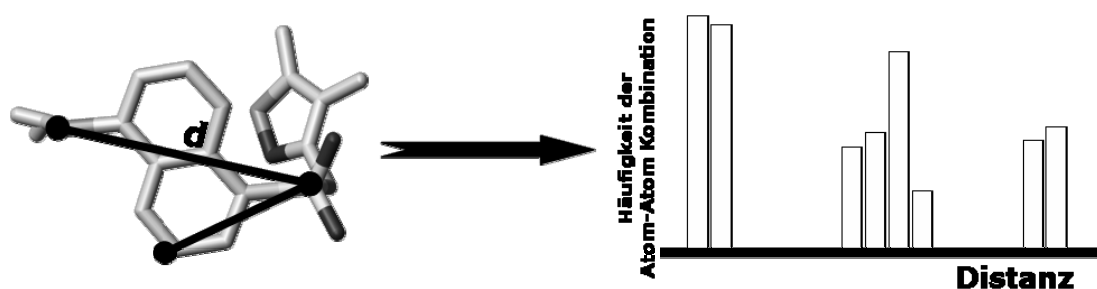


Abbildung 9: Die Berechnung des DiP-Deskriptors. Die Euklidischen Atom-Atom-Distanzen werden direkt in die Radialverteilungsfunktion umgesetzt.

Es werden die gleichen Atomtypen und –Eigenschaften wie bei SESP verwendet, jedoch sind im Vergleich dazu einige wichtige Ergänzungen notwendig. Erstmals müssen kontinuierliche Distanzdaten in Kategorien verteilt werden. Die maximale Atom-Atom-Distanz innerhalb der zu untersuchenden Moleküle bestimmt dabei die Dimension des Zählvektors, die Diskretisierung erfolgt in 1 Å-Schritten. Das bedeutet, dass alle Distanzen, die in einem Bereich um +/- 0.5 Å um ein Zentrum liegen, als identisch angesehen werden. Dies wurde auch bei der Nachfolgetechnik MaP und der im Rahmen dieser Arbeit entwickelten Technik so umgesetzt und wird dort noch umfangreicher beschrieben. Die Ergebnisse dieses Deskriptors sind sehr gut interpretierbar[10,38]. Die als wichtig identifizierten Atom-Atom-Distanzen können extrem einfach wieder in den ursprünglichen Datenraum zurückprojiziert werden.

Mapping Property Distributions of Molecular Surfaces (MaP)

Die MaP-Technik[14,19,38,80-82] (*Mapping Property Distributions of Molecular Surfaces*) ist der direkte Vorläufer der im Rahmen dieser Arbeit entwickelten neuen 4D-QSAR-Methode. Viele grundsätzliche Ideen wurden bei der Nachfolgetechnik verwendet und dort um eine vierte Dimension erweitert. Analog zum SESP-Deskriptor[53] und DiP-Deskriptor[10] ist die Bestimmung einer Radialverteilungsfunktion die Grundlage aller weiteren Berechnungen. Erstmals kommen eine unscharfe Zählweise sowie eine Berücksichtigung der molekularen Oberfläche zum Einsatz. Die Verwendung einer unscharfen Zählweise bedeutet, dass Information, die innerhalb einer Kategorie kodiert wird zu gewissen Teilen in weiteren Kategorien abgespeichert wird. Diese Verteilung erfolgt nach vorher definierten Regeln und wird später genauer beschrieben. Die gesamte Vorgehensweise ist schematisch in Abbildung 10 illustriert. Die einzelnen Schritte sollen aufgrund ihrer Bedeutung für die Entwicklung der neuen Technik im Folgenden detailliert aufgeführt und dargestellt werden.

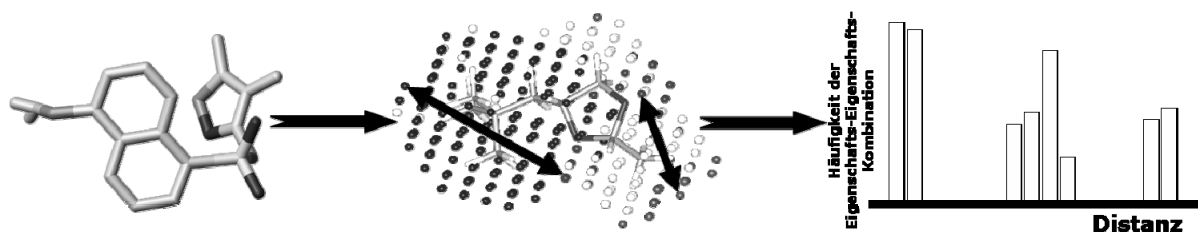


Abbildung 10: Überblick über die Berechnung des MaP-Deskriptors. Nach der Berechnung einer Oberfläche aus gleichverteilten Punkten und entsprechender Eigenschaftszuweisung werden Punkt-Eigenschafts-Punkt-Eigenschafts-Kombinationen in eine Radialverteilungsfunktion umgesetzt.

Wie bei jeder 3D-QSAR-Technik muss zunächst für jedes Molekül ein repräsentatives Konformer ausgewählt werden. Von diesem wird angenommen, dass es dem biologisch aktiven Konformer sehr ähnlich ist. Diese Einschränkung kann wie bei jeder 3D-QSAR-Technik zu einer Beeinflussung des Ergebnisses führen. Für MaP konnte gezeigt werden, dass dieser Einfluss kleiner ist als bei anderen Techniken [19]. Das ausgewählte Konformer wird bei der Berechnung des MaP-Deskriptors zunächst im Koordinatenraum entlang seiner Hauptachsen (-komponenten) ausgerichtet. Damit kommt es entlang seiner größten Ausdehnung auf der x-Achse zu liegen, entlang der zweitgrößten auf der y-Achse und entlang der drittgrößten an der z-Achse. Dieser so genannte Kanonisierungsschritt ist erforderlich, um den Einfluss des Diskretisierungsfehlers, der durch die Oberflächenberechnung entsteht, möglichst gering zu halten. Ist das ausgewählte Konformer entlang seiner Hauptachsen ausgerichtet, so wird es in einer Gitterbox mit einem Abstand zwischen den Gitterpunkten

von 0.8 Å platziert. Unter Zuhilfenahme eines modifizierten GEPOL-Algorithmus [14,83] wird eine analytische Moleküloberfläche angenähert, die aus gleichverteilten Punkten besteht. Im Anschluss wird jedem einzelnen Oberflächenpunkt auf Grundlage seines nächstliegenden Atoms eine Eigenschaft aus fünf verschiedenen Eigenschaftsklassen (Wasserstoffbrücken-Donor und -Akzeptor, Hydrophilie, starke und schwache Lipophilie) zugewiesen. Dies ist im Detail in den entsprechenden Veröffentlichungen beschrieben und wird in ähnlicher Art und Weise auch in dieser Arbeit eingesetzt. Auch MaP basiert auf der Umsetzung von Eigenschafts-Eigenschafts-Kombinationen in eine Radialverteilungsfunktion. Insgesamt gibt es, wenn man die erwähnten fünf Eigenschaftsklassen zugrunde legt, 15 mögliche Eigenschafts-Eigenschafts-Kombinationen. Zur Initialisierung der Radialverteilungsfunktion wird zunächst die maximale Euklidische Oberflächenpunkt-Oberflächenpunkt-Distanz d_{max} im zu untersuchenden Datensatz ermittelt und die nächstliegende Ganzzahl als maximale Distanzobergrenze festgelegt. Um den Deskriptor möglichst effizient zu gestalten ist es nötig, die errechneten Distanzen zu diskretisieren. Dadurch entstehen Distanzkategorien. Deshalb werden alle Distanzen, die innerhalb eines gewissen Fensters liegen als gleich angesehen. Es hat sich herausgestellt, dass hier ein Bereich von jeweils 1 Å sinnvoll ist, um in einem Histogramm die Distanzverteilung am Besten wiederzugeben. Dies wird außer bei MaP auch bei DiP und den Deskriptoren der CATS-Familie (siehe nächstes Kapitel) so gehandhabt. Als Mittelpunkte B einer Kategorie setzt MaP alle Ganzzahlwerte zwischen 1 und der zu d_{max} nächstliegenden Ganzzahl. Beispielsweise für ein $d_{max}=23.4$ Å ist dies die Zahl 23. In der jeweiligen Kategorie werden alle Werte abgespeichert, die im Bereich $d > B-0.5$ bis $d \leq B+0.5$ liegen. Der resultierende Vektor enthält für jedes Molekül $15 \times \text{int}(d_{max})$ Elemente und wird mit Nullen initialisiert.

Im nächsten Schritt wird ein zufällig ausgewählter Oberflächenpunkt als Startpunkt ausgesucht und eine erste Punkt-Punkt-Distanz berechnet. Die Information über diese Distanz wird anschließend an der Stelle im Vektor abgespeichert, die genau die entsprechende Eigenschafts-Eigenschafts-Distanzkombination beschreibt. Im Unterschied zu den Vorgängertechniken wird nicht mehr die vollständige Information in nur einer Kategorie abgespeichert. Auch die direkt benachbarte Kategorie wird mit inkrementiert. In welchem Maße das jeweils erfolgt wird mit Hilfe der exakten Euklidischen Punkt-Punkt-Distanz errechnet. Ist die Distanz zwischen zwei Punkten beispielsweise 4.66 Å, so erhält die Kategorie, welche eine Distanz von 5 Å beschreibt 66% des Inkrements von 1 (also 0.66). Der Rest (also 0.34) geht in die Kategorie, welche eine Distanz von 4 Å beschreibt. So wird bei allen möglichen Punkt-Punkt-Kombinationen verfahren. Am Ende ist eine

Radialverteilungsfunktion vorhanden, die alle Punkt-Eigenschafts-Punkt-Eigenschafts-Kombinationen beschreibt. Jedes einzelne Element des Vektors beschreibt wie bei den Vorgängertechniken SESP und DiP einen potenziellen Zweipunkt-Pharmakophor. Dieser kann ohne weiteres in den ursprünglichen Datenraum zurückprojiziert werden, womit das Ergebnis interpretierbar wird.

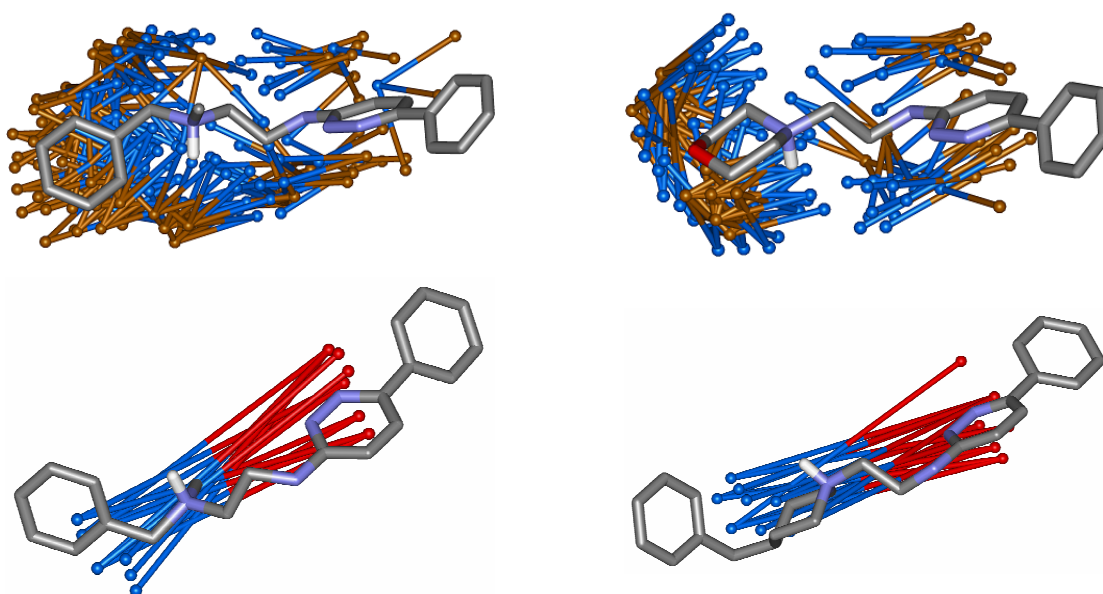


Abbildung 11: Rückprojektion der MaP-Deskriptoren in den ursprünglichen Datenraum. Hier ein Beispiel aus dem Bereich der Acetylcholinesterase-Inhibitoren. Die Richtungs- und Eigenschaftsinformationen erlauben Rückschlüsse auf pharmakophore Elemente der Moleküle.

Bei der Erstellung eines QSAR-Modells wird beispielsweise durch eine Variablenselektion einer der potenziellen Pharmakophore als wichtig identifiziert. Zur Darstellung werden alle Punkt-Punkt-Kombinationen verwendet, die die entsprechende Variable mit mehr als 0.8 inkrementieren. Zusammen mit der Verbindung zwischen den Punkten werden sie farbkodiert dargestellt. Damit enthält die Darstellung auch eine Richtungsinformation. Dieser Darstellung wird die Struktur des zu untersuchenden Moleküls überlagert. Diese Form der Rückprojektion ist in Abbildung 11 gezeigt. Auf diesen Bildern kann im Kontext des Gesamtdatensatzes bei der Interpretierung der QSAR-Gleichung sehr schnell erkannt werden, welche chemischen Gruppen des Moleküls für wichtig hinsichtlich der biologischen Aktivität befunden wurden. Durch eine Rückprojektion einzelner Variablen mehrerer Strukturen können Unterschiede gut erkannt werden und somit QSAR-Daten interpretiert werden. Dies konnte mehrfach gezeigt werden [14,19,38,80-82].

Die CATS-Familie

Die Familie der translations- und rotationsinvarianten CATS-Deskriptoren (Chemically Advanced Template Search) des Arbeitskreises Schneider von der Universität Frankfurt ist sehr eng mit den Deskriptoren des AK Baumann verwandt: SESP hat als Pendant CATS [84], DiP CATS3D [85,86] und MaP SURFCATS [85,86]. Daher werden die entsprechenden Techniken im Folgenden nur kurz erläutert und die wichtigsten Unterschiede aufgezeigt.

Die wichtigste Unterscheidung ist das Hauptanwendungsgebiet: Die CATS-Deskriptoren werden vorwiegend im virtuellen Screening eingesetzt, der Fokus der Baumann-Deskriptoren dagegen liegt in der Anwendung auf Quantitative Struktur-Wirkungs-Beziehungen. Die wesentlichen Unterschiede sind immer in der Kodierung der einzelnen Eigenschaften zu finden. CATS weist den einzelnen Atomen definierte pharmakophore Eigenschaften zu, wobei insgesamt 5 Klassen zum Einsatz kommen: Wasserstoffbrücken-Donor und -Akzeptor, positive und negative Ladung sowie Lipophilie, wobei jedem Atom zwischen 0 und 2 Eigenschaften zugewiesen werden. Auch dieser Deskriptor ist eng mit dem Atumpaare-Deskriptor verwandt: Es wird zunächst eine Pharmakophormatrix erstellt. Dabei enthält jeweils ein Eintrag genau eine Eigenschafts-Eigenschafts-Kombination. Insgesamt sind hier 15 Kombinationen möglich. Parallel wird eine Distanzmatrix erstellt, die über einen „Kürzesten-Pfad“-Algorithmus die topologischen Distanzen zwischen den einzelnen Atomen beschreibt. Dieser „Kürzeste-Pfad“-Algorithmus wurde bei verschiedenen Anwendungen von den Autoren variiert. Dies dürfte aber keinen Einfluss auf das Ergebnis haben. Jeder Eintrag der Distanzmatrix wird mit dem passenden aus der Pharmakophormatrix kombiniert, wobei alle Distanzen zwischen 0 und 9 Å berücksichtigt werden. Am Ende entsteht so ein 150-dimensionaler Deskriptor (15 Eigenschafts-Eigenschafts-Kombinationen multipliziert mit 10 Distanzkategorien) für das Molekül. Schließlich werden die Deskriptoren skaliert, wofür drei verschiedene Ansätze vorliegen: Keine Skalierung, Division durch die Anzahl der schweren Atome und Division der 15 verschiedenen potenziellen Zweipunkt-Pharmakophore durch die addierte Häufigkeit der zwei zugehörigen pharmakophoren Eigenschaften. Letzter Punkt bedeutet, dass gezählt wird, wie oft jede Eigenschaft vorliegt. Für einen potenziellen Zweipunkt-Pharmakophor wird die Häufigkeit der beiden Eigenschaften addiert und der entsprechende Deskriptor durch diese Häufigkeit dividiert. Diese letzte Methode berücksichtigt dabei insbesondere, dass seltene Pharmakophor-Kombinationen besonders wichtig in der Interaktion sein können und deswegen stärker gewichtet werden.

Diese Art der Molekülkodierung kann sehr einfach auf drei Dimensionen erweitert werden, was mit CATS3D auch umgesetzt wurde: Wie bei DiP werden hier die Euklidischen Distanzen zwischen den einzelnen Atomen errechnet und in einer Radialverteilungsfunktion,

die die einzelnen EEDs beschreibt, abgespeichert. Dabei wird eine diskretisierte Darstellung erreicht. Bei CATS3D kommen immer 20 EEDs mit einem Abstand von jeweils 1 Å zum Einsatz, womit ein Bereich von 0 bis 20 Å abgedeckt wird. Die zugewiesenen Atomtypen basieren auf der PATTY-Klassifizierung [87], was insgesamt 6 Typen beinhaltet: Kation, Anion, Wasserstoffbrücken-Donor, -Akzeptor, Polarität (Donor und Akzeptor) und Hydrophobie. Daraus ergeben sich 21 Eigenschafts-Eigenschafts-Kombinationen, insgesamt also ein 420-dimensionaler Deskriptor. Im Gegensatz zu CATS wird hier nur ein Pharmakophor-Typ pro Atom zugewiesen. Jedes Element der Radialverteilungsfunktion wird dabei nach folgender Gleichung berechnet:

$$CV_d^T = \frac{1}{N_1 + N_2} \sum_i \sum_j \frac{1}{2} \delta_{ij,d}^T \quad \text{Gleichung 2}$$

Hierbei sind i und j die Atomindices, d der Distanzbereich, T das Paar der zugewiesenen Atomtypen für die Atome i und j . N_1 und N_2 sind die Anzahl der Atome, die vom Typ i und j sind und δ_d^T (Kronecker's δ)=1 für alle Atompaare vom Typ T innerhalb des Distanzbereiches d . Der Faktor 0.5 sorgt dafür, dass automatisch doppelt gezählte Paare trotzdem nur einfach gewichtet werden.

Eine Kodierung der molekularen Oberfläche zur direkten Beschreibung der Interaktionen ist wünschenswert, da damit genau der für die Ligand-Rezeptor-Interaktionen relevante Bereich beschrieben wird: Auch die CATS-Familie hat ein Mitglied, das die molekulare Oberfläche kodiert: In Analogie zu MaP [14] wird bei SURFCATS eine Kontaktoberfläche für das Molekül berechnet. Das geschieht hier mit einem Gauss-Connolly-Algorithmus, der eine Oberfläche mit einem Punktabstand von circa 2 Å berechnet. Im Gegensatz zu MaP sind hier die Oberflächenpunkte jedoch nicht gleichverteilt, die Distanzen zwischen nächstliegenden Punkten variieren. Auch erfolgt keine Kanonisierung der Moleküle, so dass ein massiver Diskretisierungsfehler zu erwarten ist, der von den Autoren aber bisher nicht weiter untersucht wurde. Jedem einzelnen Oberflächenpunkt wird die Eigenschaft des nächstliegenden Atoms zugewiesen. Mit Hilfe der Euklidischen Distanzen zwischen jeweils zwei Oberflächenpunkten wird wie bei den beiden Vorgängertechniken eine Radialverteilungsfunktion berechnet, die die Verteilung der Punkt-Eigenschafts-Punkt-Eigenschaftskombinationen beschreibt. Auch hierzu wird die in Gleichung 1 beschriebene Autokorrelationsfunktion eingesetzt.

2.1.4 Die vierte Dimension in der QSAR

Wie bereits im vorherigen Kapitel angesprochen wurde, muss bei 3D-QSAR-Techniken immer ein Konformer ausgewählt werden. Von diesem wird angenommen, dass es dem bioaktiven Konformer möglichst nahe kommt. Im folgenden Kapitel werden verschiedene Techniken vorgestellt, die diese Beschränkung nicht mehr aufweisen. Diese Techniken können die Flexibilität von Molekülen mit einbeziehen. Trotzdem sind diese Methoden in verschiedener Weise eingeschränkt im Bezug auf den Alignment-Schritt oder ihre Interpretierbarkeit, was in den einzelnen Unterkapiteln diskutiert wird.

2.1.4.1 Die Methode von Hopfinger

Es sind fast zehn Jahre vergangen, seitdem von Hopfinger und Mitarbeitern die erste 4D-QSAR Technik vorgestellt wurde [15]. Die Technik wurde erfolgreich auf eine Vielzahl von Datensätzen angewendet und ausführlich validiert, was in einer Vielzahl von Publikationen dokumentiert ist [15,88-105].

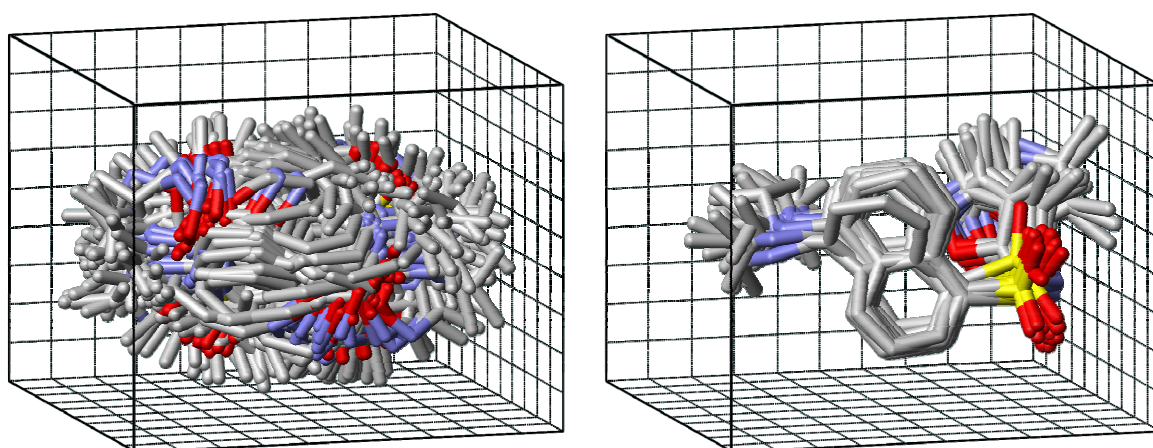


Abbildung 12: Illustration des Alignment-Problems bei Hopfinger's 4D-QSAR[15]. Neben allen Konformen eines einzelnen Moleküls müssen auch alle Moleküle sinnvoll übereinander gelegt werden. Diese Überlagerung ist hier für das Konformerensemble eines Moleküls gezeigt und illustriert die Flexibilität. Der Aufwand der Suche nach einem Alignment solcher Konformerensembles wird deutlich.

Die Grundidee dehnt im Prinzip die 3D-QSAR dahin gehend aus, dass jedes untersuchte Molekül nicht nur durch eines, sondern durch viele Konformere repräsentiert wird. So wird die Flexibilität in der Modellbildung berücksichtigt. Im ersten Schritt wird ein Konformerensemble für jedes Molekül durch Zuhilfenahme einer Moleküldynamik-Simulation (MD-Simulation) mit dem MM2-Kraftfeld berechnet [106]. Der Ablauf ist normalerweise wie folgt [105]: Es wird pro Molekül ein Zeitraum von 50 Picosekunden bei einer Temperatur von 300 Kelvin simuliert. Alle 0.001 Picosekunden wird eine Konformation abgespeichert. Folglich sind am Schluss der MD-Simulation für jedes Molekül 50000

Konformere abgespeichert worden. Konformere, die über 2 kcal/mol mehr innere Energie zeigen als das energieärmste Konformer werden von der Analyse ausgeschlossen. Damit werden sehr homogene Konformerensembles generiert, was aufgrund des darauf folgenden Alignmentschritts unbedingt erforderlich ist. Ein Beispiel zur Illustrierung des Alignmentschritts zeigt Abbildung 12 anhand des Konformerensembles eines einzelnen Moleküls. Der Alignment-Schritt erfolgt bei Hopfinger halb-automatisch [15,95,100], sein Einfluss wurde in einer Publikation ausführlich aufgezeigt [92]: Es werden drei Atome, die über alle Moleküle eines Datensatzes charakteristisch sind, ausgewählt. Anschließend werden diese Atome automatisch überlagert, so dass eine gleichsinnige Ausrichtung der Moleküle erreicht wird. Bei jeder QSAR-Analyse werden alle in Frage kommenden Alignments auf ihr Ergebnis hin überprüft. Das beste Ergebnis (also das mit dem höchsten q^2 bzw. R_{CV-1}^2 , Definitionen siehe unter Mathematische Modellierung) wird anhand der statistischen Modellqualität ausgewählt. Dieser Schritt erfordert Moleküle, die eine homogene Grundstruktur oder zumindest sehr leicht identifizierbare Bioisostere besitzen. Die Modellbildung erfolgt nach folgender Methode: Die 4D-QSAR ist eine gitterbasierte Technik. Daher werden alle im Raum überlagerten Konformeren gleichsinnig in einer Gitterbox mit passender Größe ausgerichtet. Ein Beispiel für eine Gitterbox ist in Abbildung 13 gezeigt.

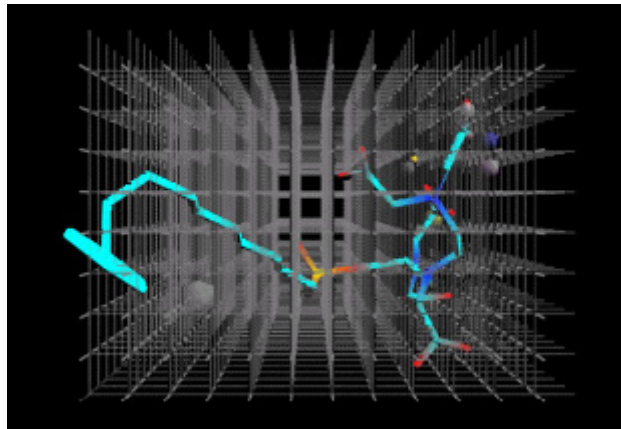


Abbildung 13: Beispiel für eine von Hopfinger verwendete Gitterbox. Hier gezeigt mit einem Konformer eines PGF_{2α}-Derivats. Darstellung entnommen aus [107].

Diese Gitterbox hat einen Punktabstand von 1 bis 2 Å. Jeweils acht dieser Punkte definieren einen Würfel mit einem Volumen von 1 oder 8 Å³. Ein Würfel wird als Zelle bezeichnet. Diese Zellen sind die Grundlage der Deskriptorberechnung. Der eigentliche Deskriptor basiert dann auf einer gewissen Anzahl von so genannten „Grid Cell Occupancy Descriptors“ (GCODs). Dieser Begriff kann am Besten mit „Häufigkeit der Besetzung von Zellen“ übersetzt werden. Diese englischsprachige Bezeichnung wird im Folgenden durchweg verwendet. Ein GCOD, der beispielsweise die Fähigkeiten zur Wasserstoffbrückenbindung

der Moleküle beschreibt, wird wie folgt berechnet: Für jede Zelle (siehe Abbildung 13) wird die Häufigkeit von dort liegenden Atomen innerhalb der Konformere eines einzelnen Moleküls, die als Wasserstoffbrücken-Akzeptoren fungieren, bestimmt. Für alle anderen Eigenschaften werden die GCODs genau so berechnet. Anschließend werden mit einer bestimmten mathematischen Modellierung die Teile der Gitterbox identifiziert, welche die wichtigsten Atome für das zu untersuchende Problem enthalten. Diese Daten können auch ohne weiteres in den ursprünglichen Datenraum rückprojiziert werden. Eine Interpretation der Daten ist damit relativ gut möglich. Mit dieser Visualisierung wird eine Identifizierung wichtiger Molekülelemente möglich.

2.1.4.2 Die Methode von Dobler und Vedani

Einen sehr interessanten Ansatz bietet die von Dobler und Vedani entwickelte mehrdimensionale QSAR-Methode [17,18]. Diese Technik stellt eine Erweiterung der Rezeptoroberflächenmodelle (engl. *Receptor surface models*) dar [108,109], durch die von QSAR-Daten Rückschlüsse auf die Struktur des Rezeptors gezogen werden [110]. Ein Beispiel für ein derartiges Strukturmodell ist in Abbildung 14 gezeigt.

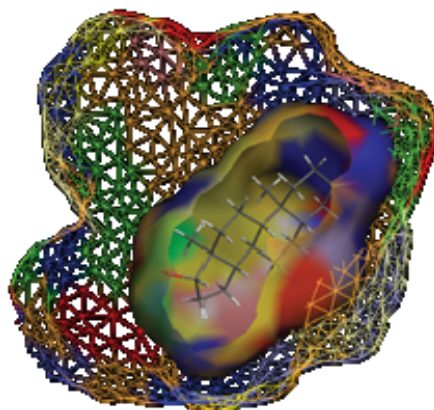


Abbildung 14: Beispiel für ein mittels der Methode von Dobler und Vedani erstelltes Modell des androgenen Rezeptors. Darstellung modifiziert nach [111].

Die Technik benötigt daher für eine sinnvolle Anwendung ebenfalls eine gleichsinnige Ausrichtung der Moleküle im Raum. Im Gegensatz zu den sehr vielen Konformeren, die bei Hopfinger's Methode verwendet werden, sind es hier maximal 64 pro Molekül. Gewöhnlich werden sogar weniger als 10 verwendet [112]. Die Technik ist für vielen Gebiete der Wirkstoffforschung anwendbar [17,18,113-117] und wurde in letzter Zeit auch in eine fünfte [115-117] und eine sechste [118] Dimension erweitert. Damit werden zusätzlich die Flexibilität des Zielproteins und Löslichkeitszustände berücksichtigt. Obwohl es eine 4D-Technik ist, muss neben dem Einfluss des Alignment-Schrittes auch bei der Auswahl der zu untersuchenden Konformere ein starker Nutzereinfluss vermutet werden. Das liegt in der

relativ geringen Anzahl von Konformeren begründet. Die Ergebnisse sind gut interpretierbar, da die Modelldaten gut mit Strukturinformationen im ursprünglichen Datenraum visualisiert werden können (siehe Abbildung 14).

2.1.4.3 Weitere Ansätze

Im Bereich der 4D-QSAR gibt es neben den beiden angesprochenen Methoden noch weitere Ansätze, die die Molekülflexibilität mit in die Modellbildung mit einbeziehen. Diese Ansätze sind nicht gut dokumentiert. An erster Stelle ist dabei die von Kuz'min und Mitarbeitern [16,119] entwickelte Methode zu nennen. Diese generiert über eine MD-Simulation in Hyperchem die zu untersuchenden Konformerensembles. Dabei wird für die innere Energie der Konformere eine Obergrenze von 5 bis 7 kcal/mol über dem energieärmsten Konformer festgelegt. Konformere, die über der Grenze liegen in der Analyse nicht berücksichtigt.

Außerdem ist die Methode von Potemkin zu erwähnen, bei der die Autoren einen eigenen Algorithmus zur Konformationssuche verwenden [120,121]. Hier wird ebenfalls eine Obergrenze von 5 kcal/mol angewendet. Ein weiterer Ansatz wurde von Bhonsle und Mitarbeiter publiziert [122]. Hierbei wird ein mit Monte-Carlo-Berechnungen erstelltes Konformerensembles als Grundlage für die Optimierung eines Alignments und damit der statistischen Güteparameter einer CoMFA-Analyse verwendet. Diese Technik ist daher im Grenzgebiet zwischen dritter und vierter Dimension anzusiedeln.

2.1.5 Fazit zu bisherigen Deskriptorentwicklungen

Fasst man die bisher vorgestellten Fakten zusammen, so sind für die Deskriptoren bzw. die QSAR drei zentrale Schwierigkeiten festzuhalten:

- (a) Das Alignment-Problem
- (b) Die Auswahl eines Konformers, von dem angenommen wird, dass es dem biologisch aktiven sehr ähnlich ist
- (c) Die Interpretierbarkeit der Ergebnisse

Die bis hierhin vorgestellten Techniken bieten jeweils Lösungen für diese Probleme. Jedoch wurde bisher noch nie der Versuch unternommen, diese Schwierigkeiten gleichzeitig mit einer einzigen Technik zu beheben. Die im Rahmen dieser Arbeit entwickelte Methode soll eine mögliche Lösung für die drei genannten Probleme anbieten. Der dazu nötige mathematische Unterbau wird im folgenden Kapitel ausführlich vorgestellt.

2.2. Mathematische Modellierung

Ziel der QSAR ist es, eine mathematische Verknüpfung zwischen der chemischen Konstitution C und der biologischen Aktivität Φ zu finden. Die Deskriptoren beschreiben dabei die Konstitution C . Die Information darüber wird in einem Datenvektor \mathbf{x} gesammelt. Im Standardfall handelt es sich hierbei nicht um einen Vektor, sondern um eine Datenmatrix \mathbf{X} . Diese enthält die unabhängigen Variablen. Die biologische Aktivität Φ wird im Zahlenvektor \mathbf{y} abgelegt. Dies ist die abhängige Variable. Diese Datenstruktur kann gut mittels der in der Chemometrik gebräuchlichen Vektor- und Matrixnotation dargestellt werden. Eine ausführliche Dokumentation der hier benutzten Methoden findet sich in verschiedenen mathematischen Standardwerken [123-126].

2.2.1 Verwendete Notation und grundlegende Operationen

Im Folgenden wird zwischen Skalaren, Vektoren und Matrizen unterschieden. Um die jeweilige Zuordnung zu verdeutlichen werden Skalare *kursiv* (x), Vektoren klein und **fett** (\mathbf{x}) sowie Matrizen GROSS und **FETT** (\mathbf{X}) dargestellt. Einzelne Vektoren werden ausschließlich als Spaltenvektoren angegeben. Ein hochgestellter Index T zeigt an, dass es sich um die jeweilige transponierte Form handelt. Eine transponierte Matrix wird zum Beispiel mit \mathbf{X}^T beschrieben. Zur Beschreibung der Dimension von Vektoren und Matrizen wird $n \times p$ verwendet. Dabei ist n die Anzahl der Zeilen und p die Anzahl der Spalten. Die i -te Spalte der Matrix \mathbf{X} ist der Vektor \mathbf{x}_i . Bei den einzelnen Elementen von Vektoren und Matrizen handelt es sich um Skalare. Diese werden auch *kursiv* dargestellt. Sie erhalten zusätzlich einen bzw. zwei tiefer gestellte Indizes. Für eine Beispielmatrix \mathbf{X} mit $n \times p$ Elementen sieht das wie folgt aus:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

Außerdem haben einige häufig verwendete Matrizen mit speziellen Eigenschaften eigene Symbole. So wird die Einheitsmatrix (Matrix der Größe $m \times m$, Diagonalelemente Einsen, alles anderen Nullen) als \mathbf{I}_m und ein nur aus Einsen bestehender Spaltenvektor als $\mathbf{1}_m$ ($m \times 1$) bezeichnet.

Eine durchgängig verwendete Methode in der QSAR und daher auch in dieser Arbeit ist die Kreuzvalidierung. Einige Grundlagen zu Bezeichnungen werden in Abbildung 15 illustriert.

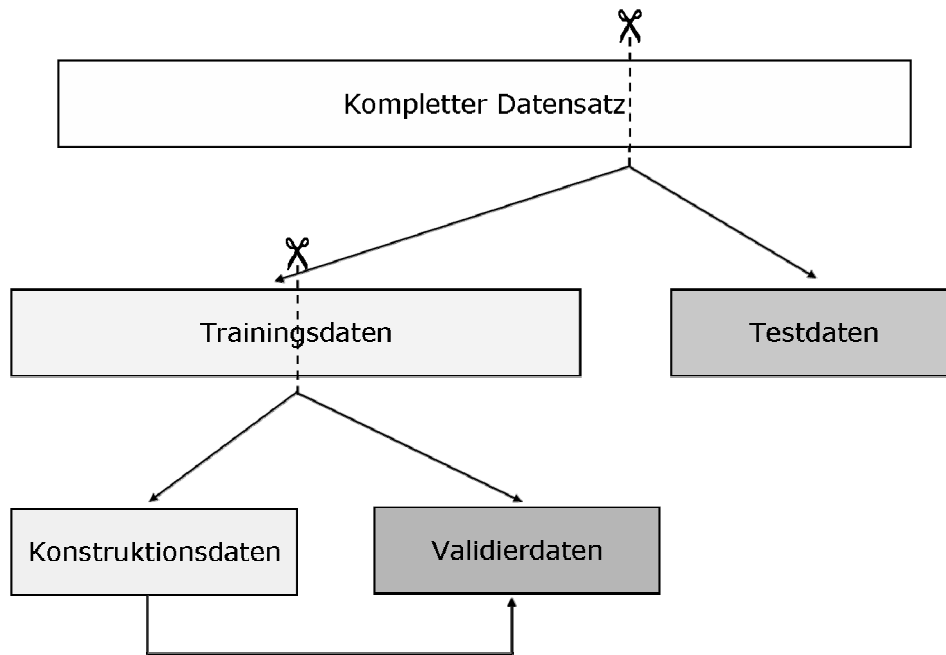


Abbildung 15: Schematische Datensatzaufteilung in der Kreuzvalidierung, Basis der Berechnung von statistischen Güteparametern.

Im Allgemeinen wird bei dieser Technik der aus n_{gesamt} Objekten bestehende Datensatz so aufgeteilt, dass zuerst ein Trainingsdatensatz ($n := n_{gesamt} - n_{Test}$ Objekte) und ein Testdatensatz (n_{Test} Objekte) erzeugt werden. Innerhalb des Trainingsdatensatzes wird dann ein Konstruktionsdatensatz mit $n-d$ Objekten und ein Validierdatensatz mit d Objekten gebildet. Auf Grundlage des Konstruktionsdatensatzes wird ein Modell berechnet, welches mit dem Validierdatensatz validiert wird (illustriert in Abb. 15) illustriert. Mit diesen Resultaten lassen sich interne statistische Güteparameter für ein Modell berechnen. Das in diesem Schritt gefundene beste Modell verwendet man, um den Testdatensatz vorherzusagen. Auf Grundlage dieser Testdatensatzvorhersage werden die externen statistischen Güteparameter berechnet. Grundsätzlich basiert die im Rahmen dieser Arbeit verwendete Validierung auf dieser Datensatzunterteilung. Die einzelnen Schritte werden jedoch in mehreren Iterationen wiederholt. Dies ist bei den Validierprotokollen im Detail gezeigt.

2.2.2 Wichtige Schritte zur Datenvorbehandlung

Die Schritte der Matrixfaktorisierung, die im Zuge der Berechnung von PCR und PLS angewendet werden sind nicht invariant gegenüber Zentrierung. Wenn nicht zentriert wird, wird Varianz komplett in den Daten belassen, obwohl sie in der Regel nicht bedeutsam ist. Deshalb werden die zu untersuchenden Daten zentriert.

2.2.2.1 Zentrierung

Die Datenmatrix \mathbf{X} enthält die jeweils p Deskriptorwerte für n Moleküle. Für jeden Deskriptorvektor \mathbf{x}_i wird der Mittelwert über alle n Moleküle berechnet, so dass sich für die ganze Matrix ein Mittelvektor $\bar{\mathbf{x}}$ der Dimension $p \times 1$ ergibt. Dieser wird dann so von der Rohdatenmatrix subtrahiert, dass jedes Element der Matrix um genau den betreffenden Mittelwert der Spalte vermindert wird:

$$\mathbf{X}_{\text{zentr}} = \mathbf{X}_{\text{orig}} - \mathbf{1}_n \cdot \bar{\mathbf{x}}^T \quad \text{Gleichung 3}$$

Identisch werden die Y-Daten, also die biologischen Aktivitäten, zentriert:

$$y_{\text{zentr}} = y_{\text{orig}} - \mathbf{1}_n \cdot \bar{y} \quad \text{Gleichung 4}$$

Hierbei ist \bar{y} der Mittelwert der abhängigen Variablen y . In dieser Arbeit sind alle vorgestellten Daten zentriert, sofern nicht explizit anders angegeben.

2.2.3 Regressionstechniken

QSAR-Methoden basieren meist auf linearen mathematischen Modellen. Wichtigste Ausnahme sind die Neuronale Netzwerke, die sich weit verbreitet haben und auch nichtlineare Zusammenhänge modellieren können. In der vorliegenden Arbeit kamen jedoch ausschließlich lineare Verfahren zum Einsatz. Bei allen Techniken ist das Ziel das Gleiche: Es sollen Regressionskoeffizienten bestimmt werden, die es ermöglichen, mit einer Regressionsgleichung die abhängige Variable (y) jedes Datensatzobjektes möglichst gut zu schätzen [127]. Die Methoden, die in dieser Arbeit zum Einsatz kamen, werden im Folgenden ausführlicher beschrieben.

2.2.3.1 Einfache Lineare Regression

Die einfach lineare Regression verknüpft die abhängige Variable y des Datensatzes mit einer unabhängigen Variablen \mathbf{x} :

$$\mathbf{y} = b_0 + b_1 \mathbf{x} + \mathbf{e} \quad \text{Gleichung 5}$$

Die Korrelation zwischen dem Deskriptor \mathbf{x} und der abhängigen Variablen y wird mittels des Regressionskoeffizienten b_1 hergestellt, der Achsenabschnitt b_0 beinhaltet die relative Verschiebung zum Nullpunkt. Nach erfolgter Datenzentrierung ist $b_0=0$. Der Fehlervektor \mathbf{e} beschreibt den Fehler des Modells. In der QSAR gilt die unabhängige Variable als fehlerfrei.

2.2.3.2 Multiple lineare Regression(MLR)

Bei QSAR-Anwendungen bestehen die \mathbf{X} -Daten fast immer aus mehr als nur einem Deskriptor. Dies führt zur Annahme, dass sich die biologische Aktivität y additiv aus der

Summe der Beiträge einzelner Deskriptoren zusammensetzt. Dabei lässt sich jeder einzelne Deskriptoreinfluss durch einen Regressionskoeffizienten quantifizieren. Dies bedeutet in der mathematischen Darstellung:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_p \cdot x_p + e \quad \text{Gleichung 6}$$

beziehungsweise in der Matrixschreibweise

$$y = \mathbf{X} \cdot \mathbf{b} + e$$

Dabei hat \mathbf{X} die Dimension $n \times p$ und enthält die p Deskriptoren der n Moleküle, \mathbf{b} hat die Dimension $p \times 1$ und beinhaltet die unbekanntenen Regressionskoeffizienten. Der unabhängige Fehler e der Objekte wiederum hat die Dimension $n \times 1$. In der MLR wird der Regressionskoeffizientenvektor als

$$\hat{\mathbf{b}} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} \quad \text{Gleichung 7}$$

geschätzt. $\hat{\mathbf{b}}$ wird dabei so geschätzt, dass die Summe der quadrierten Abweichungen (*RSS*, engl. *Residual Sum of Squares*, siehe Gleichung 15) zwischen den gemäß

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \hat{\mathbf{b}} \quad \text{Gleichung 8}$$

gefundenen und den experimentell ermittelten Werten \mathbf{y} möglichst gering ist. Gleichung 8 ermöglicht auch die Vorhersage der Aktivität neuer Moleküle \hat{y}_{neu} unter Zuhilfenahme der Deskriptorwerte \mathbf{x}_{neu} .

Gleichung 7 kann durch Matrixinversion der Matrix $\mathbf{X}^T \cdot \mathbf{X}$ prinzipiell direkt gelöst werden. Diese Lösung kann jedoch in einigen Fällen nicht verwendet werden. So können beispielsweise unterbestimmte Gleichungssysteme und Gleichungssysteme mit Multikollinearitäten ohne weitere Nebenbedingungen nicht eindeutig gelöst werden.

Unterbestimmte Gleichungssysteme. In der QSAR stehen normalerweise sehr viele Deskriptoren für wenige Moleküle zur Verfügung, so dass unterbestimmte Gleichungssysteme recht häufig anzutreffen sind. Unterbestimmte Gleichungssysteme mit $n < p$, die so genannten „fetten“ Matrizen, haben unendlich viele gleichwertige Lösungen. Die Schätzung des Regressionskoeffizientenvektors $\hat{\mathbf{b}}$ ist nicht eindeutig. Überbestimmte Gleichungssysteme mit $n > p$ sind in der 3D-/4D-QSAR dagegen eher selten.

Multikollinearitäten. Eine exakte Multikollinearität liegt vor, wenn mehrere Spalten der Datenmatrix \mathbf{X} linear voneinander abhängig sind. Wenn eine der beiden entfernt wird, so kann dies sehr leicht behoben werden. Nicht so einfach ist es, wenn zwei Spalten nur näherungsweise linear abhängig sind, also eine so genannte Nahezu-Kollinearität vorliegt.

Dann ist es nicht ausreichend, einfach eine der Spalten zu eliminieren. Es ist zwar möglich, eindeutige Regressionskoeffizienten zu bestimmen, doch wird deren Varianz möglicherweise unakzeptabel hoch. Die mathematische Begründung dafür wird im nächsten Unterkapitel erklärt.

2.2.3.3 Singulärwertzerlegung (SVD)

Die Singulärwertzerlegung (engl. *Singular value decomposition*, SVD) ist ein mathematisches Verfahren, das die Grundlage zur Lösung der genannten Problemfälle in der MLR schafft[128]. Dabei wird die Datenmatrix \mathbf{X} in zwei orthonormale Matrizen \mathbf{U} und \mathbf{V} sowie eine Diagonalmatrix \mathbf{S} zerlegt. Diese enthält die Singulärwerte, also die positiven Wurzeln der Eigenwerte. Die SVD einer Matrix \mathbf{X} liefert:

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \qquad \text{Gleichung 9}$$

Dabei entsprechen die Spalten der Matrix \mathbf{U} den Eigenvektoren von $\mathbf{X} \cdot \mathbf{X}^T$, die Diagonalmatrix \mathbf{S} enthält die Singulärwerte von $\mathbf{X}^T \cdot \mathbf{X}$ und die Spalten von \mathbf{V} geben die Eigenvektoren der Matrix $\mathbf{X}^T \cdot \mathbf{X}$ an. Es gilt $\mathbf{U}^T \cdot \mathbf{U} = \mathbf{I}_n$ und $\mathbf{V}^T \cdot \mathbf{V} = \mathbf{I}_p$, d.h. die Matrizen \mathbf{U} und \mathbf{V} sind orthonormal.

Hauptkomponentenanalyse. Die aus der SVD erhaltenen Matrizen ermöglichen direkt eine Bestimmung der für die Hauptkomponentenanalyse (engl. *Principal component analysis*, PCA) wichtigsten Größen[129]. Somit berechnen sich die Scores \mathbf{T} als

$$\mathbf{T} = \mathbf{U} \cdot \mathbf{S} \text{ bzw. } \mathbf{T} = \mathbf{X} \cdot \mathbf{V} \qquad \text{Gleichung 10}$$

Die Loadings entsprechen den Eigenvektoren \mathbf{V} , die quadrierten Diagonalelemente von \mathbf{S} sind die Eigenwerte.

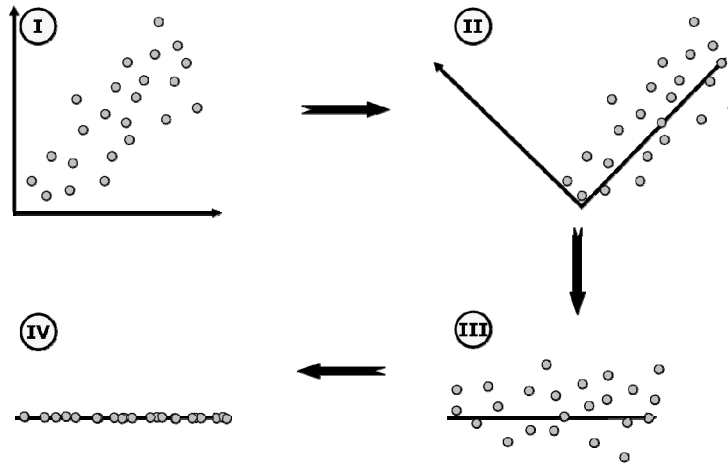


Abbildung 16: Geometrische Deutung der Dimensionsreduktion mit Hilfe der SVD: Das Koordinatensystem der Originaldaten (I) wird derartig gedreht, dass die erste Hauptachse den größten Teil der Varianz der Daten erfasst (II). Dann wird das Ganze auf eine Hauptachse reduziert (III), da so nur ein geringer Teil der Information verloren geht. Im dann entstandenen eindimensionalen Koordinatensystem (IV) bleiben alle Objekte eindeutig voneinander unterscheidbar.

Berechnung der Regressionskoeffizienten. Aus den Gleichungen 8 und 9 ergibt sich nach Anwendung der SVD auf die Matrix \mathbf{X} der Zusammenhang

$$\mathbf{y} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \cdot \mathbf{b} + \mathbf{e} \quad \text{Gleichung 11}$$

Wenn man nun berücksichtigt, dass $\mathbf{X}^{-1} = \mathbf{V} \cdot \mathbf{S}^{-1} \cdot \mathbf{U}^T$ ist der geschätzte Regressionskoeffizientenvektor

$$\hat{\mathbf{b}} = \mathbf{V} \cdot \mathbf{S}^{-1} \cdot \mathbf{U}^T \cdot \mathbf{y} \quad \text{Gleichung 12}$$

Dabei darf \mathbf{S} wegen der Matrixinversion \mathbf{S}^{-1} offensichtlich keine Singulärwerte gleich Null enthalten. Dies wäre der Fall, wenn zwei Spalten der Datenmatrix \mathbf{X} linear voneinander abhängig wären. Überdies sind sehr kleine Singulärwerte problematisch, da die Varianz eines jeden Regressionskoeffizienten $\hat{\mathbf{b}}$ umgekehrt proportional zu den verwendeten Singulärwerten ist [128]:

$$\text{var}(\hat{b}_i) = \sum_{k=1}^r \frac{v_{i,k}^2}{s_{k,k}^2} \cdot \sigma^2 \quad \text{Gleichung 13}$$

Hier sind $v_{i,k}^2$ und $s_{k,k}^2$ die entsprechenden Elemente der \mathbf{V} - bzw. \mathbf{S} -Matrix, r ist der Rang der Matrix. Finden sich also im Nenner von Gleichung 13 kleine Singulärwerte in Dimensionen in denen $v_{i,k}^2$ groß ist, so folgt daraus eine hohe Varianz der Regressionskoeffizienten. Der Informationsgehalt der entsprechenden Variablen in der Originalmatrix wird durch die Größe des Singulärwertes widerspiegelt. Deshalb ist es nur konsequent, Hauptkomponenten, die mit kleinen Singulärwerten verknüpft sind, zu vernachlässigen. Dies wird genau auf diese Weise

bei der im Folgenden beschriebenen Hauptkomponentenregression (engl. *Principal component regression*, PCR) umgesetzt.

2.2.3.4 Hauptkomponentenregression (PCR)

Aus der SVD bekommt man für eine Datenmatrix \mathbf{X} vom Rang r genau r Hauptkomponenten mit Singulärwerten größer Null. Die Hauptkomponentenregression verwendet aber nur q Hauptkomponenten, um die in Gleichung 13 beschriebene Varianz der Regressionskoeffizienten zu reduzieren. In der PCR wird analog Gleichung 12 der Regressionskoeffizientenvektor wie folgt berechnet:

$$\hat{\mathbf{b}}_{PCR} = \mathbf{V}_q \cdot \mathbf{S}_q^{-1} \cdot \mathbf{U}_q^T \cdot \mathbf{y} \quad \text{Gleichung 14}$$

Hierbei ist von entscheidender Bedeutung, wie groß q , also die Anzahl der Hauptkomponenten ist. Zwar soll die Varianz von $\hat{\mathbf{b}}$ möglichst stark reduziert werden, doch führt auch genau diese Vernachlässigung der verbleibenden $r - q$ Hauptkomponenten zu einer Ergebnisverzerrung. Der Kompromiss zwischen Verzerrung und Varianz wird meist durch Anwendung der Kreuzvalidierung bestimmt.

2.2.3.5 Partial Least Squares Regression (PLS)

Die *Partial Least Squares Regression* ist eng mit der PCR eng verwandt [130]. Auch hierbei werden ähnlich der SVD (siehe Abbildung 16) die Daten in ein neues Koordinatensystem projiziert, zu dessen Bestimmung neben den \mathbf{X} -Daten auch die \mathbf{y} -Daten herangezogen werden. Dafür werden die Eigenvektoren leicht rotiert, wodurch die Korrelation der Komponenten (latenten Variablen) mit \mathbf{y} optimiert wird. Ziel ist quasi, den optimalen Kompromiss zwischen den \mathbf{X} - und \mathbf{y} -Eigenvektoren zu finden. Dabei wird die Kovarianz von \mathbf{X} und \mathbf{y} berücksichtigt. Die PLS liefert im Vergleich mit der PCR oft weniger komplexe Modelle, die auf einer geringeren Anzahl von Komponenten beruhen. Dabei ist die Anzahl der tatsächlich verwendeten Freiheitsgrade normalerweise jedoch nahezu identisch. Daher ist die PLS, obwohl sie die etablierte Standardmethode ist, für die QSAR nicht grundsätzlich besser geeignet als die PCR [131].

2.2.4 Datenmodellierung und -validierung

Normalerweise wird mittels einer QSAR-Analyse die biologische Aktivität neuer Substanzen vorhergesagt. Das ist möglich, da auch für „virtuelle“ – also noch nicht synthetisierte – Moleküle die Deskriptoren berechnet werden können. Dies wird dazu genutzt, die Synthesen

für neue Moleküle gezielt zu priorisieren. Ziel dabei ist es, eine Kostenreduktion zu erreichen, da neben den Synthesen auch die Zahl der zu testenden Moleküle sinkt.

Um einen Überblick über die Aussagekraft der erstellten Modelle zu bekommen, ist es immens wichtig, diese gut zu validieren. Damit können in der Praxis auftretende Probleme minimiert werden. Die Qualität und Verlässlichkeit eines QSAR-Modells kann nur durch den Vergleich von Vorhersage- und tatsächlichen Werten ermittelt werden. Der Vorhersagewert wird dabei mittels eines bestehenden QSAR-Modells und den für ein Molekül vorhandenen Deskriptoren berechnet. Er stellt den vermuteten Wert für biologische Aktivität dar. Nebenbedingung dabei ist, dass der betreffende Aktivitätswert nicht in der Modellbildung zum Einsatz gekommen ist. In einer guten Validierung müssen immer Werte, die nicht in der Modellbildung verwendet wurden, zum Einsatz kommen. Dies sind die so genannten Testdaten. Oft sind die Datensätze nicht hinreichend groß, um ohne weiteres in der Modellbildung auf einen Teil der Objekte verzichten zu können. Die Modellbildung würde sonst instabil. Deshalb wird häufig eine Methode der wiederholten Stichprobenziehung, die Kreuzvalidierung (engl. *Cross-validation*, CV), eingesetzt [132]. Dafür ist eine Unterteilung der bestehenden Daten in Trainings- und Testdaten notwendig. Wie diese jeweilige Unterteilung erfolgt, wurde in Abbildung 15 gezeigt.

Bestimmt man die Vorhersagekraft eines QSAR-Modells, so ist nicht die Bestimmung eines Qualitätsmaßes das ausschließliche Ziel. Vielmehr werden darüber auch wichtige Modellparameter ausgewählt. Die Anzahl q der für die Modellbildung verwendeten Hauptkomponenten wird in der Regel dadurch ermittelt, dass Modelle für unterschiedliche Werte von q berechnet und dann verglichen werden. Das Modell mit der besten Vorhersagekraft liefert den optimalen Wert für q und wird für weitere Berechnungen verwendet. Wird eine derartige datengestützte Auswahl getroffen, bei der die Information der abhängigen Variablen genutzt wird, so nennt man diese Validierung eine interne Validierung. Den damit verknüpften Vorhersagefehler nennt man entsprechend internen Vorhersagefehler. Davon abzugrenzen ist die externe Validierung. Bei dieser wird die datengestützte Modellauswahl ohne Information über die abhängige Variable eines Teils der Daten (nämlich der Testdaten) getroffen. Folglich sind die Testdaten völlig unabhängig von der gesamten Modellselektion. Dementsprechend wird der damit verknüpfte Vorhersagefehler als externer Vorhersagefehler bezeichnet.

Diese Entscheidungen werden unter Zuhilfenahme der vorliegenden Daten anhand einer internen Validierung getroffen, bei der alle Daten auch in die Modellselektion eingehen. Gleiches gilt für die Methode der Variablenselektion, die nur einen Teil der vorhandenen

Deskriptoren für die Modellerstellung nutzt. Hierbei muss die optimale Kombination der dann verwendeten Variablen durch die Validierung unzähliger einzelner Modelle herausgefunden werden, die aus verschiedenen Untermengen von Variablen hervorgegangen sind.

2.2.4.1 Kreuzvalidierung

Im Rahmen einer Kreuzvalidierung wird dem Datensatz als Erstes eine gewisse Anzahl von Objekten entnommen, bevor mit den übrigen Objekten ein Modell erzeugt wird. Dieses Modell verwendet man anschließend, um die abhängige Variable der ausgelassenen Objekte vorherzusagen. Dieser vorhergesagte Wert wird mit dem im Laborexperiment bestimmten Wert verglichen und der Grad der Übereinstimmung wird mit einer Gütefunktion charakterisiert. Die verwendete Gütefunktion ist normalerweise die Fehlerquadratsumme (siehe Gleichung 21). Der gesamte Vorgang wird mehrmals unter Auslassung verschiedener Objekte wiederholt. Die genaue Art der Umsetzung dieser Datensatzaufteilung führt dann zu verschiedenen Arten der Kreuzvalidierung.

“Leave-one-out”-Kreuzvalidierung (LOO-CV). Bei dieser Art der Kreuzvalidierung wird dem aus n Objekten bestehenden Datensatz genau eines als „Validierdatensatz“ entnommen. Mit den übrigen $n - 1$ Objekten, also den Konstruktionsdaten, wird ein Modell gebildet. Dieses Modell nutzt man zur Vorhersage der abhängigen Variablen des ausgelassenen Objektes (siehe Abb. 17). Die gesamte Prozedur wird genau n -mal wiederholt, so dass für jedes Objekt genau einmal die abhängige Variable vorhergesagt wurde. Es wird also genau n -mal ein Modell berechnet. Der Einsatz der LOO-CV in dieser Arbeit wird durch die Kennzeichnung der entsprechenden Parameter mit dem Subskript „CV-1“ ersichtlich.

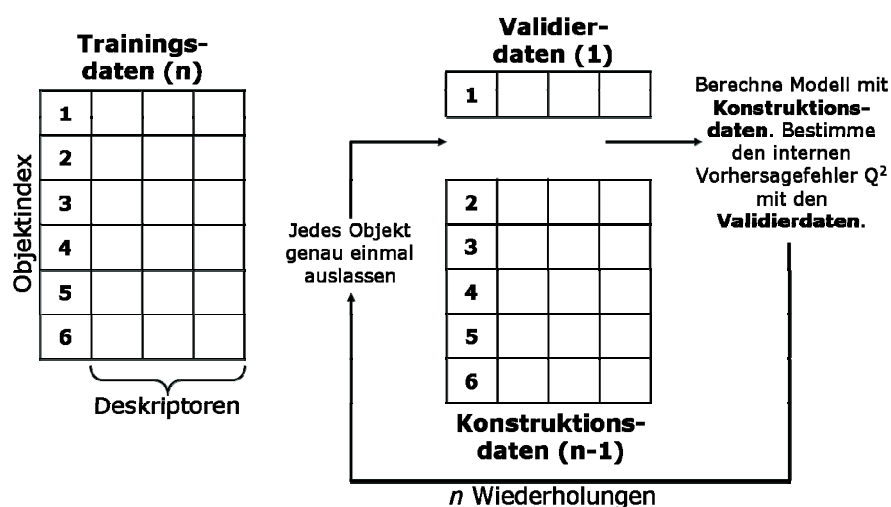


Abbildung 17: Schematischer Ablauf der Leave-one-out-Kreuzvalidierung

ν -fache Kreuzvalidierung (ν -CV). Bei der ν -fachen Kreuzvalidierung wird der Datensatz aus n Objekten in ν möglichst gleich große Gruppen unterteilt. Davon werden $\nu - 1$ Gruppen genommen und aus diesen ein Modell erstellt. Damit werden die abhängigen Variablen der verbleibenden Gruppe, die hier als Validierdatensatz verwendet wird, vorhergesagt. Das hat zur Folge, dass bei der ν -CV genau ν Modelle erstellt werden, da jede Gruppe genau einmal ausgelassen wird. Bei kleinen Werten von ν reduziert sich also die Anzahl der berechneten Modelle und somit der Rechenaufwand. Der Grenzfall $\nu = n$ entspricht der LOO-CV.

“Leave-multiple-out“-Kreuzvalidierung (LMO-CV). Einen Überblick über die LMO-CV zeigt Abbildung 18. Wird nicht nur ein Objekt, sondern $d > 1$ Objekte ausgelassen, so spricht man von der Leave-multiple-out-Kreuzvalidierung. Die $n - d$ Objekte des Konstruktionsdatensatzes werden zur Modellbildung verwendet. Die d Objekte aus dem Validierdatensatz werden vorhergesagt [133]. Diese Prozedur wird ebenfalls mehrfach wiederholt. Im Gegensatz zur ν -CV werden aber die d auszulassenden Objekte jeweils durch eine erneute Zufallsauswahl bestimmt, was eine relativ hohe Anzahl von B Wiederholungen erfordert. Dabei haben sich Werte zwischen $B = 2 \cdot n$ und $B = 3 \cdot n$ als robuste Anhaltspunkte herausgestellt [134]. Subskripte mit „CV- d “ beschreiben in dieser Arbeit den Einsatz der LMO-CV, meistens ist das „CV-50%“, da standardmäßig die Hälfte der Objekte ausgelassen wurde.

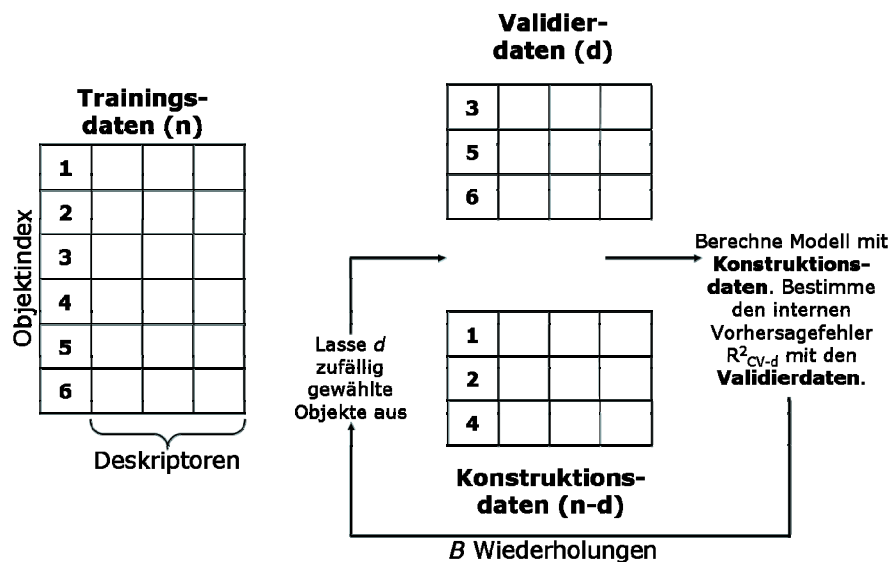


Abbildung 18: Schematischer Ablauf der Leave-multiple-out-Kreuzvalidierung (LMO-CV).

2.2.4.2 Gütekriterien

Die zentrale Unterscheidung, die bei der Validierung eines Modells getroffen werden muss, ist diejenige zwischen interner und externer Validierung. Eine interne Validierung wie die Kreuzvalidierung lässt alle vorhandenen Daten in die Modellauswahl einfließen. Dabei nutzt diese Art der Validierung sehr effizient die Daten, trifft jedoch tendenziell zu optimistische Aussagen bezüglich der Modellqualität. Daher ist eine wirklich belastbare Modellvalidierung nur durch die Vorhersage der abhängigen Variablen eines externen Testdatensatzes möglich. Dieser muss auf Daten beruhen, die nicht bereits in der Modellauswahl Anwendung fanden. Daher existieren für die verschiedenen Arten der Validierung auch unterschiedliche Gütekriterien. Diese sind eng miteinander verwandt und oft sogar identisch, tragen aber häufig gesonderte Bezeichnungen, um ihre verschiedenen Grundlagen zu verdeutlichen.

Wie bereits erwähnt dient ein QSAR-Modell zur Vorhersage biologischer Aktivitäten. Daher wird die Präzision der Vorhersagen zur Bestimmung von Gütekriterien genutzt. Diese Präzision wird durch die quadrierte Differenz zwischen dem berechneten und dem wahren Wert charakterisiert. So wird beschrieben, wie gut ein Modell die Werte wiedergeben kann, auf denen es basiert. Die Summe über alle quadrierten Differenzen wird als Fehlerquadratsumme oder Summe der Abweichungsquadrate (eng. *Residual sum of squares*, *RSS*) bezeichnet und wie folgt berechnet:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{Gleichung 15}$$

Wurden Werte für einen externen Datensatz vorhergesagt, so spricht man von der *Predictive residual sum of squares* (*PRESS*).

$$PRESS = RSS_{Test} = \sum_{i=1}^n (y_{i,Test} - \hat{y}_{i,Test})^2 \quad \text{Gleichung 16}$$

Der auch als Bestimmtheitsmaß bekannte quadrierte Korrelationskoeffizient R^2 beschreibt, inwieweit das Modell die Varianz der abhängigen Variablen erklärt. Für die Datenanpassung gilt $0 \leq R^2 \leq 1$ und

$$R^2 = 1 - \frac{RSS}{SYY} \quad \text{Gleichung 17}$$

mit

$$SYY = \sum_{i=1}^n (y_i - \bar{y})^2 . \quad \text{Gleichung 18}$$

Der quadrierte Korrelationskoeffizient der LOO-CV ist in der gängigen Literatur [135] als Q^2 oder q^2 bekannt.

$$Q^2 = R_{CV-1}^2 = 1 - \frac{PRESS}{SYY} \quad \text{Gleichung 19}$$

Der mittlere quadrierte Fehler der Abweichung zwischen berechnetem und tatsächlichem Wert (engl. *Mean squared error of prediction*, *MSEP*) wird neben der Kreuzvalidierung auch bei der externen Testdatensatzvorhersage verwendet.

$$MSEP = \frac{1}{n_{Test}} \cdot PRESS = \frac{1}{n_{Test}} \cdot \sum_{i=1}^{n_{Test}} (y_{i,Test} - \hat{y}_{i,Test})^2 \quad \text{Gleichung 20}$$

Normalerweise findet man aber den *RMSEP* (engl. *Root mean squared error of prediction*), der die Quadratwurzel dieses Fehlers darstellt:

$$RMSEP = \sqrt{MSEP} = \sqrt{\frac{1}{n_{Test}} \cdot \sum_{i=1}^{n_{Test}} (y_{i,Test} - \hat{y}_{i,Test})^2} \quad \text{Gleichung 21}$$

Über das Subskript, also $RMSEP_{CV-1}$ für die LOO-CV oder $RMSEP_{Test}$ für eine externe Testdatenvorhersage, wird für diese Größe die Art der Validierung spezifiziert.

2.2.4.3 Ensemble averaging

Zentrales Ziel einer jeden QSAR-Analyse ist eine gute externe Vorhersagekraft der gefundenen Modelle. Diese Vorhersagekraft kann durch die Verwendung von Ensemble-Techniken verbessert werden. Dabei wird nicht nur ein einzelnes Modell erstellt, sondern ein ganzes Ensemble von Modellen, die zusammen für die Vorhersage verwendet werden. Dafür ist es nötig, zunächst den Trainingsdatensatz in einer bestimmten Art und Weise zu manipulieren. Damit erhält man einen neuen leicht veränderten Datensatz, der trotzdem die Charakteristik der ursprünglichen Daten besitzt. Beispielsweise kann dafür die abhängige Größe mit einem zusätzlichen Rauschen versehen werden, das dem Zufallsfehler der experimentellen Messung dieser Größe entspricht. Diese Manipulation der Ursprungsdaten wird mehrfach durchgeführt, so dass man ein Ensemble von leicht veränderten und untereinander verschiedenen Trainingsdatensätzen erhält. Aus diesen Daten wird jeweils ein QSAR-Modell erstellt; man erhält ein Ensemble von Modellen.

Jedes der so erzeugten k Modelle des Ensembles wird zur Vorhersage verwendet, so dass schließlich für jede vorherzusagende abhängige Variable \hat{y}_{-i} eine Anzahl k an Ensemble-Vorhersagewerten $\hat{y}_{ens,1}, \hat{y}_{ens,2}, \dots, \hat{y}_{ens,k}$ vorliegt. Das Subskript $-i$ bezeichnet dabei ein Objekt, das in der Modellauswahl nicht verwendet wurde. Der endgültige Vorhersagewert wird über den Mittelwert aller Ensemble-Vorhersagen berechnet:

$$\hat{y}_{ens} = \frac{1}{k} \cdot \sum_{i=1}^k \hat{y}_{ens,i}$$

Gleichung 22

Die den Ensemble-Modellen zugrunde liegenden manipulierten Datensätze können mit unterschiedlichen Methoden erzeugt werden. Diese Methoden unterscheiden sich in der Art und Weise wie sie die Daten verändern. In dieser Arbeit wurde dafür ausschließlich die Methode des Bagging verwendet, die im Folgenden detaillierter erläutert wird. Umfassende Beschreibungen über für diesen Zweck geeignete Methoden finden sich bei Dietterich [136,137] und besonders ausführlich bei Busemann [138].

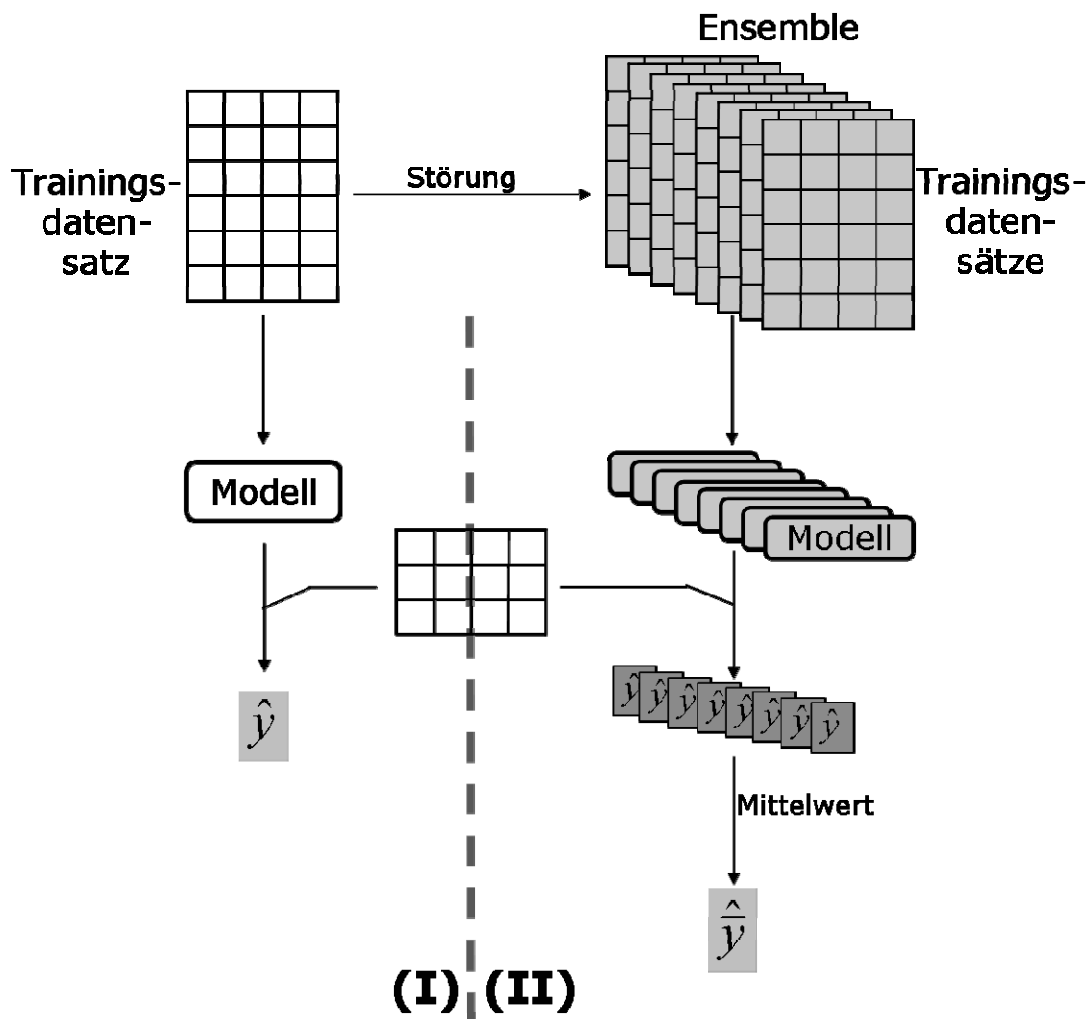


Abbildung 19: Vorhersage von \hat{y} -Werten eines Testdatensatzes mit einem Einzelmodell (I) sowie mit einem Ensemble von Modellen (II)

Das Bagging ist eine Bootstrap-Methode [139], also eine wiederholte Stichprobenziehung mit Zurücklegen. Aus dem Datensatz wird dabei n Mal ein Objekt x mitsamt seiner zugehörigen Variable y gezogen, dem neuen Datensatz hinzugefügt und dann wieder zurückgelegt. Aufgrund dieses Zurücklegens besteht die Möglichkeit, dass einige Objekte mehrfach gezogen werden, so dass ein Objekt im neuen Datensatz auch mehrfach enthalten sein könnte.

Es lässt sich zeigen, dass der neue Datensatz etwa 63 Prozent Unikate enthält. Es gelangen also 37 Prozent der Daten nicht in den neuen Datensatz. Dabei werden die Objekte selbst im Gegensatz zu anderen Ensembletechniken nicht verändert. Die Manipulation des Datensatz basiert ausschließlich auf seiner veränderten Zusammenstellung. Speziell eingesetzt wurde eine Unterart des Bagging, das so genannte Su-Bagging [140,141], das sich als sehr effektiv herausgestellt hat. Dabei werden Objekte, die zwei- oder dreifach gezogen wurden, aus dem Trainingsdatensatz entfernt. Die Objekte, die nicht gezogen wurden, stellen den Testdatensatz dar. Dieses Entfernen von mehrfach gezogenen Objekten ist der einzige Unterschied zwischen Bagging und Su-Bagging. Ein allgemeiner Überblick über das Ensemble-Averaging findet sich in Abbildung 19.

2.3. Variablenselektion

Bei modernen QSAR-Techniken ist es üblich, dass ein Molekül mit sehr vielen verschiedenen Deskriptoren beschrieben wird. Schnell sind hier mehrere hundert Werte berechnet, von denen natürlich nicht alle relevante Information für die jeweilige Fragestellung tragen. Methoden wie die Hauptkomponentenanalyse können diese Daten reduzieren, um so wesentliche Informationen zu extrahieren. Sie benutzen jedoch oft viel irrelevante Information (Rauschen) [142], so dass die entstehenden Modelle nicht optimal sind. Im Optimalfall identifiziert die verwendete Methode aus den vorhandenen Variablen diejenigen, die am Besten mit der abhängigen Variablen korreliert sind. Damit wird eine Beschreibung der Unterschiede zwischen Molekülen ermöglicht. Diese Variablen zu finden ist das Ziel einer Variablenselektion, die aus den vorhandenen p Variablen x_1, x_2, \dots, x_p diejenigen auswählt, welche tatsächlich relevant sind. Für diese Auswahl gibt es 2^p Möglichkeiten, so dass nicht alle daraus resultierenden QSAR-Modelle sinnvollerweise evaluiert werden können. Demgegenüber ergibt sich aus der Vielzahl der Modelle die Gefahr von Zufallskorrelationen [143]. Dies bedeutet, dass die abhängige Variable sehr gut vorhergesagt werden kann, obwohl die ausgewählten unabhängigen Variablen keine sinnvolle Beschreibung des Systems darstellen, da sie zufällig mit der abhängigen Variablen korrelieren. Das bekannteste Beispiel für eine Zufallskorrelation dürften die vielzitierten Störche in Schleswig-Holstein sein [144]. In den Jahren 1965 bis 1980 korrelierte dort die Anzahl der brütenden Storchenpaare sehr gut mit der Anzahl der neugeborenen Babies. Eine

mathematische Lösung zur Steigerung der fallenden Geburtenrate wäre also gewesen, einfach mehr Störche anzusiedeln. Da genau dies offensichtlich versucht wurde, haben die Jahre seit 1980 aber gezeigt, dass eben diese Korrelation nur zufällig war [145]. Während sich die Zahl der Störche nämlich seitdem annähernd verdoppelt hat ist die Geburtenrate faktisch konstant geblieben. Es wird also deutlich, dass die damalige Korrelation nur zufällig war und nicht, wie mehrfach kolportiert wurde, den Beweis liefert, dass der Storch kleine Kinder bringt [145].

Für eine sinnvolle Variablenselektion müssen also zwei wichtige Punkte berücksichtigt werden: Einerseits ist dies eine effiziente Suche nach den „richtigen“ Variablen sowie andererseits eine besonders strenge Modellvalidierung, um Zufallskorrelationen zu verhindern. Hierbei ist vor allem der zweite Punkt sehr wichtig, da erst die Validierung Information darüber liefern kann, ob die vom verwendeten Suchalgorithmus vorgeschlagenen Variablen tatsächlich mit der abhängigen Variable korrelieren. Das bedeutet, dass auch der schnellste Suchalgorithmus keinen Vorteil bringt, wenn die Beurteilung der selektierten Variablen über die so genannte Gütefunktion (engl. *Objective function*) nicht zuverlässig ist.

Mathematisch betrachtet ist eine Variablen-Submenge α zu finden, die die Gütefunktion optimiert. Dabei sei $\alpha \in A$ und A die Menge aller 2^p möglichen Lösungen. Dabei bestehe α aus einer definierten Anzahl von 0 bis p Variablen, wobei die zur Verfügung stehenden Variablen als $\{1, \dots, p\}$ definiert seien. Wird als Gütefunktion der Vorhersagefehler $PE(\alpha)$ des Modells verwendet, so ist folgendes Optimierungsproblem zu lösen:

$$\text{minimiere } PE(\alpha) : \alpha \in A \qquad \text{Gleichung 23}$$

2.3.1 Der Suchalgorithmus

Im Rahmen dieser Arbeit wurde zur Variablenselektion ausschließlich die Tabu-Suche (TS) verwendet. Diese Methode wählt die Variablen schrittweise aus. In jedem Schritt werden ausgehend von der aktuellen Lösung alle Nachbarschaftslösungen erzeugt. Initial startet der Algorithmus an dem Punkt, bei dem alle Variablen den Zustand „nicht enthalten“ haben. Die entsprechenden Lösungen unterscheiden sich nur durch eine einzige Variable (eine mehr oder weniger) von der aktuellen Lösung. Ausgehend von dieser Menge wird der Schritt ausgeführt, der $PE(\alpha)$ nach

$$\Delta PE = PE(\alpha_{\text{Nachbar}}) - PE(\alpha) \qquad \text{Gleichung 24}$$

minimiert. Dieser Schritt zieht die stärkste Verbesserung (engl. *Steepest descent*) der Modellgüte nach sich. Sollten nur die Qualität verschlechternde Schritte möglich sein, so wird derjenige ausgeführt, der die geringste Verschlechterung (engl. *Mildest ascent*) hervorruft. Da auch Verschlechterungen des Modells möglich sind, bleibt der Suchalgorithmus nicht in lokalen Minima stecken und kann theoretisch das globale Minimum finden. Der Name Tabu-Suche rührt daher, dass eine bereits in einem vorherigen Schritt ausgewählte Lösung in späteren Schritten nicht erneut gewählt werden kann. Dies wird bei einer strikten Tabu-Suche so gehandhabt. Es gibt auch weniger strikte Arten der Tabu-Suche, bei denen früher gewählte Lösungen wieder gewählt werden können. In dieser Arbeit wurde jedoch nur die strikte Tabu-Suche eingesetzt. Für die Tabu-Suche zur Variablenselektion muss der Anwender mittels einer Gütefunktion ein Abbruchkriterium definieren.

2.3.2 Die Gütefunktion

Bei den im Rahmen dieser Arbeit durchgeführten Variablenselektionen wurde die zur Definition des Abbruchkriteriums erforderliche Gütefunktion jeweils $3 \cdot p$ -mal durchlaufen. Dieser Wert war bei vorangehenden Untersuchungen als sinnvoll identifiziert worden [38,146,147]. Die Steuerung des Suchalgorithmus, also die Bewertung der möglichen Lösungen, kann nur durch eine Gütefunktion erfolgen. Demzufolge ist die Gütefunktion der wichtigste Bestandteil einer Variablenselektion. Sie ist quasi das Navigationssystem des Suchalgorithmus und weist die Richtung zu einer Verbesserung des resultierenden Modells.

Ferner gibt es eine sehr enge Verknüpfung zwischen der Art der Gütefunktion und der Wahrscheinlichkeit von Zufallskorrelationen. Dies konnte durch Simulationsstudien nachgewiesen werden [38,131,148]. Auch ein analytischer Nachweis für diese Verknüpfung konnte von Shao erbracht werden [134]. Um die Wahrscheinlichkeit für Zufallskorrelationen so weit als möglich zu reduzieren, sollten folgende Regeln eingehalten werden:

- Das Verhältnis Objekte zu Variablen sollte keinesfalls kleiner als sechs sein
- Der Datensatz sollte mindestens 20 Objekte enthalten
- Es sollten zwischen 40 und 60% der Objekte ausgelassen werden, falls eine Kreuzvalidierung als Gütefunktion zum Einsatz kommt

Diese Regeln lassen sich ableiten aus Ergebnissen von Topliss [143,149]. Dessen Untersuchungen wurden mit multipler linearer Regression und R^2 als Gütefunktion durchgeführt. Die dritte Regel macht deutlich, dass die LOO-CV nicht als Gütefunktion für die Variablenselektion geeignet ist. Ausreichend streng in der Modellvalidierung ist demzufolge erst die LMO-CV. In dieser Arbeit wurde bei der Variablenselektion daher stets die LMO-CV mit 50% ausgelassenen Objekten verwendet (also die L50%O-CV), auch die anderen beiden Bedingungen wurden in allen Fällen eingehalten.

3. Ergebnisse und Diskussion

Dieser Teil der Arbeit präsentiert die neu entwickelte xMaP-Methodik, sowie die damit erzielten Ergebnisse. Zusätzlich werden bei dem D₁-Datensatz (Dopaminrezeptor-Antagonisten) über das reine QSAR-Modell hinausgehende Ergebnisse diskutiert. Diese beruhen auf einer Sequenzanalyse sowie einem Homologiemodell des Dopamin-D₁-Rezeptors.

3.1. xMaP – eine interpretierbare alignment-freie 4D-QSAR-Methodik

Der Name der neu entwickelten Technik setzt sich wie folgt zusammen: Der Namensbestandteil MaP weist auf die Vorgängertechnik [14] („*Mapping Property Distributions of Molecular Surfaces*“) hin. Für das „x“ gibt es verschiedene Erklärungen: Ursprünglich als Beschreibung für „extended“ gedacht, so wurde im Laufe der Zeit daraus „flexible“. Grundsätzlich steht das „x“ als Beschreibung für die 4. Dimension, um die die 3D-Technik MaP erweitert wurde.

Die Zielvorgabe wurde bereits in der Einführung dieser Arbeit vorgestellt: xMaP soll eine 4D-QSAR-Methodik sein, die nicht auf einer räumlichen Überlagerung der Moleküle beruht. Die Technik soll ohne den Alignment-Schritt auskommen. Darüber hinaus sollen die Ergebnisse interpretierbar sein. Damit soll die Kommunikation mit medizinischen Chemikern vereinfacht werden. Wie diese Vorbedingungen im Rahmen dieser Arbeit umgesetzt wurden, wird in den folgenden Kapiteln beschrieben.

3.2. Verwandtschaft zu den Vorgängertechniken

Bei xMaP handelt es sich um eine Weiterentwicklung der Vorgängertechniken MaP, DiP und SESP. Deshalb sind einige grundlegende Methoden sehr ähnlich. xMaP

- verwendet distanzabhängige Histogramme
- benutzt eine Variablenselektion, um die Modellinterpretation zu erleichtern
- kommt ohne ein Alignment der Moleküle aus

- nimmt als Ausgangspunkt die Oberflächenberechnung und -eigenschaftszuweisung von MaP
- benutzt die gleichen Oberflächeneigenschaften wie MaP
- ist analog zu MaP gut zu visualisieren und zu interpretieren

Es gibt jedoch mehrere fundamentale Unterschiede, die im Folgenden herausgearbeitet werden:

- Berücksichtigung der Flexibilität der untersuchten Moleküle. Es ist nicht mehr nötig, ein Konformer auszuwählen, von dem angenommen wird, dass es dem biologisch aktiven sehr ähnlich ist.
- Verschmelzung von Oberflächenarealen mit identischen Oberflächeneigenschaften zu Oberflächenarealen, die im Folgenden kurz „Patches“ genannt werden.
- Eine erweiterte unscharfe Zählweise für die Bestimmung des Deskriptorvektors
- Wegfall des Kanonisierungsschrittes

Diese Punkte werden im Folgenden diskutiert.

3.3. Die Berechnung des Deskriptors

Ein Überblick über die Berechnung ist in Abbildung 20 gezeigt. Im Folgenden sollen die einzelnen Schritte, die zur Berechnung der Deskriptoren nötig sind, erklärt werden.

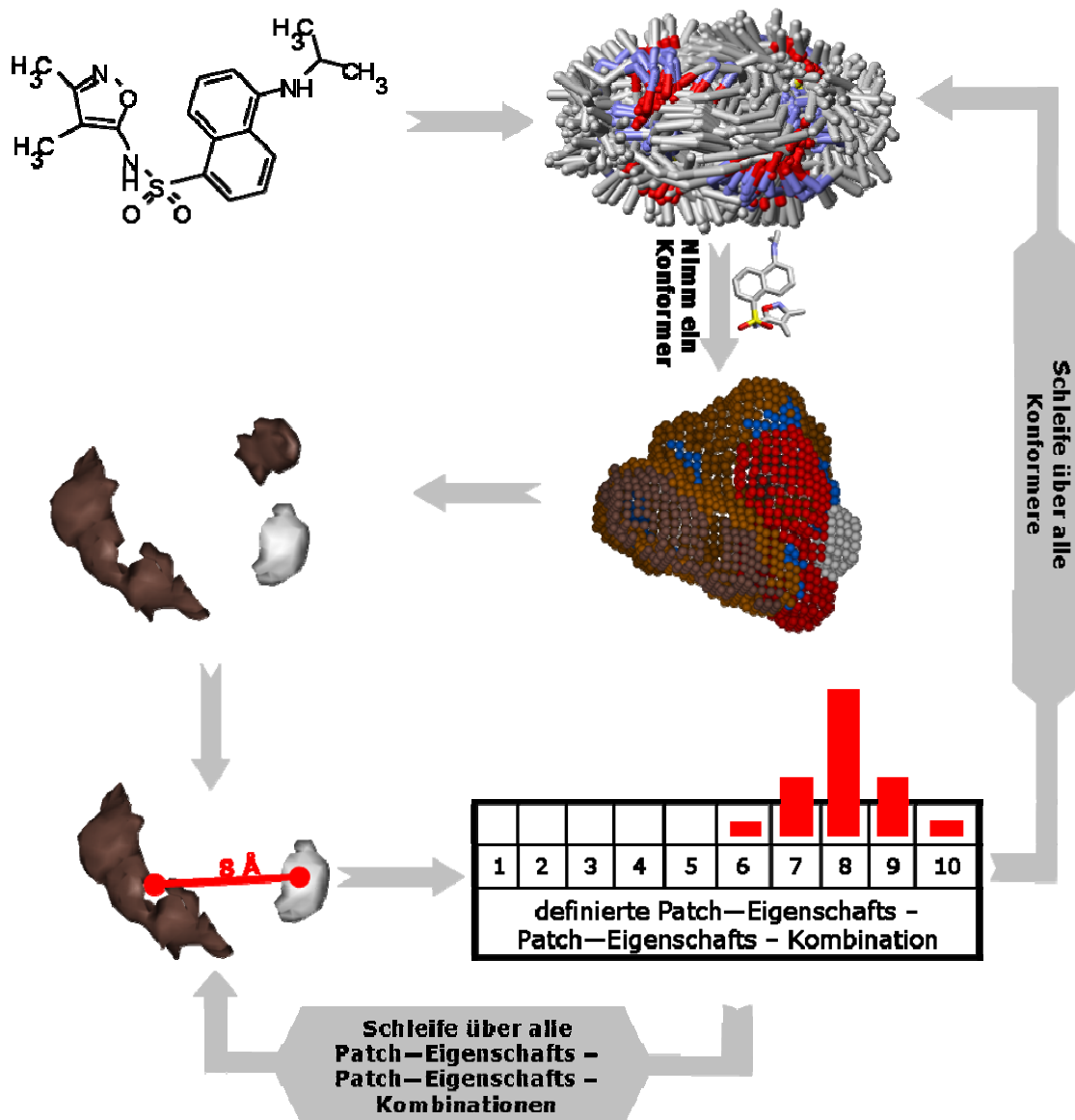


Abbildung 20: Die Berechnung der xMaP-Deskriptoren: Anhand einer 2D-Struktur eines Moleküls wird ein Konformerensemble errechnet. Danach wird sukzessive für jedes Konformer eine molekulare Oberfläche mit gleichverteilten Punkten berechnet. Jedem Punkt werden maximal zwei von fünf Eigenschaften zugewiesen. Regionen mit Punkten, die gleiche Eigenschaften tragen, werden zu so genannten „Patches“ verschmolzen. Für jeden Patch werden die Oberfläche sowie der geometrische Schwerpunkt bestimmt. Diese Information wird durch potenzielle Zweipunkt-Pharmakophore als eine Radialverteilungsfunktion charakterisiert. Dabei wird eine erweiterte unscharfe Zählweise verwendet. Diese Schritte werden für alle Konformere aller Moleküle aus dem Datensatz wiederholt.

3.3.1 Berechnung der Konformerensembles

Wie bereits zuvor erläutert wurde, erfordert die Nutzung der vierten Dimension für die Deskriptorberechnung die Bestimmung eines ganzen Konformerensembles für jedes Molekül. Um dieses zu berechnen werden etablierte Standardmethoden eingesetzt, von denen bekannt ist, dass sie den Strukturraum eines Moleküls gut abdecken. Dies ist einerseits das Programm Catalyst Confirm [150] der Firma Accelrys, sowie andererseits das Programm Omega [151] der Firma Openeye. Beide Programme wurden vor kurzem in einer Arbeit von Kirchmair und Mitarbeitern ausführlich evaluiert [152]. Zur Validierung der xMaP-Technik wurde der Einfluss der Konformerenauswahl untersucht. Es wurden dabei verschiedene Parameter wie die Größe des untersuchten „Energiefensters“ oder die „Konformerengewichtung“ variiert. Damit wurden die daraus folgenden Auswirkungen studiert. Details zu diesen Validierungen sind in Kapitel 3.8. beschrieben. Für die Anwendung des Standardprotokolls zur Deskriptorberechnung werden die Standardparameter der einzelnen Programme eingesetzt. Dies bedeutet, dass bei Catalyst der so genannte „BEST“-Algorithmus verwendet wird. Die Alternative, der so genannte „FAST“-Algorithmus, ist schneller als der „BEST“-Algorithmus, liefert jedoch homogenere Konformere und deckt damit den Konformationsraum der Moleküle weniger gut ab [153-155]. Die Maximalzahl an Konformeren pro Molekül ist auf 255 festgelegt. Außerdem wird ein Schwellenwert für die innere Energie eines Konformers von 20 kcal/mol oberhalb der energieärmsten Konformation definiert. Konformere, die eine innere Energie aufweisen, die oberhalb dieses Schwellenwertes liegt, werden nicht in die danach folgenden Berechnungen einbezogen.

Wenn Omega eingesetzt wird gilt eine Maximalzahl von 400 Konformeren pro Molekül und eine Obergrenze von 15 kcal/mol. Außerdem wird bei Omega bei Konformeren die zueinander einen RMSD-Wert von geringer als 0.8 Ångström aufweisen nur jeweils eines belassen. Alle anderen werden verworfen. Die Abkürzung RMSD steht für „Root Mean Squared Deviation“. Dieser Wert beruht auf Vergleichen einzelner Atome und ihrer Position im Raum. Je ähnlicher die jeweilige Position wird, desto kleiner wird der Wert für die RMSD. Bei sehr ähnlichen Konformeren wird der Wert klein, da die Atompositionen sehr ähnlich sind. Nach einer rigiden Überlagerung der untersuchten Konformeren im Raum ist die mathematische Definition der RMSD wie folgt:

$$RMSD(\mathbf{V}, \mathbf{W}) = \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{v}_i - \mathbf{w}_i\|^2} \quad \text{Gleichung 25}$$

Dabei stehen \mathbf{V} und \mathbf{W} für die Atomkoordinaten jeweils eines Konformers. n ist die Gesamtzahl an Atomen und $\|\mathbf{v}_i - \mathbf{w}_i\| = \sqrt{\sum (\mathbf{v}_i - \mathbf{w}_i)^2}$. Diese muss bei \mathbf{V} und \mathbf{W} identisch

sein. Catalyst führt automatisch ein ähnliches Verfahren durch. Dies geschieht durch Verwendung des so genannten „*Poling*“-Algorithmus [153-155].

Wie bereits früher in dieser Arbeit beschrieben, wurde großer Wert darauf gelegt, die gewählten Parameter der verschiedenen Prozeduren möglichst konstant zu halten. Deshalb wurden sowohl Catalyst als auch Omega immer mit ihren Standardparametern eingesetzt. Bei sehr flexiblen Molekülen war bei Omega in einigen Fällen eine Erhöhung des Wertes für die Anzahl der maximal rotierbaren Bindungen nötig. Diese Erhöhung ermöglicht, dass für die entsprechend großen Moleküle Konformerensembles errechnet werden können. Bei Molekülen mit weniger rotierbaren Bindungen hat die Veränderung des Wertes für rotierbare Bindungen keinen Einfluss. Lediglich die Berechnung wird etwas verlangsamt. Wenn eine Anpassung dieses Wertes bei einem Datensatz nötig war, so ist es vermerkt. Nach der Konformerberechnung liegt für jedes Molekül aus dem Datensatz ein komplettes Konformerensemble vor. Dieses beschreibt den Konformationsraum der einzelnen Moleküle ausreichend und enthält mit großer Wahrscheinlichkeit auch das bioaktive Konformer [156].

3.3.2 Oberflächenberechnung für die Konformere

Das Konformerensemble ist der Startpunkt für alle weiteren Berechnungen. Für jedes einzelne Konformer wird eine Oberfläche aus gleichverteilten Punkten berechnet. Dies geschieht analog zur MaP-Technik, bei der jeweils die Oberfläche für das vermutlich biologisch aktive Konformer jedes Moleküls errechnet wurde. Die Oberfläche wird unter Zuhilfenahme eines modifizierten GEPOL-Algorithmus berechnet. Dieser wurde von Nikolaus Stiefl im Rahmen seiner Dissertation entwickelt [19]. Die Arbeitsweise des Algorithmus ist wie folgt: Das untersuchte Konformer wird in einen Quader gelegt, der es vom Volumen her aufnehmen kann. Dieser Quader ist in viele kleine gleich große Würfel unterteilt. Die Seiten und Eckpunkte der Würfel überlappen. Ein derartiger Quader wird als Gitterbox bezeichnet, die Schnittpunkte an den Kanten von jeweils vier Würfeln werden als Gitterpunkte bezeichnet. Diese Gitterbox ist mit der von Hopfinger verwendeten vergleichbar (vgl. Abb.13). Für den GEPOL-Algorithmus sind jedoch die Gitterpunkte und nicht die Würfelvolumina relevant. Für jeden Gitterpunkt entscheidet der Algorithmus, ob dieser Punkt zur Oberfläche des untersuchten Konformers gehört oder nicht. Nachdem diese Entscheidung für alle Punkte getroffen ist, liegt eine Oberfläche für das untersuchte Konformer vor. Diese Oberfläche besteht aus gleichverteilten Punkten im Raum. Der Algorithmus ist in den entsprechenden Publikationen umfangreich dokumentiert [14,19,38,81]. Auf die Darstellung von Details wird daher verzichtet. Die Unterschiede in der Oberflächenberechnung zwischen

MaP und xMaP sind wie folgt: MaP verwendet eine Kantenlänge von 0.8 \AA für die Würfel. Der maximale Punkt-Abstand auf der Oberfläche liegt daher bei $\sqrt{2} \cdot 0.8 \text{ \AA}$. Dieser Wert wurde für xMaP reduziert. Hier wird ein Abstand zwischen den Punkten von 0.4 \AA eingesetzt. Der maximale Abstand zwischen zwei Punkten liegt demzufolge bei $\sqrt{2} \cdot 0.4 \text{ \AA}$. Die Punktdichte auf der Oberfläche wird deshalb deutlich höher. Positiver Nebeneffekt dabei ist die sehr starke Minimierung des Einflusses des Diskretisierungsfehlers. Dieser Diskretisierungsfehler tritt auf, wenn die untersuchten Moleküle in der Gitterbox bewegt werden. Der Algorithmus nimmt dann andere Punkte als Oberflächenpunkte. Die errechnete Oberfläche verändert sich je nach Position des Moleküls im Raum geringfügig. Bei MaP hat dieser Diskretisierungsfehler einen deutlichen Einfluss auf die Deskriptorberechnung und wird dort durch den Kanonisierungsschritt umgangen. Reduziert man den Punkt-Abstand, so sinkt die Wahrscheinlichkeit für den Diskretisierungsfehler. Bei xMaP wären auch größere Punkt-Abstände kein Problem. Durch die Verschmelzung von Oberflächenbereichen (siehe Kapitel 3.3.3.) zu so genannten „Patches“ wird dieser Diskretisierungsfehler faktisch völlig eliminiert.

Im nächsten Schritt werden jedem Punkt maximal zwei von fünf Eigenschaften (Wasserstoffbrücken-Donor und -Akzeptor, Hydrophilie, starke und schwache Lipophilie) zugewiesen. Eine der Zuweisungen beschreibt die Hydrophilie bzw. Lipophilie des Oberflächenpunktes, die zweite Zuweisung trägt die Information über die Wasserstoffbrückenbindungskapazität. Die Zuweisungen erfolgen regelbasiert auf dem lipophilen Potenzial der dem jeweiligen Punkt nächstliegenden Atome [14,19]. Es werden die von Ghose aufgestellten Regeln verwendet [157,158]. Bei MaP wurde jedem Punkt nur eine Eigenschaft zugewiesen. Bei Punkten mit Wasserstoffbrückenbildungseigenschaften wurde die Zuweisung über die Hydrophilie bzw. Lipophilie entfernt. Es gibt jedoch die Erweiterung MaP+, bei der auch zwei Eigenschaften zugewiesen wurden. Dies hatte aber so gut wie keinen Einfluss auf die Ergebnisse [19]. Für xMaP hat sich in Vorversuchen herausgestellt, dass die Variante mit mehr als einer Zuweisung deutlich bessere Ergebnisse liefert, da die Hydrophilie respektive Hydrophobie kontinuierlich über die ganze Moleküloberfläche beschrieben wird. Daher wird diese Variante durchgängig verwendet.

Ergebnis der Oberflächenberechnung ist für jedes Konformer jeden Moleküls eine Oberfläche mit gleichverteilten Punkten. Jeder Punkt trägt dabei Information über die Eigenschaften der Oberfläche. Damit können die Eigenschaften des Moleküls charakterisiert werden.

3.3.3 Verschmelzung von Oberflächenbereichen mit identischen Eigenschaften

Im nächsten Schritt kommt ein rekursiver Suchalgorithmus zum Einsatz, der Oberflächenbereiche mit Punkten identischer Eigenschaften identifiziert. Dazu startet der Algorithmus auf einem zufällig ausgewählten Oberflächenpunkt. Dessen Umgebung wird so lange abgesucht, bis alle Oberflächenpunkte in der Nachbarschaft identifiziert sind, die die gleiche Eigenschaft tragen. Dieser Schritt wird für alle im vorangegangenen Schritt gefundenen Punkte wiederholt. Können keine weiteren Punkte gefunden werden, ist der gesamte in sich abgeschlossene Bereich, der eine definierte Eigenschaft trägt, identifiziert. Der Suchvorgang wird an einem neuen zufällig ausgewählten Punkt, der noch nicht Bestandteil eines definierten Bereichs ist, fortgeführt. Die Suche wird so lange wiederholt bis alle Punkte Bestandteil mindestens eines Oberflächenbereichs sind. Es kann auf der Moleküloberfläche an unterschiedlichen Stellen Bereiche mit identischen Eigenschaften geben. Diese Areale sind dann räumlich auf der Oberfläche getrennt und werden in den weiteren Berechnungen separat behandelt. Aufgrund der Zuweisung von maximal zwei Eigenschaften zu einem Oberflächenpunkt ist es möglich, dass ein Punkt zu zwei verschiedenen Bereichen (beispielsweise Donor und Hydrophilie) gehört. Alle identifizierten Punkte werden zu einem Oberflächenareal (so genannter „Patch“) verschmolzen. Damit wird eine massive Reduktion der Daten erreicht. Diese Reduktion ist nötig, da ein Konformerensemble extrem viel Information trägt. Darüber hinaus wird die Interpretation der Ergebnisse vereinfacht. Details dazu finden sich in Kapitel 3.5.

3.3.4 Charakterisierung der Oberflächeneigenschaften durch Deskriptoren

Für jedes einzelne Oberflächenareal wird im nächsten Schritt die Oberfläche A nach folgender Formel berechnet:

$$A = n_s \cdot s^2 \qquad \text{Gleichung 26}$$

Hierbei beschreibt n_s die Anzahl an Oberflächenpunkten, die das Areal beschreiben und s den Abstand der einzelnen Punkte zueinander in Å. Die Größe der Oberfläche wird in später folgenden Schritten zur Bewertung der Wichtigkeit des entsprechenden Areals genutzt.

Für die Berechnung des Deskriptors ist es darüber hinaus erforderlich, den Spaltenvektor des geometrischen Schwerpunkts \mathbf{m} des Areals mit seinen Koordinaten zu bestimmen. Diese Berechnung basiert auf den kartesischen Koordinaten (x, y, z) aller n_s Oberflächenpunkte eines Areals:

$$\mathbf{m} = \begin{bmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i / n_s \\ \sum_{i=1}^n y_i / n_s \\ \sum_{i=1}^n z_i / n_s \end{bmatrix} \quad \text{Gleichung 27}$$

Als nächstes werden sämtliche Euklidischen Distanzen zwischen allen Schwerpunkten der verschiedenen Areale berechnet. Diese Information wird in einer Distanzmatrix abgespeichert, die für die Deskriptorberechnung benötigt wird. Diese Euklidische Distanz $d_{i,j}$ zwischen den Schwerpunkten des i ten und j ten Patch wird wie folgt berechnet:

$$d_{i,j} = \sqrt{(\bar{x}_i - \bar{x}_j)^2 + (\bar{y}_i - \bar{y}_j)^2 + (\bar{z}_i - \bar{z}_j)^2} = \|\mathbf{m}_i - \mathbf{m}_j\| \quad \text{Gleichung 28}$$

Durch Nutzung dieser Distanzmatrix werden potenzielle Zweipunkt-Pharmakophore generiert. Diese werden durch die Größe der zugrunde liegenden Areale und der Distanz zwischen den geometrischen Schwerpunkten dieser Areale charakterisiert. Da insgesamt 5 verschiedene Eigenschaften verwendet werden, um die Molekül- bzw. Konformerenergieoberfläche zu beschreiben, gibt es 15 verschiedenen Areal-Eigenschafts-Areal-Eigenschaftskombinationen. Diese Kombinationen sind in Tabelle 1 gezeigt. Jeder Eigenschaft ist dabei eine Farbe zugewiesen, die durchwegs in allen Darstellungen in dieser Arbeit verwendet wird. So werden **Wasserstoffbrücken-Akzeptoren** rot eingefärbt und **Wasserstoffbrücken-Donoren** weiß. **Hydrophile Areale** erhalten eine blaue Farbe, **schwach lipophile Areale** werden hellbraun eingefärbt und **stark lipophile Bereiche** dunkelbraun.

Tabelle 1: Überblick über alle möglichen Eigenschafts-Eigenschafts-Kombinationen während der Deskriptorberechnung und die Abkürzungen die für sie im Rahmen dieser Arbeit verwendet werden. Zusätzlich sind die Spaltenbenennungen in den durchgängig verwendeten Farbcodes markiert.

	H-Brücken-Akzeptor (A)	H-Brücken-Donor (D)	hydrophil(H)	schwach lipophil (Lw)	stark lipophil(Ls)
H-Brücken- Akzeptor (A)	AA				
H-Brücken-Donor (D)	DA	DD			
hydrophil (H)	HA	HD	HH		
schwach lipophil (Lw)	LwA	LwD	LwH	LwLw	
stark lipophil (Ls)	LsA	LsD	LsH	LsLw	LsLs

Für ein einzelnes Konformer besteht der Deskriptor aus einer gewichteten Zählstatistik der einzelnen potenziellen Zweipunkt-Pharmakophore, die in einer Radialverteilungsfunktion abgespeichert werden. Wenn man eine feste Areal-Eigenschafts-Kombination zu Grunde legt, so ist die charakteristische Eigenschaft eines potenziellen Zweipunkt-Pharmakophors die Distanz zwischen den geometrischen Schwerpunkten. Diese Distanz wird durch eine

kontinuierliche Variable beschrieben. Eine Diskretisierung in Distanzkategorien ist erforderlich, um einen Deskriptor zu erhalten, der effizient angewendet werden kann. Dies wird analog bei CATS, DiP und MaP so durchgeführt. Die Distanzkategorien haben eine Auflösung von 1 Å. Alle potenziellen Zweipunkt-Pharmakophore, deren Abstand der Schwerpunkte innerhalb der oberen und unteren Distanzgrenze einer Distanzkategorie liegt, werden in dieser Kategorie zusammen erfasst. Die erste Kategorie startet bei 0.0 Å und es werden genau so viele Kategorien zusätzlich generiert, dass die Maximaldistanz zwischen zwei Arealen im kompletten Datensatz beschrieben werden kann. Die Obergrenze der letzten Kategorie wird mit d_{max} bezeichnet. Sie beschreibt gleichzeitig die tatsächliche Anzahl an Deskriptorvariablen und damit die Dimension p des xMaP-Vektors durch folgende Formel:

$$p = \left(\frac{E \cdot (E + 1)}{2} \right) \cdot d_{max} \quad \text{Gleichung 29}$$

Für fünf verschiedene Eigenschaften ($E=5$) ergeben sich demzufolge $p = 15 \cdot d_{max}$ einzelne Vektorelemente. Im einfachsten Fall könnte man das Vorhandensein der potenziellen Zweipunkt-Pharmakophore zählen. Somit entstünde ein Histogramm dessen Abzisse die Distanzkategorie darstellt. Die Ordinate gäbe die Häufigkeit der potenziellen Zweipunkt-Pharmakophore an. Allerdings würde dieses Vorgehen die unterschiedliche Größe der Oberflächenareale missachten. Deshalb ist das einfache Zählen der potenziellen Zweipunkt-Pharmakophore einem aufwändigeren Wichtungsschema gewichen, die potenziellen Zweipunkt-Pharmakophore werden über einen Wichtungsfaktor W gewichtet. Die untersuchten Konformere unterscheiden sich sowohl in der Anzahl als auch der Verteilung und Größe der einzelnen Areale.

Das Gewicht W eines jeden einzelnen potenziellen Zweipunkt-Pharmakophors wird durch das Produkt der beiden Oberflächen A_i und A_j der zugrunde liegenden Areale bestimmt:

$$W = A_i \times A_j \quad \text{Gleichung 30}$$

Diese Art der Gewichtung gibt potenziellen Zweipunkt-Pharmakophoren, die aus etwa gleich großen Arealen bestehen ein höheres Gewicht als solchen, die sich aus unterschiedlich großen zusammensetzen.

Wie bereits erwähnt, muss die Distanzachse diskretisiert werden. Um einen Diskretisierungsfehler und starke Deskriptorveränderungen für sehr ähnliche Konformere zu vermeiden, wird das Gesamtgewicht eines potenziellen Zweipunkt-Pharmakophors auf mehrere Kategorien verteilt, d.h. mehrere Kategorien werden inkrementiert. Der Begriff „inkrementiert“ bedeutet dabei, dass neue Werte zu den bereits in der entsprechenden

Kategorie befindlichen addiert werden. Neben der Kategorie, die genau die Distanz zwischen den Schwerpunkten der beiden Areale beschreibt, werden auch deren Nachbarn inkrementiert. Dieses Vorgehen wird als unscharfe Zählweise (engl.: *fuzzy increment*) bezeichnet. Die genau die Distanz beschreibende Kategorie wird um $0.5 \times W$ inkrementiert. Die direkten Nachbarn darüber und darunter werden jeweils um $0.2 \times W$ inkrementiert. Die übernächsten Nachbarn um jeweils $0.05 \times W$. Insgesamt wird also um $1.0 \times W$ inkrementiert. Inkremente für negative Distanzkategorien (die beispielsweise bei einer Areal-Areal-Distanz von 1 \AA vorkommen würden) werden verworfen. Dagegen wird für Distanzen größer als d_{max} (die unscharfe Zählweise bei $d=d_{max}$) die Anzahl an Kategorien erhöht. Es werden insgesamt $d_{max}+2$ Distanzkategorien erzeugt. Die genaue Anzahl an Variablen ist $m = 15 \times (d_{max} + 2)$. Dieser Wert ergibt sich aus dem Produkt der verschiedenen Eigenschafts-Eigenschafts-Kombinationen mit der Maximaldistanz. Neben der Reduktion von Diskretisierungsfehlern spiegelt diese Methode der unscharfen Zählweise über fünf Distanzkategorien des gewichteten Zählvektors auch die Tatsache wider, dass die untersuchten Areale eine gewisse Größe haben und nicht nur aus ihrem Schwerpunkt bestehen.

3.3.5 Umsetzung der Konformerendaten zu Moleküldeskriptoren

Für alle Konformere werden mögliche Areal-Eigenschafts-Areal-Eigenschafts-Kombinationen bestimmt. Die entsprechenden Kategorien werden um das den Kombinationen zugrunde liegende Gewicht inkrementiert. Die resultierende Radialverteilungsfunktion beschreibt ein Konformer über die Verteilung und Größe der Oberflächenareale. Die beschriebene Prozedur wird für alle Moleküle aus dem Datensatz wiederholt. Als Ergebnis erhält man eine Matrix der Größe $n_c \times m$. Dabei steht n_c für die Anzahl der Konformere und m für die Anzahl der Variablen. In dieser Matrix beschreibt jede Zeile ein Konformer und jede Spalte das gewichtete Vorkommen eines potenziellen Zweipunkt-Pharmakophors. Für weitere Berechnungen wird diese Matrix zu einem Vektor reduziert, der das gesamte Konformerensemble beschreibt. Dies wird durch die Berechnung des Spaltenmittelwerts über alle Konformere n_c des entsprechenden Moleküls erreicht. Der resultierende Vektor beschreibt die Verteilung der Oberflächeneigenschaften sämtlicher Konformere. Hierbei wird die Flexibilität der Moleküle wie folgt kodiert: Bei Arealen aus sehr flexiblen Bereichen des Moleküls reichert sich das Gewicht nicht in einer einzelnen Distanzkategorie an. Das Gewicht verteilt sich über mehrere Distanzkategorien. Diese Verteilung basiert auf unterschiedlichen Areal-Areal-Distanzen in unterschiedlichen Konformeren und darf nicht mit der unscharfen

Zählweise verwechselt werden. Areale aus sehr rigiden Molekülbereichen akkumulieren dagegen in einer einzigen Distanzkategorie. Der errechnete Durchschnittsvektor charakterisiert somit sowohl die Verteilung der Oberflächeneigenschaften als auch die Flexibilität des untersuchten Moleküls. Um die Durchschnittsbildung zu evaluieren, wurden im Rahmen dieser Arbeit Untersuchungen im Bezug auf unterschiedliche Arten der Wichtung durchgeführt. Die dazugehörigen Ergebnisse sind in Kapitel 3.8.3. gezeigt.

3.4. Erstellung und Validierung der Modelle

Nicht alle errechneten potenziellen Zweipunkt-Pharmakophore sind mit der biologischen Aktivität korreliert. Daher ist es notwendig, die wichtigsten potenziellen Zweipunkt-Pharmakophore mit Hilfe der Variablenselektion zu identifizieren. Die Grundlage dieser Selektion und Validierung bilden Methoden, die bereits erfolgreich eingesetzt und im Theorieteil der Arbeit eingeführt wurden [14,19,38,80,81,146,159]. Die beschriebene Validierung legt den Schwerpunkt auf eine ausführliche und aussagekräftige Testdatenvalidierung. Eine Verbindung von LMO und Ensemble-Averaging-Methoden führt zu einer extrem harten Modellvalidierung. Schematisch ist dies in Tabelle 2 gezeigt. Im Rahmen der Ergebnisdiskussion wird mehrfach darauf hingewiesen, dass eine Variable bei den verschiedenen Unterteilungen der Datensätze häufiger selektiert wurde. Diese Aussage bezieht sich auf Schritt 2 in Tabelle 2. Dieser Schritt wird jeweils 100mal durchgeführt. Eine Variable könnte im Extremfall 100mal selektiert werden. Je öfter eine Variable innerhalb der Iterationen selektiert wird, desto mehr ist eine Diskussion und Interpretation dieser Variable gerechtfertigt, da sie auch nach Veränderung des Trainingsdatensatzes die biologische Aktivität gut erklären kann. Dies weist darauf hin, dass diese Variable nicht zufällig in einer bestimmten Datensatzunterteilung wichtig ist, sondern ein wichtiges Phänomen der Struktur-Wirkungsbeziehungen unabhängig von der Datensatzunterteilung erklärt. In diesem Fall ist zu erwarten, dass die Variable sehr gut mit der biologischen Aktivität korreliert und deswegen eine große Vorhersagekraft hat.

Tabelle 2: Überblick über das Standard-Validierprotokoll bei der Erstellung von xMaP-Modellen

1	Zufällige Aufteilung der vorhandenen Daten in Test- (etwa 37%) und Trainingsdatensatz (etwa 63%). Dieser und alle folgenden Schritte werden 100mal wiederholt
2	Variablenselektion: Tabu-Suche (REM-TS); Abbruch, wenn das Hinzufügen oder Entfernen einer Variable aus/zum Modell im kreuzvalidierten $RMSEP$ ($RMSEP_{CV}$) zu einer Änderung von weniger als 3% oder zur Verschlechterung führt
3	Regressionstechnik: PCR (Standard) oder PLS.
4	Gütefunktion: L50%OCV mit $B = 3 \times n$ Unterteilungen des Trainingsdatensatzes zu Konstruktions- und Validierdatensatz
5	Beschränke die Maximalzahl von Variablen im endgültigen Modell auf $n / 6$
6	Überprüfe die interne Robustheit des selektierten Modells über interne Güteparameter (z.B. $RMSEP_{Test}$ und der zugehörige R^2_{CV} -Wert).
7	Überprüfe das Modell auf zufrieden stellende externe Vorhersagekraft durch Vorhersage der Testdaten aus Schritt 1 nachdem das endgültige Modell bestimmt wurde. Das bedeutet, dass die Testdaten unabhängig vom Modellselektionsschritt sind.
8	Stabilisiere externe Vorhersagen durch Nutzung eines Ensembles von Modellen für die endgültige Vorhersage

3.5. Interpretation und Visualisierung der Modelle

Es ist von großer Wichtigkeit, dass QSAR-Modelle nach ihrer Erstellung interpretier- und kommunizierbar sind. Auf diese Weise wird eine Interaktion mit medizinischen Chemikern erst möglich. Nur ein interpretierbares Modell erlaubt die direkte Erkennung von Regionen am Molekül, deren Modifikation eine verbesserte biologische Aktivität verspricht. Wie bei allen TRI-Techniken ist es bei xMaP nötig, dass jedes einzelne Molekül zusammen mit den Informationen der als wichtig identifizierten Variablen separat untersucht wird. So ist es möglich, Änderungen der biologischen Aktivität zu verstehen und zu erfassen. Dieser Schritt ist relativ zeitaufwändig. Neben der Darstellung einzelner Molekülinformationen muss auch die Information, die ein ganzes Konformerensemble in sich trägt, dargestellt werden. Dazu wurden zwei komplementäre Ansätze entwickelt, die Modellinformationen auf die Strukturen zurückzuprojizieren.

3.5.1 Dreidimensionale Rückprojektion der xMaP-Daten

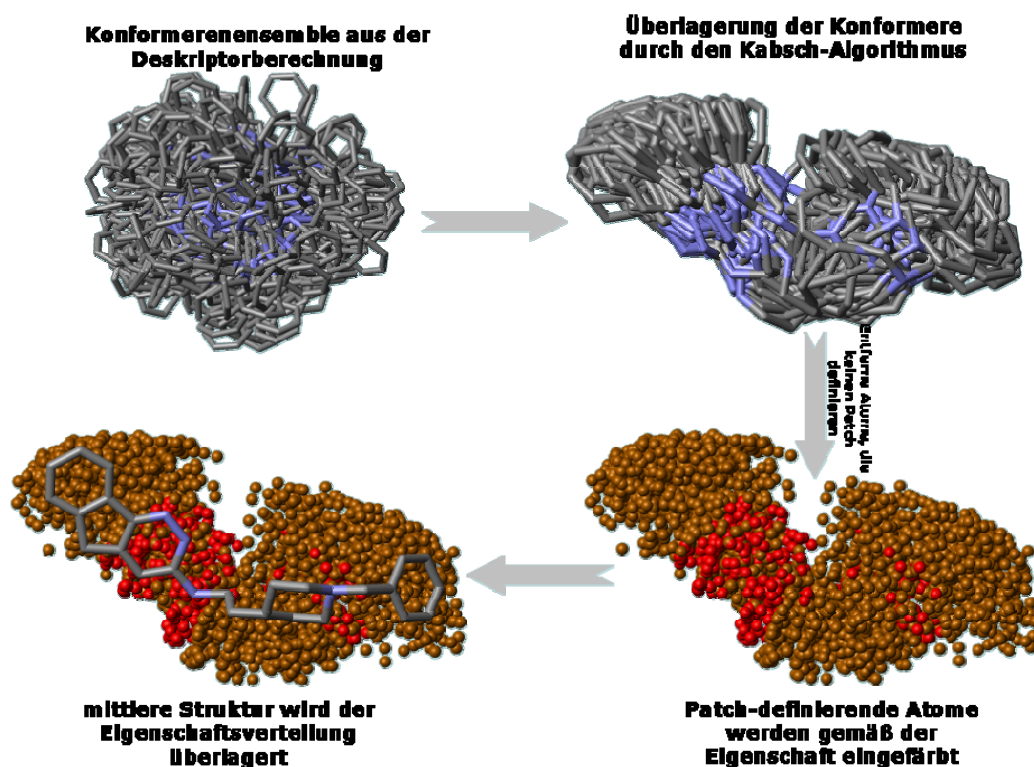


Abbildung 21: Der Weg zur dreidimensionalen Rückprojektion einer xMaP-Variable; Abbildung erzeugt mit MOLMOL [160].

Der exakte Ablauf einer dreidimensionalen Rückprojektion ist in Abbildung 21 gezeigt und soll im Folgenden erläutert werden. Ausgehend von dem für das Molekül berechneten Konformerensemble erfolgen mehrere Schritte, die in MOLMOL [160] durchgeführt werden: Zunächst werden die Konformere eines einzelnen Moleküls mit dem Kabsch-Algorithmus [161] so überlagert, dass ihre räumlichen Unterschiede minimiert werden. Die Zielgröße der Minimierung ist dabei die in Formel 25 beschriebene RMSD zwischen einzelnen Konformeren. Das ist faktisch ein Alignment der verwendeten Konformere aufeinander, darf aber nicht mit dem Alignment-Schritt der 3D-QSAR verwechselt werden. Diese hier durchgeführte Überlagerung erfolgt erst nach der Deskriptorberechnung und betrifft nur die Konformere eines einzelnen Moleküls. Aus der Oberflächenberechnung ist bekannt, welche Atome jeweils für die Generierung der entsprechenden Areale verantwortlich sind. Ist aus der Variablenselektion von einem potenziellen Zweipunkt-Pharmakophor der große Einfluss auf die biologische Aktivität bekannt, so sind bei allen untersuchten Molekülen die dafür verantwortlichen Atome direkt bekannt. Ausgehend von der Darstellung des überlagerten Konformerensembles werden alle Atome entfernt, die nicht zu dem potenziellen Zweipunkt-Pharmakophor gehören. Zudem werden alle Bindungen entfernt. Die

verbleibenden Atome werden gemäß dem vorher definierten Farbcode eingefärbt (siehe Tabelle 1): **Wasserstoffbrücken-Akzeptoren** werden rot eingefärbt und **Wasserstoffbrücken-Donoren** weiß. **Hydrophile Areale** erhalten eine blaue Farbe, **schwach lipophile Areale** werden hellbraun eingefärbt und **stark lipophile Bereiche** dunkelbraun. Diese Darstellung wird mit dem zum Mittelwertkonformer ähnlichsten Konformer ($\min(RMSD)$) aus dem errechneten Ensemble überlagert. Dies ist der letzte in Abbildung 21 gezeigte Schritt. Aus dieser Abbildung wird neben einer Identifizierung der wichtigen Strukturelemente auch die Flexibilität der beschriebenen Molekülbereiche ersichtlich. Eine Interpretation einer Reihe von Darstellungen für verschiedene Moleküle ergibt, ob bestimmte Bereiche flexibel oder rigide sein müssen. Außerdem zeigt sich, wie entsprechende Molekülteile oder -gruppen zueinander stehen. Diese Darstellung ist relativ ähnlich zu der bei MaP gewählten [14,19,38].

3.5.2 Zweidimensionale Rückprojektion der xMaP-Daten

Bei xMaP ist es zusätzlich möglich, die Daten auf eine zweidimensionale Moleküldarstellung zu projizieren. Dies kommt dem Bild, das Chemiker von Molekülen haben, am nächsten und verspricht eine gute Kommunizierbarkeit der erstellten Modelle. Der Ablauf der Berechnung einer solchen Darstellung ist in Abbildung 22 gezeigt:

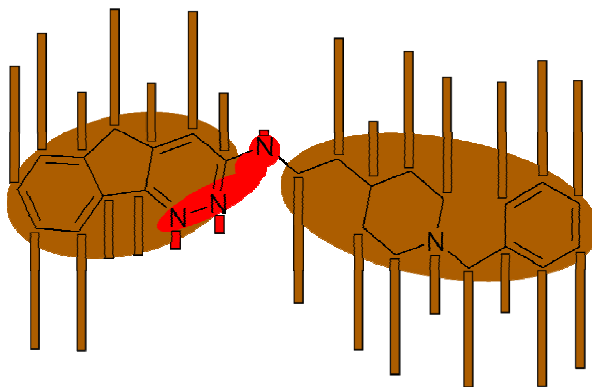


Abbildung 22: Die zweidimensionale Rückprojektion einer xMaP-Variable. Hier ist eine hypothetische LwD-Variable gezeigt.

Bei allen Konformeren ist bekannt, wie stark jedes Atom zu der Generierung eines Areals und damit zu einem potenziellen Zweipunkt-Pharmakophor beiträgt. In der 2D-Darstellung werden alle Atome, die zu dem entsprechenden Areal beitragen, farblich hinterlegt. Die Farbe richtet sich nach dem in Tabelle 1 definierten Farbcode. Das Ergebnis wäre Abbildung 22 ohne Balken. Um zusätzlich die Wichtigkeit der einzelnen Atome zu charakterisieren wird berechnet, wie stark sie im Schnitt zu dem jeweiligen Areal beitragen. Dafür wird folgende Formel genutzt:

$$c_{Atom} = \frac{\sum_{i=1}^{n_{conformers}} \frac{W_{P2PP_i}}{n_{Atom}}}{n_{conformers}} \quad \text{Gleichung 31}$$

Hier beschreibt c_{Atom} den Beitrag eines einzelnen Atoms zu dem entsprechenden Areal innerhalb der kompletten Konformerensemble. W_{P2PP_i} steht für das Gewicht des untersuchten potenziellen Zweipunkt-Pharmakophors. n_{Atom} ist die Gesamtzahl an Atomen, die einen Beitrag zu dem entsprechenden Areal leisten und $n_{conformers}$ gibt die Anzahl an Konformeren des untersuchten Moleküls an. Ergebnis ist eine Darstellung bei der die Wichtigkeit jedes Atoms außer mit der Farbe auch mit einem Balken der Höhe $h=c_{Atom}/c_{max}$ illustriert wird. Hierbei steht c_{max} für den höchsten Beitrag, den ein Atom für ein Areal liefert. Damit wird die Stärke des Beitrages eines einzelnen Atoms zu dem als wichtig identifizierten Areal dargestellt.

3.5.3 Fazit zur Interpretierbarkeit

Die Kombination dieser zwei- und dreidimensionalen Darstellungen liefert eine sehr aussagekräftige, aber einfache Darstellung der potenziellen Zweipunkt-Pharmakophore des Moleküls, die zur Erklärung der biologischen Aktivität wichtig sind. Am untersuchten Molekül können Regionen und Substituenten identifiziert und anschließend Änderungen vorgeschlagen werden, um eine bessere vorhergesagte biologische Aktivität zu erreichen.

3.6. Einbindung von Informationen über die Targetstruktur

Aufgrund des enormen Fortschrittes in der Strukturbiologie [2,3] und der Modellierung von Proteinstrukturen [162-164] ist die Struktur des Zielenzym oder -rezeptors oftmals im atomaren Detail bekannt. Diese Information kann genutzt werden, um Wirkstoffmoleküle zu optimieren. Folglich sollte versucht werden, diese Information auch in QSAR-Analysen einfließen zu lassen. Die neu entwickelte xMaP-Technik beruht auf Konformerensembles für einzelne Moleküle. Die Konformerensembles können daher nicht nur mit Konformationssuchealgorithmen ohne geometrische Beschränkung, sondern auch auf Basis der Zielstruktur berechnet werden. Am Besten dafür geeignet sind Dockingprogramme, die versuchen, die einzelnen Liganden optimal in die vorliegende Bindetasche einzupassen. Dieses Einpassen (engl. *Posing*) funktioniert mittlerweile recht gut [165-170], auch wenn es

nie eine einzelne eindeutige Lösung geben wird (siehe zum Vergleich auch Abbildung 26). Ein Dockingprogramm generiert immer viele mögliche Lösungen, wie ein Ligand im Protein positioniert sein kann. Diese Lösungen werden dann anhand verschiedener integrierter Gütefunktionen (engl. *Scoring*) bewertet, so dass ein Ranking der entstandenen Lösungen vorgenommen werden kann. In strukturbasierten 3D-QSAR-Ansätzen wird oft die mittels dieser empirischen *Scoring*-Funktion bestimmte „beste“ Lösung als Start für die Analyse ausgewählt. Wie bei jeder 3D-QSAR-Analyse stellt dies eine subjektive Auswahl dar und vernachlässigt wichtige Informationen. Fasst man die Vielzahl dieser Lösungen als Konformerensemble auf, so kann diese als Startpunkt einer xMaP-Analyse eingesetzt werden. Es ist interessant, welchen Einfluss die Qualität der Proteinstruktur auf die Qualität und Aussagekraft der QSAR-Gleichung hat. Exemplarisch ist dies in Abbildung 23 für ein Molekül aus dem AZT-Datensatz (siehe Anhang II.1) gezeigt. Rechts ist das Ergebnis der Konformationssuche und links ist das Ergebnis des Dockings gezeigt. Es fällt auf, dass die im Docking generierte Familie deutlich homogenere Konformere beinhaltet.

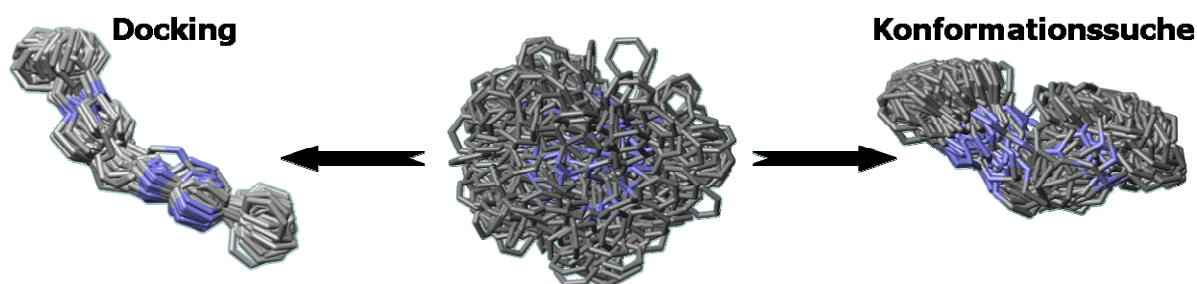


Abbildung 23: Vergleich von Konformerensembles aus dem Docking und aus einer klassischen Konformationssuche ohne räumliche Beschränkung.

Der Grund dafür ist in der Einschränkung des Konformationsraums durch die Bindetasche zu suchen. Die Einschränkung des Konformationsraums durch die Bindetasche erfolgt mit Informationen über tatsächlich bekannte Gegebenheiten als Fundament. Das steht im Gegensatz zu willkürlichen Beschränkungen des Konformationsraums. Deshalb ist eine konformelle Einschränkung *in vivo* zu erwarten, die der im Docking getroffenen sehr ähnlich ist. Die Verwendung strukturbasierter Information ist daher wünschenswert. Der Einfluss dieser Information auf das Ergebnis der QSAR-Analysen wird in folgenden Kapiteln diskutiert.

3.7. Untersuchte Datensätze

Im Rahmen dieser Arbeit wurde eine Vielzahl pharmazeutisch relevanter Datensätze untersucht. Das Ziel war ein möglichst deutliches Bild über die Anwendbarkeit von xMaP zu erhalten. Für alle Datensätze wurde das etablierte Standardprotokoll verwendet und das entsprechende Modell jeweils interpretiert. Wenn Änderungen oder Ergänzungen nötig waren, so ist dies bei den entsprechenden Datensätzen vermerkt. Zunächst werden die untersuchten Datensätze mit ihrem Hintergrund vorgestellt. Im darauf folgenden Überkapitel wird detailliert auf die Ergebnisse eingegangen.

Die in Klammern nach den Unterüberschriften im folgenden Kapitel angeführten Abkürzungen für die Datensatznamen werden aus Übersichtlichkeitsgründen in dieser Arbeit durchweg verwendet.

3.7.1. Inhibitoren der Acetylcholinesterase (Abkürzung: AZT)

Dieser Datensatz wurde bereits erfolgreich mit verschiedenen QSAR-Methoden wie MaP und GRID/GOLPE beforscht und eignet sich daher sehr gut als Benchmarkdatensatz [80,171]. Des Weiteren ist die Struktur der Acetylcholinesterase (AChE) im atomaren Detail bekannt [172-178]. Zudem ist die bioaktive Konformation verschiedener Liganden aus den entsprechenden Publikationen bekannt. Daher eignet sich dieser Datensatz sehr gut für QSAR-Studien, die Informationen über die Proteinstruktur mit berücksichtigen. Ein solches strukturbasiertes Modell wurde im Rahmen dieser Arbeit erstellt. Der Einfluss der strukturbasierten Information auf das Modell wurde evaluiert. Der Datensatz besteht aus insgesamt 49 flexiblen Aminopyridazinen [171]. Diese sind von der antidepressiv wirkenden Leitstruktur Minaprin [179] abgeleitet und stellen Hybride des Minaprins und des Donepezils [180] dar. Die fünf verschiedenen Grundstrukturen sind in Abbildung 24 gezeigt.

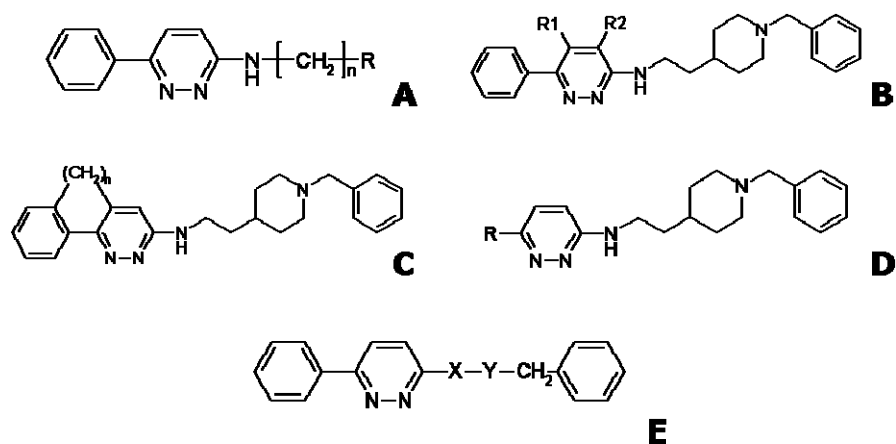


Abbildung 24: Grundgerüste der Strukturen des AZT-Datensatzes. Alle Strukturen zusammen mit ihren Bezeichnungen finden sich im Anhang II.1.

Acetylcholinesterasehemmer gehören in die Gruppe der indirekten Parasympathomimetika, was bedeutet, dass es zu einer erhöhten AcetylcholinKonzentration kommt. Dies wiederum führt zu einem erhöhten Parasympathikustonus und einem erhöhten Tonus der quergestreiften Muskulatur. Das am stärksten im Fokus der Forschung liegende Gebiet im Bereich der AChE-Inhibitoren sind Substanzen wie z.B. Donepezil und Rivastigmin, die bei einer Alzheimer-Therapie eingesetzt werden können. Außerdem gehören zu dieser Substanzkategorie Antidote (z.B. Neostigminbromid) für Vergiftungen durch stabilisierende Muskelrelaxantien.

Die biologische Aktivität dieser Substanzen wurde an der AChE des Kalifornischen Zitterrochenes *Torpedo californica* nach der Methode von Ellman bestimmt [181].

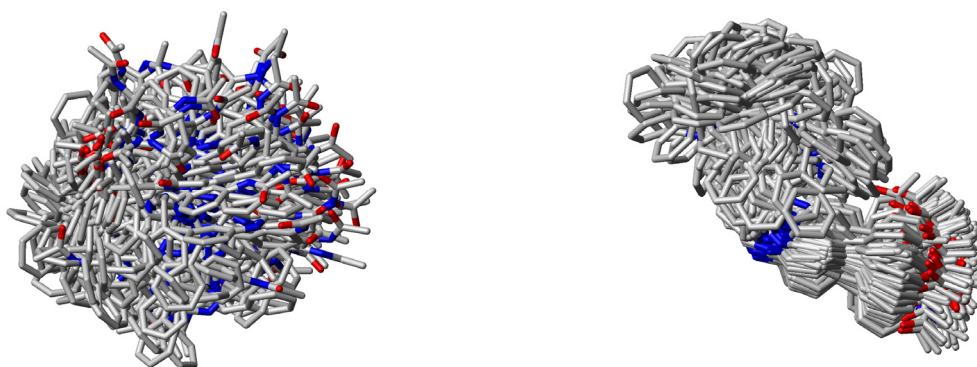


Abbildung 25: Die Substanz 24 aus dem AZT-Datensatz als Konformerensemble direkt aus Catalyst (links) und nach Überlagerung mit dem Kabsch-Algorithmus [161] (rechts).

Als Ausgangspunkt für alle weiteren Analysen wurden die im Rahmen von mehreren Arbeiten erzeugten 3D-Strukturen der Verbindungen verwendet [19,80]. Einerseits wurden sie in Catalyst eingelesen und dort für jedes Molekül mehrere Konformerensembles mit dem BEST-Algorithmus und unterschiedlich großen Energiefenstern (5, 10 und 20 kcal) erzeugt. Andererseits wurden die Verbindungen in OMEGA eingelesen und dort mit den

Standardparametern ein Konformerensemble erzeugt. Die Flexibilität dieser Substanzen ist exemplarisch anhand von Verbindung 24 in Abbildung 25 dargestellt. Diese Verbindung ist in Anhang II.1. gezeigt. Ein weiteres Konformerensemble wurde durch Docking erzeugt. Die einzelnen Inhibitoren wurden in der Kristallstruktur der Acetylcholinesterase [182,183] (PDB-ID: 1EVE) in die bekannte und sehr gut definierte Bindetasche (siehe Abbildung 26) mittels des Programms FlexX [184-186] gedockt. Die jeweils besten 50 Lösungen pro Molekül nach FlexX-Gütefunktion wurden als Konformerensemble in der weiteren Analyse verwendet.

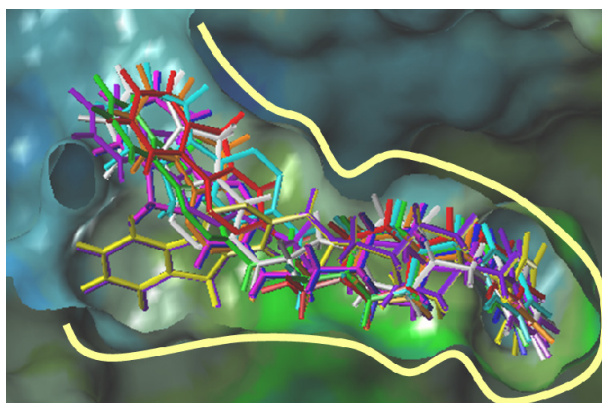


Abbildung 26: Die Bindetasche der Acetylcholinesterase mit verschiedenen Dockinglösungen zu Substanz Nr. 30 aus dem AZT-Datensatz.

Für diesen Datensatz lagen insgesamt fünf auf verschiedene Art und Weise erzeugte Konformerensembles für jedes der Moleküle vor. Diese Konformerensembles waren die Grundlage für verschiedene Berechnungen von xMaP-Deskriptoren. Die Ergebnisse sind in den entsprechenden Kapiteln gezeigt.

3.7.2. Prostaglandin $F_{2\alpha}$ -Analoga ($PGF_{2\alpha}$)

Dieser Datensatz [187-189] wurde ausgewählt, da er einer der ersten war, der je mit einer 4D-QSAR-Methode untersucht wurde [15]. Außerdem wurde er mit einer weiteren neu entwickelten QSAR-Methode untersucht, die die Flexibilität über eine automatische Konformerenauswahl modelliert [190]. Der Datensatz besteht aus 38 Analoga des Prostaglandins $F_{2\alpha}$ und wurde aus zwei Datensätzen erstellt, die den antinidatorischen Effekt der Moleküle als ED_{50} -Wert bei Ratten bzw. Hamster messen [187-189]. Die Grundstruktur ist in Abbildung 27 gezeigt. Strukturvariationen erfolgten sowohl an der α - als auch an der ω -Kette der Moleküle im Datensatz. Einige der biologischen Aktivitäten liegen jeweils nur für eines der beiden Tiere vor. Daher wurden die ED_{50} -Werte relativ zu den Werten von $PGF_{2\alpha}$ skaliert, für das beide Werte vorliegen. Um die Modellbildung zu vereinfachen wurden die

logarithmierten Werte des Verhältnisses des ED_{50} -Wertes von $PGF_{2\alpha}$ zu denen der untersuchten Substanz als abhängige Variablen verwendet. Dies wurde analog von Hopfinger so durchgeführt [15].

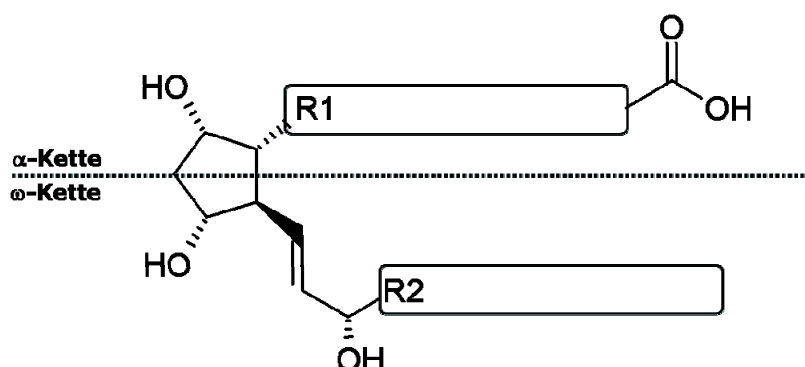


Abbildung 27: Grundgerüst der Strukturen aus dem $PGF_{2\alpha}$ -Datensatz. Der komplette Datensatz ist in Anhang II.2. gezeigt.

Im ersten Schritt wurden die Strukturen mit ISIS Draw [191] gezeichnet und anschließend mit Corina [192-194] zu dreidimensionalen Strukturen konvertiert. Diese 3D-Strukturen wurden in Catalyst [150] eingelesen und mittels der Standardparameter ein Konformerensemble für jedes der untersuchten Moleküle erzeugt. Analog dazu wurde für die Deskriptorvalidierung ein Konformerensemble mit Omega [195,196] erzeugt. Mit den beiden Konformerensembles als Ausgangspunkt wurden xMaP-Deskriptoren berechnet, die Gegenstand weiterer Untersuchungen waren. Die Strukturen sind im Anhang in Kapitel II.2. gezeigt.

3.7.3. Modulatoren des muskarinischen M_2 -Rezeptors (M_2)

Dieser Datensatz wurde aufgrund der extrem hohen Flexibilität seiner Moleküle ausgewählt. Insgesamt besteht er aus 44 allosteren Modulatoren des muskarinischen M_2 -Rezeptors. Die Grundstruktur und die Leitstruktur W84 sind in Abbildung 28 zu sehen.

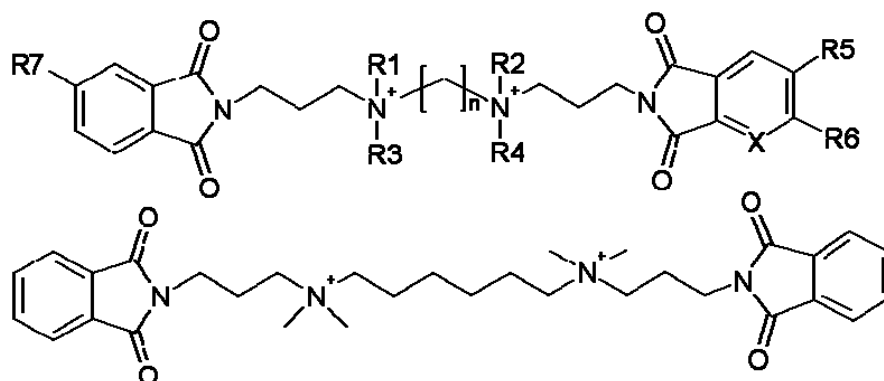


Abbildung 28: Oben das Grundgerüst der untersuchten Substanzen, unten die Leitstruktur W84.

Als biologische Aktivität wird hier der pEC_{50} -Wert der Substanzen verwendet. Dieser beschreibt, wie stark die entsprechende Substanz die Dissoziation des Radioliganden [3H]-N-Methylscopolamin vom M_2 -Rezeptor verhindern kann. Die Messung erfolgt hierbei in einem *in-vitro*-Testsystem, das auf einer Suspension myocardialer Membranen von Schweineherzen basiert. Der erhaltene Wert ist jedoch abhängig vom Medium, daher wurden nur Daten benutzt, die in einem $MgHPO_4$ -Puffer Testsystem gemessen wurden.

Allostere Modulatoren erhöhen oder vermindern die Affinität des orthosteren Liganden, dem Liganden der eigentlichen Bindungsstelle, zum Rezeptor. Dies beeinflusst sowohl Assoziation als auch Dissoziation. Damit ergibt sich eine zusätzliche Möglichkeit, die Aktivität bereits bekannter Liganden zu beeinflussen. Daraus ergibt sich der Vorteil, dass eine verminderte Dosis des orthosteren Liganden verwendet und damit mögliche unerwünschte Arzneimittelwirkungen verhindert werden können. Zum Beispiel kann bei Organophosphat-Vergiftungen die Schutzwirkung des Antagonisten Atropin am cholinergen Muskarin-Rezeptor durch die Gabe eines allosteren Modulators erhöht werden. Auch bei Alzheimer erscheint ein Einsatz allosterer Modulatoren denkbar [197].

Ausgangspunkt der hier erfolgten Analysen waren die für MaP generierten Ausgangskonformere für die 44 Moleküle [14,19,38]. Diese wurden in Catalyst beziehungsweise Omega eingelesen. Mit Hilfe des jeweiligen Programms wurden entsprechende Konformerensembles mit den Standardparametern erzeugt. Bei Omega war es nötig, die Anzahl maximaler rotierbarer Bindungen auf 35 anzuheben. Ansonsten hätten nicht alle Strukturen verarbeitet werden können. Dies hat jedoch keinen Einfluss auf das Ergebnis, erhöht aber die Rechenzeit geringfügig. Auch hier wurden jeweils für beide Konformerensembles die xMaP-Deskriptoren berechnet. Die Strukturen und Substituenten der Strukturen mitsamt ihren pEC_{50} -Werten sind im Anhang in Kapitel II.3. aufgelistet.

3.7.4. Inhibitoren der HIV-1 Reversen Transkriptase (HEPT)

Dieser Datensatz besteht aus 80 Derivaten von 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymin (HEPT), die als Inhibitoren der HIV-1 Reversen Transkriptase(RT) aktiv sind. Die biologische Aktivität wird als Logarithmus $1/C$ angegeben. C ist hierbei die molekulare Wirkstoffkonzentration, die nötig ist, um einen 50%igen Schutz von MT-4-Zellen gegen die zytopathogenen Effekte des HIV-1 (HTLV-IIIb-Stamm) zu erreichen [198]. Die Reverse Transkriptase dient Retroviren dazu, ihre einsträngige RNA in einzelsträngige DNA

zu übersetzen. Direkt danach können sie dann einen komplementären Strang DNA erzeugen. Die daraus resultierende doppelsträngige DNA kann in die Wirts-DNA eingebaut werden. Damit ist die Infektion gestartet und die Viren werden repliziert. Die Grundstruktur der untersuchten Substanzen ist in Abbildung 29 zu sehen.

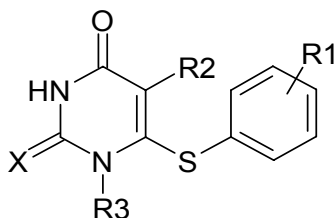


Abbildung 29: Grundstruktur der hier verwendeten HEPT-Derivate.

Auch für diesen Datensatz ist die Struktur des Zielenzym im molekularen Detail bekannt (PDB: 1C1C) [199,200]. Deswegen wurde er zur Untersuchung mit xMaP ausgewählt. Als Ausgangsstrukturen wurden die im Rahmen der Diplomarbeit von Matthias Busemann verwendeten Strukturen benutzt [201]. Diese Strukturen wurden direkt in Catalyst respektive Omega eingelesen und die dort erhaltenen Strukturfamilien als Ausgangspunkt für die Berechnungen von xMaP-Deskriptoren eingesetzt. Außerdem wurde mit dem Programm FlexX [184-186] in die Bindetasche gedockt. Mit dem derart generierten Konformerensemble wurde ein weiteres Modell erstellt. Strukturen und biologische Daten sind im Anhang in Kapitel II.4. gezeigt.

3.7.5. Dopamin-Antagonisten (D₁)

Dieser Datensatz wird aktuell in der Gruppe von Herrn Professor Lehmann am Institut für Pharmazeutische Chemie der Friedrich-Schiller-Universität Jena beforscht [202-218]. Es handelt sich um insgesamt 196 Substanzen mit antagonistischer Aktivität an den verschiedenen Dopaminrezeptoren (D₁ bis D₅). Im Speziellen waren es beim ersten Ansatz 77 Verbindungen mit einer Aktivität am D₁-Rezeptor. Die Leitstruktur LE-300 und die dahinter stehende Grundidee sind in Abbildung 30 gezeigt. Die biologische Aktivität wird dabei als K_i-Wert gemessen. Für die Berechnung wurde der negative dekadische Logarithmus dieser Werte als abhängige Variable verwendet.

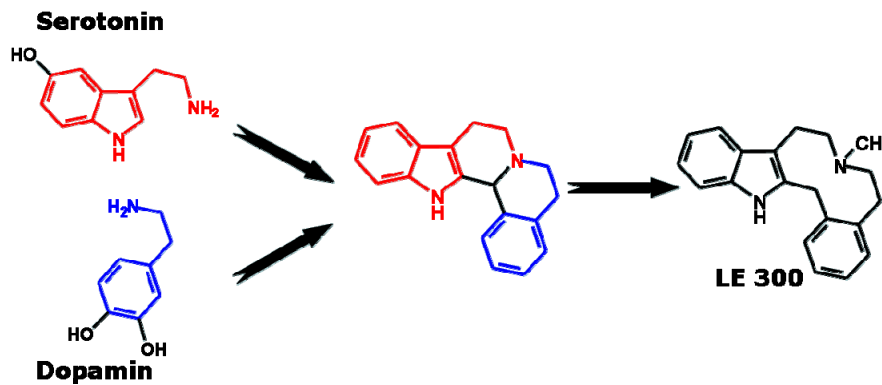


Abbildung 30: Die Leitstruktur der Dopamin-Antagonisten

Dopamin-Rezeptoren können in zwei große Unterklassen unterteilt werden. Einerseits ist dies die D₁-Familie, die aus dem D₁- und D₅-Rezeptor besteht. Eine Stimulation dieser Rezeptoren stimuliert über ein G-Protein das Enzym Adenylatcyclase, das dann wiederum in einer Signalkaskade weitere Enzyme aktiviert. Eine Anregung dieser Rezeptoren wirkt aktivierend auf die Zelle. Rezeptoren der D₂-Familie (D₂ bis D₄) dagegen wirken hemmend. Bei ihrer Stimulation wird ein inhibitorisches G-Protein freigesetzt, das die Adenylatcyclase in ihrer Aktivität hemmt. Damit wird die nachfolgende Signalkaskade blockiert. Zusätzlich werden Kalium-Kanäle aktiviert, was das Ruhepotenzial von Nervenzellen stabilisiert [219,220]. Dopamin-Antagonisten verhindern die Bindung von Dopamin und anderen Agonisten an die entsprechenden Rezeptoren. Für die Therapie bedeutet dies, dass Antagonisten der Dopamin-Rezeptoren vorwiegend bei psychiatrischen Erkrankungen wie Schizophrenie oder Psychosen eingesetzt werden. Aber auch in der Neurologie werden sie als Antikonvulsiva und in der Gastroenterologie als Medikamente zur Regulation der Darmtätigkeit (Prokinetika) verwendet. Substanzen mit antagonistischer Wirkung gegen Dopamin-Rezeptoren sind sehr interessant und werden deswegen stark beforscht. Die hier vorliegenden Strukturen liefern dabei einige neue Ideen und sehr interessante Selektivitäten. Für eine weitere Optimierung scheint eine Anwendung computerbasierter Methoden sehr Erfolg versprechend und interessant.

Bis dato gab es zu diesem Datensatz keinerlei theoretische Untersuchungen, da sich aufgrund der unterschiedlichen Ringgrößen und deren Flexibilität ein Alignment und eine Auswahl des bioaktiven Konformers für 3D-QSAR-Techniken extrem schwierig gestaltet. Im Rahmen dieser Arbeit wurde auch ein kurzer Versuch zur Erstellung eines dreidimensionalen Modells unternommen, aber aufgrund der vorgenannten Gründe bald aufgegeben. Die Moleküle wurden als maschinenlesbare 2D-Strukturen von Christoph Enzensperger von der Universität Jena zur Verfügung gestellt. Sie wurden mittels des Accelrys DS Viewer Pro[221] zu dreidimensionalen Strukturen konvertiert und dann direkt in Catalyst eingelesen. Dort wurden

entsprechende Konformerensembles erzeugt und als Ausgangspunkt für eine xMaP-Analyse verwendet.

Zusätzlich wurde bei diesem Datensatz ausgehend von einem Homologiemodell des Dopamin D₁-Rezeptors ein strukturbasiertes QSAR-Modell erstellt. Dabei wurden die oben erzeugten Konformere mittels der Software FRED [222,223] in das Modell gedockt. Genauere Details dazu sind an der entsprechenden Stelle im Ergebnisteil aufgeführt. Außerdem wurden anhand der Daten für die weiteren Dopaminrezeptoren (D₂ bis D₅) QSAR-Modelle erstellt. Die Strukturen sämtlicher untersuchter Substanzen in Verbindung mit ihren biologischen Aktivitäten sind im Anhang in Kapitel II.5. zu sehen.

3.7.6. Glukokortikoide (GK)

Diese Substanzen werden aktuell in der Gruppe von Frau Professor Högger am Institut für Pharmazie der Universität Würzburg untersucht [224-226]. Konkret basieren die hier verwendeten biologischen Aktivitäten auf Werten aus der Dissertation von Anagnostis Valotis [227]. Der Datensatz besteht aus insgesamt 31 Glukokortikoiden, die im Anhang in Kapitel II.6. gezeigt sind. Von 30 dieser Strukturen ist die relative Rezeptoraffinität (RRA) im Vergleich zum Dexamethason (Abb. 31) am Glukokortikoidrezeptor bestimmt worden. Der dekadische Logarithmus dieser Werte diente als abhängige Variable in den QSAR-Analysen. Glukokortikoid-Präparate, etwa Cortison, haben entzündungshemmende Wirkung. Sie werden zur Therapie u.a. sowohl bei allergischem Schnupfen als auch bei Asthma bronchiale verwendet. Hochwirksame Varianten werden bei akuten Notfällen (Anaphylaxie, Sepsis, Schock) eingesetzt [228,229].

Die Struktur-Wirkungsbeziehungen der Glukokortikoide sind gut untersucht (siehe Abb. 31): Die Ketogruppe an Kohlenstoff 3, die Hydroxylierung an C11 und C17 sowie die Doppelbindung zwischen C4 und C5 im Ring A sind entscheidend für die Aktivität als Glukokortikoid. Da es für den vorliegenden Datensatz bisher keine QSAR-Untersuchungen gibt, wurden neben dem xMaP-Modell auch CoMFA- und CoMSIA-Modelle erstellt.

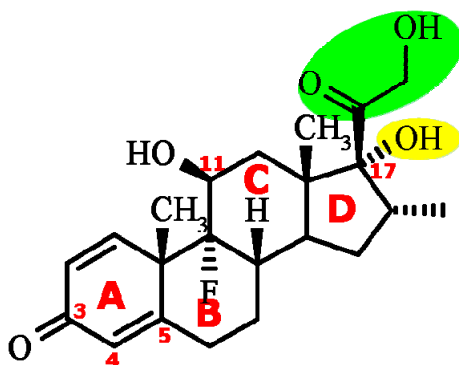


Abbildung 31: Die Struktur des Dexamethasons. Alle Rezeptoraffinitäten wurden relativ zu der des Dexamethasons bestimmt. Im Folgenden wird folgende Nomenklatur verwendet: In grün ist die β -Position für Substituenten markiert, in gelb die α -Position.

Die Strukturen wurden in ISIS Draw [191] gezeichnet und anschließend unter Verwendung des Accelrys DS Viewer Pro [221] in eine dreidimensionale Darstellung konvertiert. Für die xMaP-Analyse wurden diese Strukturen in Catalyst eingelesen und dort die entsprechenden Strukturfamilien für jedes einzelne Molekül berechnet. Dies stellte die Basis für alle weiteren Berechnungen dar.

Für die CoMFA- und CoMSIA-Untersuchungen wurde zunächst die Struktur der Referenzsubstanz Dexamethason in Tripos SYBYL eingelesen. Dort erfolgte eine Energieminimierung mit den Standardparametern im Tripos-Kraftfeld [230]. Dieses Grundgerüst diente als Referenz für das Alignment. Dafür wurden zwei verschiedene Methoden angewendet: Einerseits wurde mit FlexS [231,232] eine automatische Überlagerung berechnet. Dafür wurden die Standardparameter dieser Software eingesetzt. Eine Überlagerung der Moleküle war relativ gut möglich.

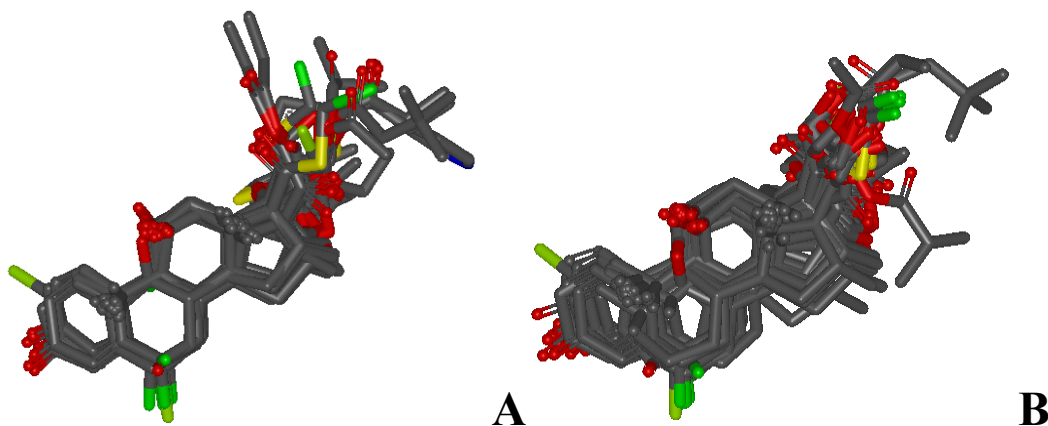


Abbildung 32: Die Überlagerungen der Strukturen im GC-Datensatz, links per Hand, rechts automatisiert per FlexS; Die deutlich höhere Homogenität des Hand-Alignments wird offensichtlich.

Andererseits wurde unter Zuhilfenahme der Multifit-Prozedur von SYBYL ein sehr zeitaufwändiges manuelles Alignment durchgeführt. Hier diente wiederum die Grundstruktur des Dexamethasons als Referenz auf die alle anderen Strukturen per Hand gelegt wurden. Diese beiden Überlagerungen sind in Abbildung 32 gezeigt. Die größere Homogenität des in Bildteil 41A gezeigten Hand-Alignments wird dabei offensichtlich. Mit diesen beiden Überlagerungen als Grundlage wurden jeweils eine CoMFA- und eine CoMSIA-Analyse durchgeführt. Die Ergebnisse und Interpretation dazu sind im Ergebnisteil im Vergleich mit den xMaP-Ergebnissen gezeigt. Alle 31 Strukturen sind zusammen mit ihren relativen Rezeptoraffinitäten im Anhang in Kapitel II.6. gezeigt.

3.7.7. Naphtylisochinolin-Alkaloide (NIQ)

Die Substanzen dieser Klasse werden seit vielen Jahren im Arbeitskreis von Professor Bringmann an der Universität Würzburg untersucht. Es gibt zu einem Teil dieser Moleküle, die Aktivität gegen den Malariaerreger *Plasmodium falciparum* zeigen, erfolgreiche 3D-QSAR-Untersuchungen [81,233]. Für genau diesen Teil dieser Strukturen wurde auch ein 4D-QSAR-Modell erstellt. Dafür wurden mittels Catalyst Konformerensembles erzeugt. Aufgrund von experimentellen Daten wurden aus jedem Konformerensembles Rotamere entfernt, die *in-vitro* nicht vorliegen. Die resultierende Konformerensembles wurde als Ausgangspunkt für die xMaP-Analyse verwendet.

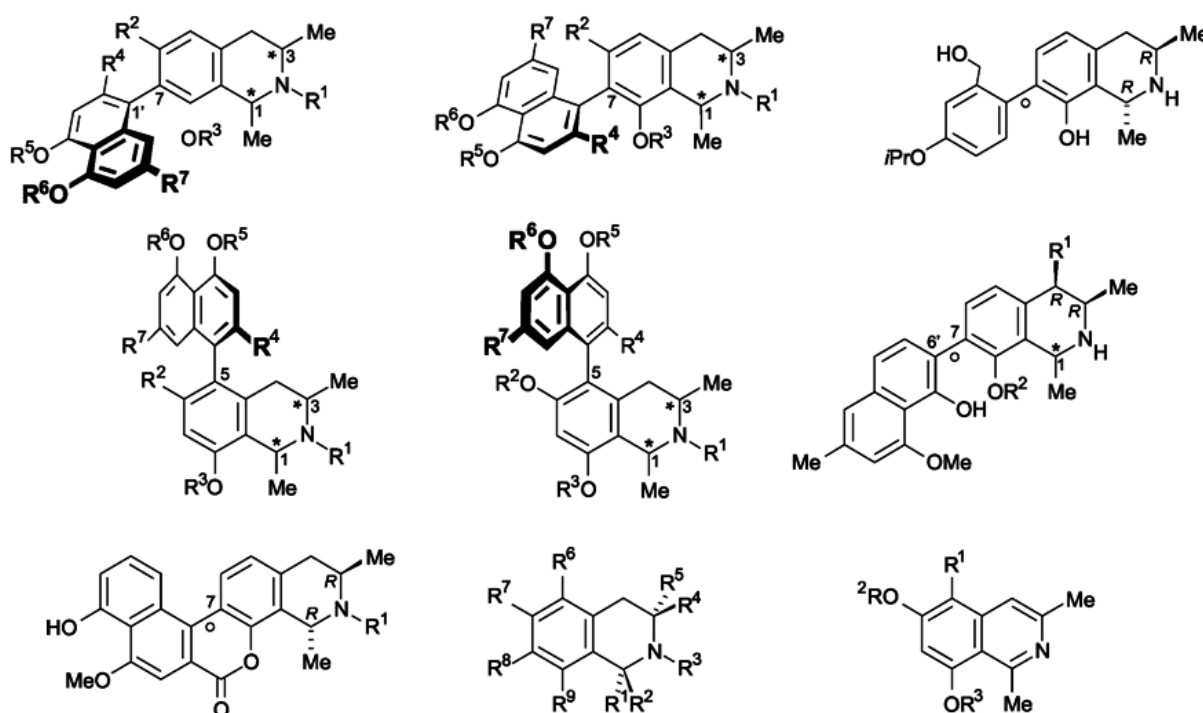


Abbildung 33: Grundgerüste der Verbindungen im NIQ-Datensatz

Außerdem wurden noch weitere Untersuchungen bezüglich Aktivitäten an weiteren Krankheitserregern durchgeführt. Dazu wurden für neu synthetisierte Substanzen Deskriptoren errechnet und weiter verarbeitet. Die Grundstrukturen sind in Abbildung 33 zu sehen, alle Strukturen im Überblick finden sich im Anhang dieser Arbeit in Kapitel II.7.

3.7.8. Weitere Datensätze

Um die neu entwickelte Methode noch eingehender zu validieren wurden verschiedene Standarddatensätze der QSAR herangezogen. Damit sollte ein Überblick darüber erhalten werden, wie breit xMaP auf Datensätze mit unterschiedlicher Zusammensetzung angewendet werden kann. Diese Datensätze werden nicht im Detail diskutiert, die statistischen Ergebnisse aber kurz präsentiert.

Der Standarddatensatz der QSAR-Methodenvalidierung wurde untersucht: Der Steroid-Datensatz, der von der Gruppe von Johann Gasteiger zur Verfügung gestellt wird [34,234]. Dann wurden für den ECETOC [235]- und den ET_A-Datensatz [236-238] Modelle erzeugt. Ferner wurden xMaP-Modelle für einen Datensatz von Artemisinin-Analoga [239] berechnet. Es wurden jeweils aus früheren Publikationen vorhandene Strukturen in Catalyst eingelesen. Damit wurde mit den Standardparametern ein Konformerensemble für jedes Molekül erzeugt. Diese Konformerensembles dienten als Basis für die jeweilige Modellbildung. Zusätzlich wurde mit xMaP noch ein Datensatz [240,241] von Inhibitoren des Transkriptionsfaktors NF- κ B beforscht, der von Frau Professor Merfort von der Universität Freiburg zur Verfügung gestellt wurde. Auch hier wurden zur Modellbildung die Standardparameter verwendet. Zusätzlich wurden Trainings- und Testdatensatz aus den bisherigen Publikationen verschmolzen, da die Standard-xMaP-Validierung zum Einsatz kam.

3.8. Standardparameter und Parametervariationen

Wie bereits in dieser Arbeit angesprochen wurde, ist es wichtig, dass der Nutzereinfluss auf die Modellbildung so gering wie nur irgendwie möglich gehalten wird. Dafür ist es bei der Veränderung gewisser Parameter wichtig, dass der Einfluss auf die Modellbildung möglichst gering ist. Außerdem muss bekannt sein, welche Parameter am Besten konstant gehalten

werden. Um dies für die xMaP-Technik zu überprüfen, wurden verschiedene Parameter variiert, um daraus ein Standardprotokoll zu erarbeiten. Dieses Standardprotokoll soll tolerant gegenüber kleinen Änderungen im Versuchsdesign sein. Die Resultate dieser Untersuchungen werden im Folgenden diskutiert. Am Schluss steht ein Protokoll, das bei allen untersuchten Datensätzen in identischer Form eingesetzt wurde. Folgende Datensätze wurden im Hinblick auf Parameterevaluation untersucht, in Klammern sind die verwendeten Kürzel gezeigt: M₂-Modulatoren (M₂), Endothelin_A-Inhibitoren (ET_A), Inhibitoren der Reversen Transkriptase von HIV-1 (HEPT), Analoga des Prostaglandins F_{2α} (PGF_{2α}), Inhibitoren der Acetylcholinesterase (AZT) sowie Antagonisten des Dopamin-D₁-Rezeptors (D₁). Diese Datensätze werden ab Kapitel 3.7. detailliert vorgestellt.

Zu den hier gezeigten Ergebnissen ist folgendes anzumerken: Bei Modellen, die sich in nur einem Parameter unterscheiden, wurden die Unterschiede in den Vorhersagen auf ihre statistische Signifikanz hin evaluiert. Dafür wurde die Methode von van der Voet mit einer Irrtumswahrscheinlichkeit von 5% genutzt [242]. Wenn sich Ergebnisse signifikant unterscheiden, dann sind die Ergebnisse durch einen hochgestellten Stern „*“ am Datensatznamen markiert. Im Text ansonsten angesprochene Unterschiede sind zwar statistisch nicht signifikant, trotzdem aber praktisch relevant.

3.8.1. Einfluss der Konformerberechnung

Die QSAR ist eine so genannte ligandbasierte Methode und wird üblicherweise dann eingesetzt, wenn keine Informationen über die Zielstruktur zur Verfügung stehen. Folglich wurden Algorithmen zur Konformerbestimmung evaluiert, die ohne Informationen über die Zielstruktur auskommen. Der Einfluss einer strukturbasierten Konformationssuche wird an anderer Stelle diskutiert. Bekannte Algorithmen zur Konformerengeneration sind Accelrys Catalyst [150] und Openeye Omega [151,195,196]. Für Catalyst wurde vor kurzem gezeigt, dass es in einem definierten Energiefenster, welches von der inneren Energie des energieärmsten Konformers bis 20 kcal/mol darüber reicht, die tatsächliche bioaktive Konformation sehr gut berechnen kann [156]. Das heißt, dass das generierte Konformerensemble mit hoher Wahrscheinlichkeit auch das bioaktive Konformer beinhaltet. Wie in Kapitel 3.3.1. beschrieben wurde, wurde im Rahmen dieser Arbeit ausschließlich der BEST-Algorithmus von Catalyst zur Konformationssuche verwendet. In der Folgepublikation zu [156] wurden Catalyst und Omega ausführlich verglichen und auf Parametervariationen hin evaluiert [152].

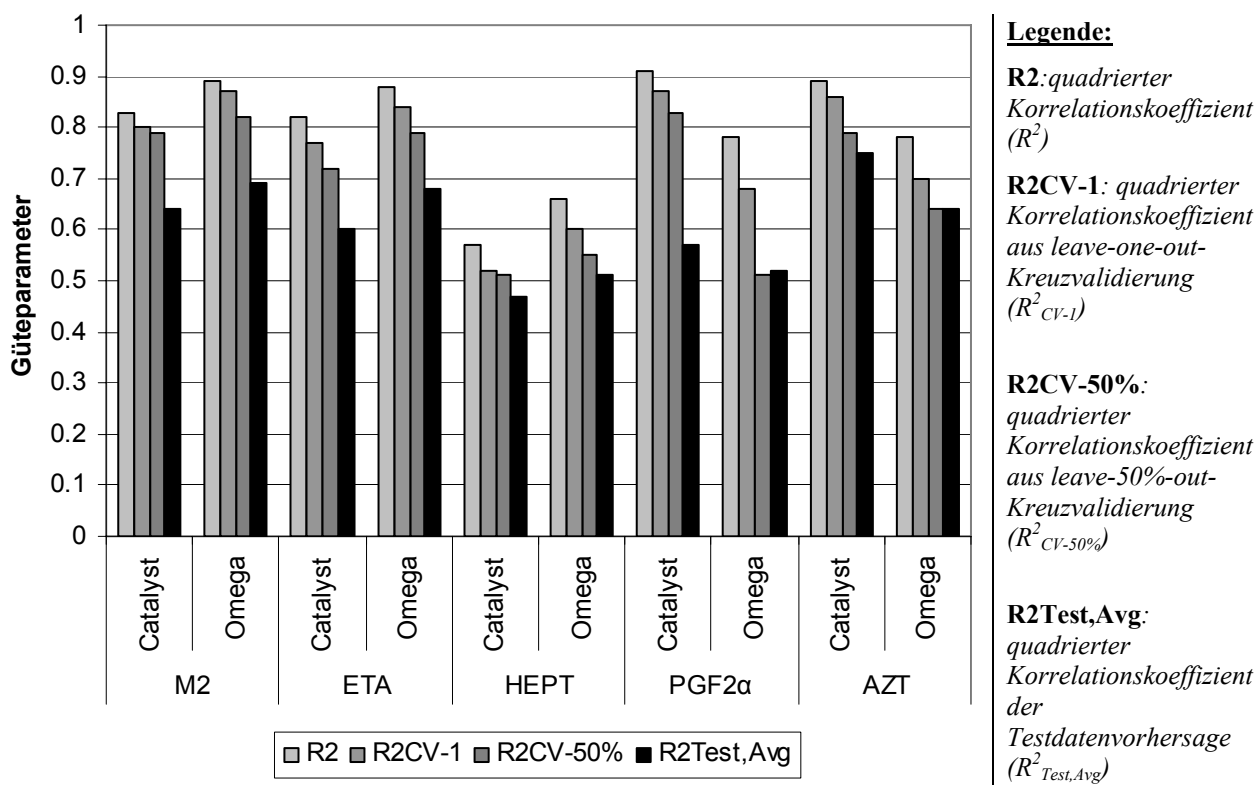


Abbildung 34: Balkendiagramm zur Illustration des Einflusses verschiedener Konformationssuche-Algorithmen auf die statistische Qualität der mit dieser Grundlage erstellten xMaP-Modelle (Die verwendeten Werte sind in Tabelle 3 gezeigt).

Diese Ergebnisse decken sich mit den im Rahmen dieser Arbeit gemachten Erfahrungen. Für insgesamt fünf Datensätze – M₂, ET_A, HEPT, PGF_{2α} sowie AZT – wurden bei allen Molekülen mit den Standardparametern von Catalyst und Omega Konformerensemble generiert. Mit diesen Daten wurden zwei verschiedene Modelle errechnet, die sich lediglich in diesem Parameter der Konformationssuche unterschieden. Alle anderen Schritte der Deskriptorberechnung und Modellbildung waren identisch. Abbildung 34 und Tabelle 3 zeigen einen Überblick über die statistische Qualität der resultierenden Modelle.

Tabelle 3: Überblick über die Evaluation des Einflusses verschiedener Konformationssuchealgorithmen auf die statistische Qualität der damit erzeugten xMaP-Modelle.

Datensatz	Methode der Konformerberechnung	R^2	R^2_{CV-1}	$R^2_{CV-50\%}$	$R^2_{Test,Avg}$
M ₂ *	Catalyst	0.83	0.80	0.79	0.64
	Omega	0.89	0.87	0.82	0.69
ET _A	Catalyst	0.82	0.77	0.72	0.60
	Omega	0.88	0.84	0.79	0.68
HEPT	Catalyst	0.57	0.52	0.51	0.47
	Omega	0.66	0.60	0.55	0.51
PGF _{2α}	Catalyst	0.91	0.87	0.83	0.57
	Omega	0.78	0.68	0.51	0.52
AZT	Catalyst	0.89	0.86	0.79	0.75
	Omega	0.78	0.70	0.64	0.64

Es zeigt sich, dass die berechneten Deskriptoren sehr ähnlich sind und deshalb Modelle mit sehr ähnlicher statistischer Qualität liefern. Der statistisch signifikante Unterschied in den Modellen für den M₂-Datensatz liegt in der Symmetrie der Moleküle begründet. Omega diskriminiert hier zusätzlich bei der Bestimmung der Konformerähnlichkeit. Die Konformerensembles werden daher diverser als bei Catalyst, das symmetrische Moleküle nicht zusätzlich diskriminiert. Dieser Unterschied hat jedoch keine praktische Relevanz.

Für zwei Moleküle aus dem AZT-Datensatz sind die verschiedenen Deskriptoren in Abbildung 35 gezeigt. Damit wird der Einfluss der Konformationssuche illustriert.

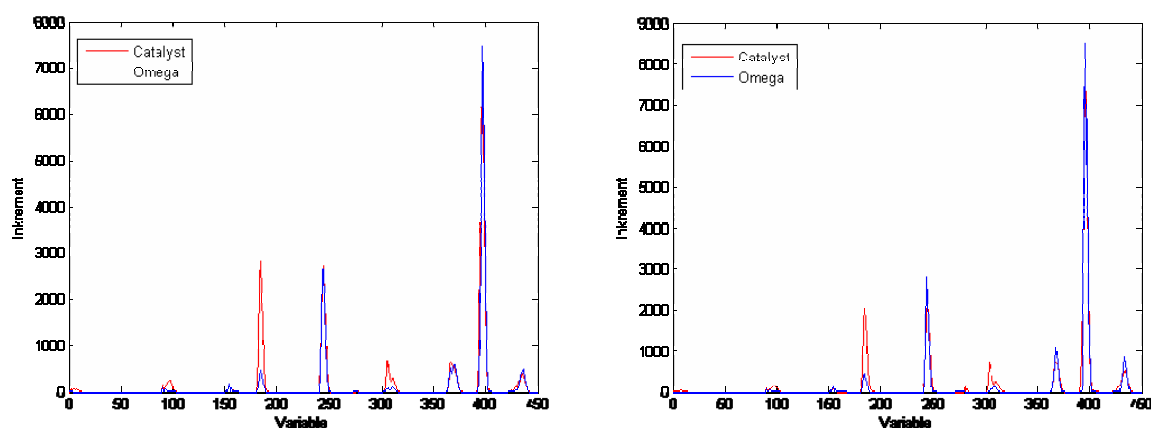


Abbildung 35: Die auf Catalyst(rot) und Omega(blau) basierenden xMaP-Deskriptoren für die Moleküle 1 und 2 des AZT-Datensatzes.

In Abbildung 35 zeigt sich, dass der Einfluss der verschiedenen Programme nicht sehr groß ist. Deshalb tragen die Deskriptoren im Wesentlichen die gleiche Information. Einzelne Variablen sind zwar stärker inkrementiert, die grundsätzliche Verteilung bleibt aber gleich. Daraus kann man schließen, dass die xMaP-Deskriptoren sehr robust im Hinblick auf die verwendete Konformationssuche sind, da die Ergebnisse nur leicht beeinflusst werden. Anwender, denen die hier genannten Programme nicht zur Verfügung stehen, können folglich auf andere Produkte ausweichen. Im Rahmen dieser Arbeit wurde durchwegs Catalyst verwendet, da es während der gesamten Zeit der Anfertigung dieser Dissertation zur Verfügung stand.

3.8.2. Einfluss des „Energiefensters“

Ein sehr wichtiger zu überprüfender Faktor ist der Einfluss eines auf unterschiedlichen Maximalenergien beruhenden Konformerensembles auf die Deskriptorberechnung. Dies lässt sich gut anhand von Catalyst evaluieren, da der Schwellenwert für die Obergrenze der

zulässigen inneren Energie einfach gesetzt werden kann. Diese Obergrenze beeinflusst die Konformerberechnung wie folgt: Es wird versucht, möglichst viele diverse Konformere zu generieren, die den kompletten Bereich des zugestandenen Energiefensters abdecken.

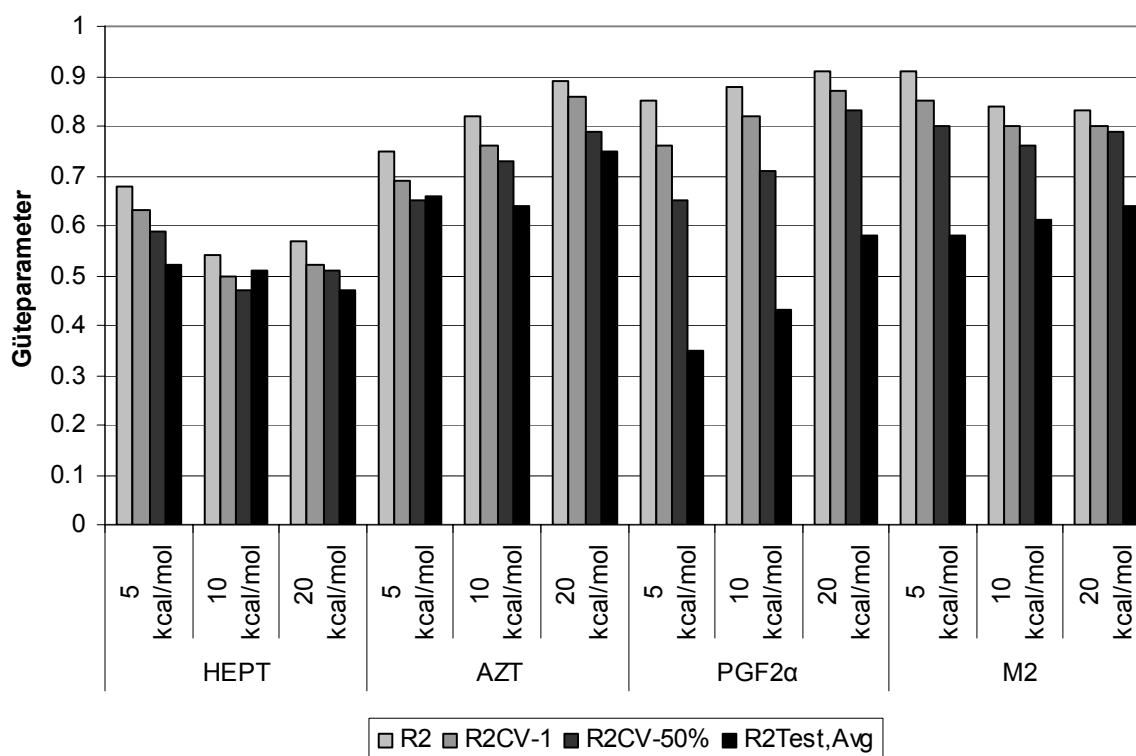


Abbildung 36: Balkendiagramm zur Illustration des Einflusses verschiedener Energiefenster in der Konformationssuche auf die statistische Qualität der damit erstellten xMaP-Modelle (Die verwendeten Werte sind in Tabelle 4 gezeigt, eine Legende findet sich bei Abbildung 34).

Für vier Datensätze - HEPT, AZT, PGF_{2α} und M₂ - wurde getestet, welchen Einfluss eine Veränderung des zugestandenen Energiefensters auf die Qualität und Aussage der Modelle hat. Mit Obergrenzen von 5, 10 und 20 kcal/mol für die maximale zusätzliche innere Energie der Konformere über dem energieärmsten Konformer wurden Konformerensembles für alle Moleküle aus den Datensätzen errechnet. Ausgehend davon wurden xMaP-Deskriptoren errechnet und QSAR-Modelle erstellt. Die gefundenen statistischen Güteparameter sind in Tabelle 4 und Abbildung 36 gezeigt. Es fällt auf, dass jeweils zu niedrigerer Energie hin beim HEPT-Datensatz (eher starre Grundgerüste) die internen Güteparameter (R^2 und R^2_{CV-1}) signifikant besser werden. Der Einfluss auf die externe Vorhersagekraft ist jedoch frappierend gering. Dies lässt sich sehr gut erklären: Je kleiner das Energiefenster wird, desto homogener werden aufgrund der relativ niedrigen Molekülflexibilität die berechneten Konformerensembles.

Tabelle 4: Überblick über die Evaluation des Einflusses verschiedener Obergrenzen für die maximale zusätzliche innere Energie von Konformeren auf die statistische Qualität der damit erzeugten xMaP-Modelle

Datensatz	Energiefenster (kcal/mol)	R^2	R_{CV-1}^2	$R_{CV-50\%}^2$	$R_{Test,Avg}^2$
HEPT	5	0.68	0.63	0.59	0.52
	10	0.54	0.50	0.47	0.51
	20	0.57	0.52	0.51	0.47
AZT	5	0.75	0.69	0.65	0.66
	10	0.82	0.76	0.73	0.64
	20	0.89	0.86	0.79	0.75
PGF _{2α}	5	0.85	0.76	0.65	0.35
	10	0.88	0.82	0.71	0.43
	20	0.91	0.87	0.83	0.58
M ₂	5	0.91	0.85	0.80	0.58
	10	0.84	0.80	0.76	0.61
	20	0.83	0.80	0.79	0.64

Dies ist in Abbildung 36 zu sehen. Dort sind die Konformerensembles bei verschiedenen Größen des Energiefensters exemplarisch für eine Struktur aus dem AZT-Datensatz gezeigt. Dieser Effekt ist bei den starrereren Strukturen im HEPT-Datensatz noch deutlicher ausgeprägt (nicht gezeigt). Betrachtet man nun die immer größer werdende Ähnlichkeit, so ist klar, dass im Deskriptor weniger Information kodiert wird. Es werden weniger diverse Konformere modelliert. Dadurch wird die Information schärfer aber auch weniger. Es kann aufgrund der hohen Ähnlichkeit der Konformere untereinander und auch unter den verschiedenen untersuchten Molekülen ein sehr gutes internes Modell herausgearbeitet werden. Da die Deskriptoren aber weniger Gesamtinformation tragen ist ihre externe Vorhersagekraft geringer. Der Grund dafür ist, dass sehr viel schwerer auf leicht abweichende Moleküle hin extrapoliert werden kann. Dies fällt noch deutlicher bei den hochflexiblen Molekülen aus den AZT-, PGF_{2 α} - und M₂-Datensätzen auf. Dort werden sogar die internen Modelle schlechter. Die externe Vorhersagekraft ist jeweils bei einem Energiefenster von 20 kcal/mol am höchsten. Demzufolge ist hier der Informationsverlust bei kleineren Energiefenstern im Deskriptor so groß, dass nur mehr schlechtere interne Modelle erstellt werden können. Der Grund dafür ist die noch niedrigere Ähnlichkeit der einzelnen Konformerensembles untereinander. Erst bei einer Obergrenze von 20 kcal/mol tragen die Deskriptoren genügend Information, so dass auch extern gute Vorhersagen getroffen werden können.

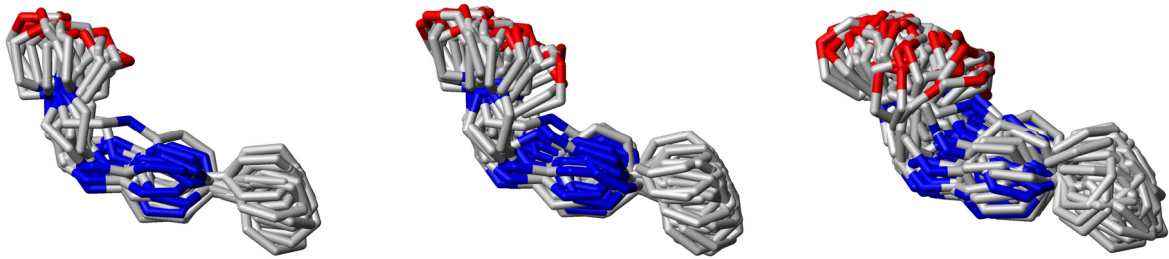


Abbildung 37: Exemplarisch für alle Moleküle aus beiden Datensätzen: Strukturfamilien für Struktur 10 aus dem AZT-Datensatz bei verschiedenen Größen des Energiefensters (links: 5 kcal/mol, 18 Konformere, RMSD=1.157 Å; Mitte: 10 kcal/mol, 33 Konformere, RMSD=1.252 Å; rechts: 20 kcal/mol, 57 Konformere, RMSD=1.437 Å)

Dieses Ergebnis kann wie folgt zusammengefasst werden: Wünschenswert ist eine möglichst hohe externe Vorhersagekraft der Modelle. Diese kann am Besten dadurch erreicht werden, dass eine möglichst diverse Konformerensemble pro Molekül eingesetzt wird. Das bedeutet, dass als Standardparameter eine Obergrenze von 20 kcal/mol angemessen ist. Hierbei kann davon ausgegangen werden, dass das bioaktive Konformer und ihm sehr ähnliche Konformere Eingang in die Deskriptorberechnung finden. Bei kleineren Energiefenstern wäre das nicht zwangsläufig der Fall [156]. Bei allen anderen Validierschritten kam folglich diese Obergrenze zum Einsatz. Für zukünftige Anwendungen sollte ein möglichst hoher aber noch sinnvoller Schwellenwert für die Obergrenze der inneren Energie zum Einsatz kommen.

3.8.3. Einfluss der Konformerenwichtung

Zur Berechnung dieser Modelle wurden die Omega-Konformer verwendet, da hier die (geschätzte) Energie einzelner berechneter Konformere im Gegensatz zu Catalyst sehr leicht zugänglich ist. Es wurden drei Datensätze untersucht: AZT, M₂ und PGF_{2α}.

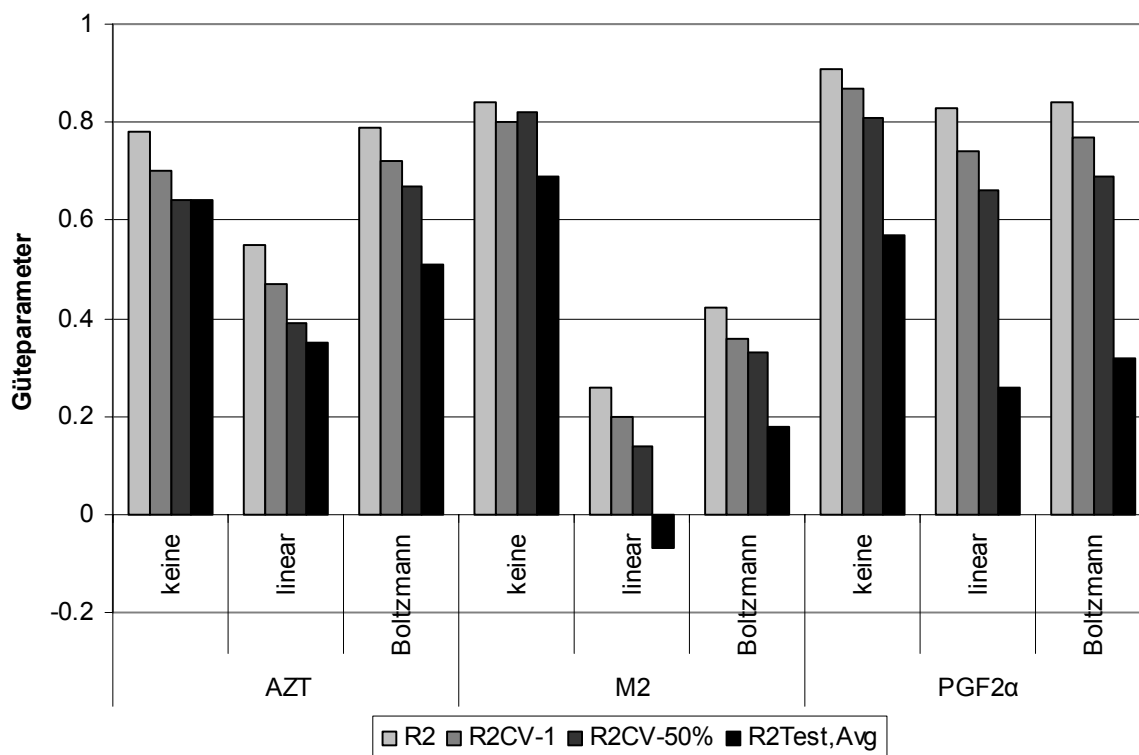


Abbildung 38: Balkendiagramm zur Illustration des Einflusses verschiedener Methoden der Konformerenwichtung auf die statistische Qualität der darauf basierenden xMaP-Modelle (Die verwendeten Werte sind in Tabelle 5 gezeigt, eine Legende findet sich bei Abbildung 34).

Ziel dieser Validierung war die Identifizierung des Einflusses einer energiebasierten Wichtung auf die statistische Qualität der erstellten Modelle. Konformere liegen in Abhängigkeit von ihrer Energie mit unterschiedlicher Wahrscheinlichkeit vor. Anhand dieser Wahrscheinlichkeit kann ein Wichtungsfaktor in der Durchschnittsbildung über alle Konformere berücksichtigt werden. Einzelne Konformere werden dadurch entsprechend ihrer Energie gewichtet. Der Durchschnittswert eines einzelnen Zweipunkt-Pharmakophors über alle Konformere wird dabei nach folgender Formel berechnet:

$$W_{ges} = \frac{\sum_{i=1}^{n_{conformers}} (w_i \cdot W_i)}{\sum_{i=1}^{n_{conformers}} w_i} \quad \text{Gleichung 32}$$

Dabei ist W_{ges} ein einzelner xMaP-Deskriptor für ein komplettes Konformerensemble, $n_{conformers}$ die Anzahl der Konformere bei dem untersuchten Molekül. W_i ist der Deskriptor für ein einzelnes Konformer und w_i sein Wichtungsfaktor, mit dem er in den Gesamtdeskriptor W_{ges} einfließt.

Tabelle 5: Überblick über die Evaluation des Einflusses verschiedener Methoden der Konformerewichtung auf die statistische Qualität der damit erzeugten xMaP-Modelle.

Datensatz	Methode der Wichtung	R^2	R_{CV-1}^2	$R_{CV-50\%}^2$	$R_{Test,Avg}^2$
AZT*	keine	0.78	0.70	0.64	0.64
	linear	0.55	0.47	0.39	0.35
	Boltzmann	0.79	0.72	0.67	0.51
M ₂ *	keine	0.84	0.80	0.82	0.69
	linear	0.26	0.20	0.14	-0.07
	Boltzmann	0.42	0.36	0.33	0.18
PGF _{2α} *	keine	0.91	0.87	0.81	0.57
	linear	0.83	0.74	0.66	0.26
	Boltzmann	0.84	0.77	0.69	0.32

Für den Fall, dass keine Wichtung erfolgt und alle Konformere als gleich wichtig gelten, ist w_i als 1 anzunehmen. Dies ist der Wert, der bei allen anderen Validierungen eingesetzt wurde. Es wurden zwei weitere Arten der Wichtung getestet. In einem Ansatz wurde der Wichtungsfaktor w_i so berechnet, dass die Konformere im Verhältnis zu dem mit der niedrigsten Energie linear gewichtet werden. Je energiereicher ein Konformer ist, desto niedriger wird es gewichtet, der Wichtungsfaktor wird wie folgt berechnet:

$$w_{linear} = \frac{E_0}{E_i} \quad \text{Gleichung 33}$$

Dabei steht E_0 für die Energie des energieärmsten und E_i für die des zu untersuchenden Konformers. Noch extremer in diese Richtung geht die Wichtung nach Boltzmann [243]:

$$w_{Boltzmann} = e^{-E_i/RT} \quad \text{Gleichung 34}$$

Dabei ist E_i identisch zu oben. R ist die allgemeine Gaskonstante mit einem Wert von $R=8.31 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$. Der Wert von T beschreibt die Temperatur im Experiment in Kelvin. Bei den Untersuchungen hier wurde eine Raumtemperatur von $T=298 \text{ K}$ ($25 \text{ }^\circ\text{C}$) angenommen.

Betrachtet man die statistische Qualität der in Tabelle 5 und Abbildung 38 gezeigten Modelle, so wird folgendes deutlich: Die externe Vorhersagekraft der Modelle ist am höchsten, wenn alle Konformere gleich gewichtet werden. Am schlechtesten ist die externe Vorhersagekraft bei linearer Wichtung. Ähnliche Zusammenhänge ergeben sich für die interne Modellgüte. Der Grund dafür ist die unterschiedliche Konformation des jeweils energieärmsten

Konformers, das über die Boltzmann- oder lineare Wichtung am höchsten gewichtet wird. Dadurch beschreibt der Deskriptor faktisch fast nur dieses eine Konformer für das zu untersuchende Molekül. Der 4D-Deskriptor verliert so viel Information, dass es faktisch wieder ein 3D-Deskriptor ist. Aus diesen Deskriptoren können damit die Zusammenhänge deutlich schlechter herausgearbeitet werden, da wesentliche Informationen über die Flexibilität vernachlässigt werden. Auch sind die hochgewichteten Konformere meist untereinander sehr divers, was die QSAR-Analyse deutlich erschwert. Ein weiteres Argument gegen eine energieabhängige Wichtung ist die Tatsache, dass unterschiedliche Konformationssuchalgorithmen auch unterschiedliche energieärmste Konformere generieren. Da bei der Durchführung von QSAR-Analysen eine möglichst hohe externe Vorhersagekraft der Modelle erwünscht ist, sollte bei künftigen Anwendungen keine Wichtung durchgeführt werden. Dies beschleunigt außerdem die Rechenzeit. Bei allen anderen Validierschritten im Rahmen dieser Arbeit wurde daher keine Wichtung verwendet.

3.8.4. Einfluss der strukturbasierten Konformerberechnung

Wie oben erwähnt führt eine Berechnung mit einer Proteinstruktur als Ausgangspunkt zu deutlich homogeneren Konformerensembles. Dies zeigt sich auch in den darauf basierenden Modellen.

Tabelle 6: Überblick über die Evaluation des Einflusses einer strukturbasierten Konformationssuche auf die statistische Qualität der damit erzeugten xMaP-Modelle.

Datensatz	Methode der Konformerberechnung	R^2	R_{CV-1}^2	$R_{CV-50\%}^2$	$R_{Test,Avg}^2$
HEPT	ligandbasiert	0.57	0.52	0.51	0.47
	strukturbasiert	0.75	0.69	0.65	0.52
AZT	ligandbasiert	0.89	0.86	0.79	0.75
	strukturbasiert	0.89	0.85	0.83	0.58
D ₁ *	ligandbasiert	0.77	0.71	0.64	0.47
	strukturbasiert	0.48	0.42	0.38	0.28

Dieser Parameter hat nicht nur auf die statistische Modellgüte massiven Einfluss. Die ausgewählten Variablen werden zudem erheblich beeinflusst. Deshalb wird die Auswirkung des Dockings als Konformerengenerator bei den entsprechenden Datensätzen diskutiert. Für den HEPT-Datensatz fällt auf, dass die internen strukturbasierten Modelle viel besser sind. Der Grund hierfür ist der Gleiche wie bei einem reduzierten Energiefenster. Es werden homogenere Konformere generiert, was dazu führt, dass diese untereinander relativ ähnlich sind. Interessanterweise wird hier auch die externe Vorhersagekraft geringfügig besser. Das ist darauf zurückzuführen, dass die Bindetasche alle Moleküle in eine ähnliche Konformation

drängt. Im Gegensatz zu einem reduzierten Energiefenster ist auch bei Extrapolationen eine hohe Ähnlichkeit der Konformationen der Testdatenmoleküle zu denen des Trainingsdatensatzes gegeben. Beim AZT-Datensatz dagegen führt dieses in-eine-Konformation-drängen intern zu identischen Modellqualitäten. Extern dagegen ist im strukturbasierten Modell eine deutlich niedrigere Modellqualität zu verzeichnen. Der Grund dafür dürfte sein, dass die homogeneren Konformere die Information auf nur wenige eindeutige für die biologische Aktivität wichtige Parameter fokussieren. Diese können weniger von der Gesamt-Variabilität in den Daten erklären, beschreiben aber einzelne wichtige Elemente gut. Außerdem ist von einem starken Einfluss durch das Docking auszugehen, da meist auch Moleküle die nicht optimal binden, trotzdem gut in die Bindetasche eingepasst werden. Wertvolle Information über mögliche schlechte Interaktionen geht dadurch verloren. Am gravierendsten ist die Tendenz zu statistisch schlechteren Modellen bei den D₁-Daten zu sehen. Dort wurde in ein Homologiemodell des Dopaminrezeptors gedockt, was den Verlust an Modellqualität erklären könnte. Der Konformationsraum wird hier nur grob eingeschränkt und auf eine Art und Weise, die nur näherungsweise mit den physiologischen Bedingungen übereinstimmen wird.

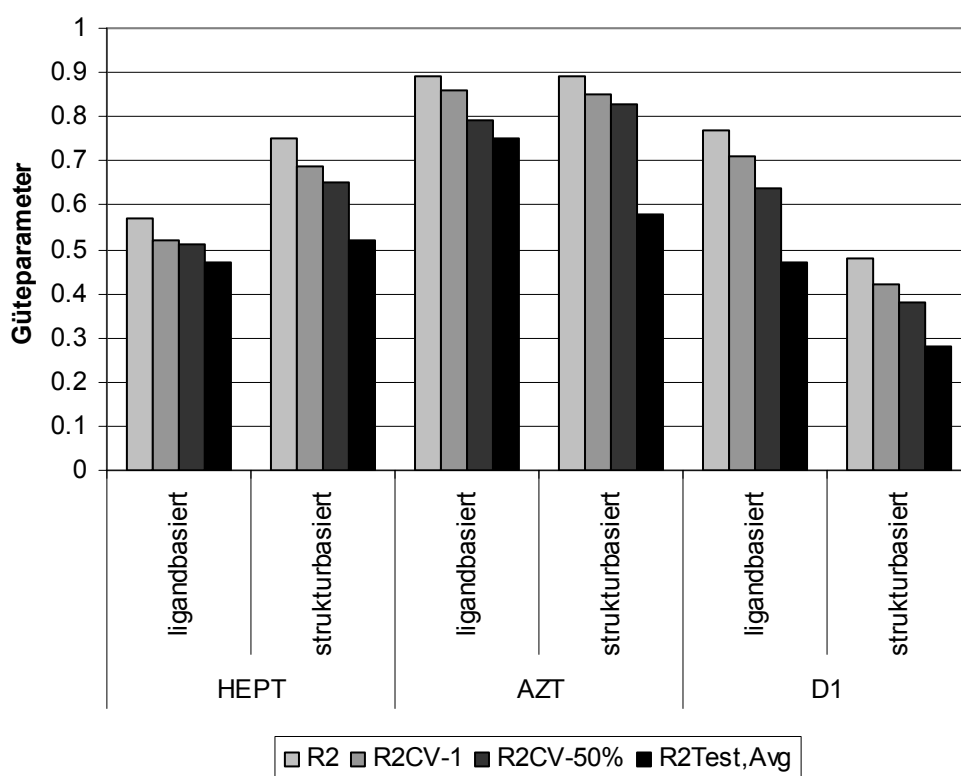


Abbildung 39: Balkendiagramm zur Illustration des Einflusses einer strukturbasierten Konformerberechnung auf die statistische Qualität der darauf basierenden xMaP-Modelle (Die verwendeten Werte sind in Tabelle 6 gezeigt, eine Legende findet sich bei Abbildung 35).

Nichtsdestotrotz bieten strukturbasierte Modelle meist große Vorteile in ihrer Interpretierbarkeit, da auch die Umgebung im Protein berücksichtigt werden kann. Vor diesem Hintergrund werden Veränderungen der statistischen Modellqualität meist sekundär. Es ist daher empfehlenswert, Strukturinformation zu nutzen, wann immer sie vorhanden ist. Dabei bietet sich auch ein Vergleich von ligand- und strukturbasiertem Modell für den zu untersuchenden Datensatz an, um den Einfluss der Strukturinformation abschätzen zu können. Bei allen drei Datensätzen ist die genaue Interpretation bei der Datensatzdiskussion in Kapitel 3.9. genauer erläutert.

3.8.5. Einfluss der Regressionstechnik

Der Standard in der QSAR ist die Verwendung der PLS, jedoch zeigt die PCR meist sehr ähnliche Resultate. Die im Rahmen dieser Arbeit verwendete Implementation der PCR ist schneller als die der PLS. Deshalb galt es herauszufinden, ob auch bei den xMaP-Deskriptoren der Einfluss der verwendeten Regressionstechnik gering ist. Tabelle 7 und Abbildung 40 zeigen den Einfluss der Regressionstechnik auf die statistische Modellgüte.

Tabelle 7: Überblick über die Evaluation des Einflusses verschiedener Regressionsmethoden auf die statistische Qualität der damit erzeugten xMaP-Modelle.

Datensatz	Regressionstechnik	R^2	R_{CV-1}^2	$R_{CV-50\%}^2$	$R_{Test,Avg}^2$
M ₂	PCR	0.83	0.80	0.79	0.64
	PLS	0.84	0.80	0.75	0.69
PGF _{2α}	PCR	0.91	0.87	0.83	0.57
	PLS	0.91	0.87	0.83	0.58
AZT	PCR	0.89	0.86	0.79	0.75
	PLS	0.89	0.86	0.79	0.75

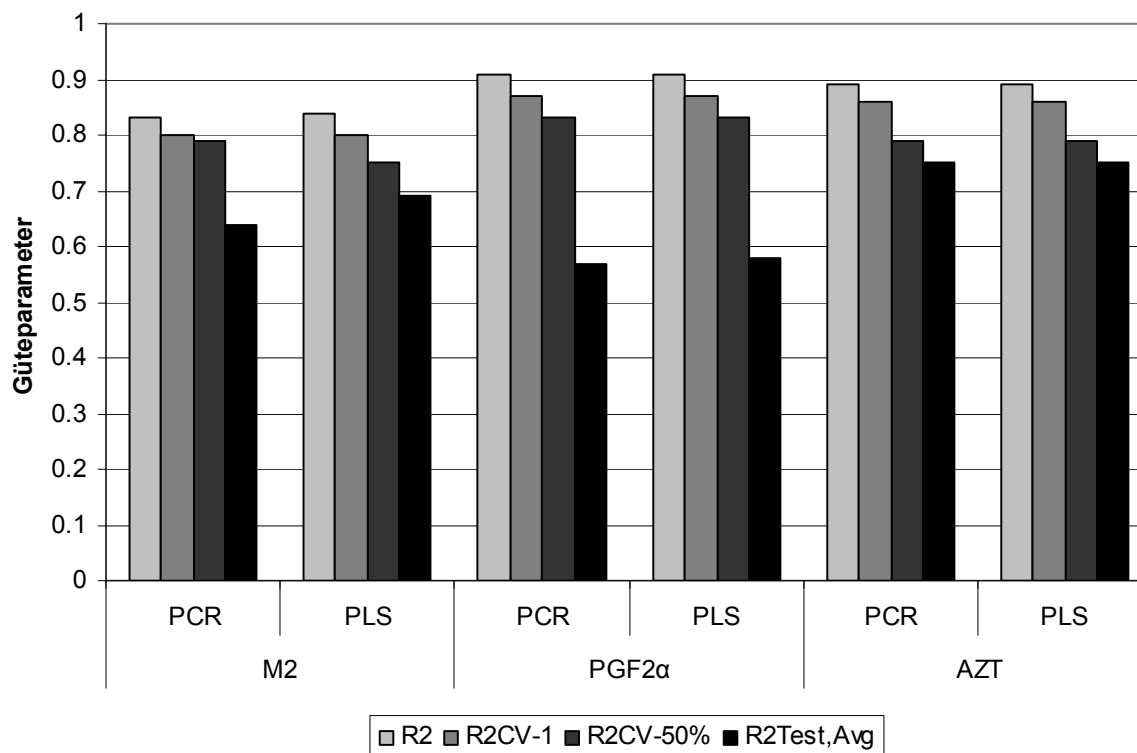


Abbildung 40: Balkendiagramm zur Illustration des Einflusses der verwendeten Regressionstechnik auf die statistische Qualität der damit erzeugten xMaP-Modelle (Die verwendeten Werte sind in Tabelle 7 gezeigt, eine Legende findet sich bei Abbildung 34).

Man erkennt, dass die Unterschiede gering sind. Deshalb können beide Techniken gleichermaßen verwendet werden. In dieser Arbeit wird mit PCR gearbeitet.

3.8.6. Einfluss der Variablenselektion

Das primäre Ziel der Variablenselektion bei QSAR-Modellen liegt darin, deren Ergebnisse leichter interpretier- und darstellbar zu machen. Die wichtigsten Variablen sollen herausgearbeitet und in den ursprünglichen Datenraum rückprojiziert werden. Es ist also von vorneherein zu erwarten, dass eine Verwendung dieser Techniken der Variablenselektion zu einer deutlichen Verbesserung der statistischen Modellqualität führt.

Tabelle 8: Überblick über die Evaluation des Einflusses der Variablenselektion mittels REM-TS auf die statistische Qualität der damit erzeugten xMaP-Modelle.

Datensatz	Variablenselektion	R^2	R^2_{CV-1}	$R^2_{CV-50\%}$	$R^2_{Test,Avg}$
M ₂ [*]	REM-TS	0.83	0.80	0.79	0.64
	Keine	0.76	0.66	0.57	0.59
PGF _{2α} [*]	REM-TS	0.91	0.87	0.83	0.57
	Keine	0.64	0.39	0.26	0.41
AZT [*]	REM-TS	0.89	0.86	0.79	0.75
	Keine	0.86	0.73	0.60	0.58

Dies ist hier wie erwartet und in Tabelle 8 und Abbildung 41 im Detail für drei Datensätze gezeigt. Alle statistischen Güteparameter verbessern sich durch die Anwendung der Variablenselektion. Deshalb ist auch der Unterschied in den verglichenen Modellen statistisch signifikant. Für künftige Anwendung wird die Verwendung der Tabusuche zur Variablenselektion angeraten. Dadurch wird die Interpretierbarkeit der Modelle gewährleistet.

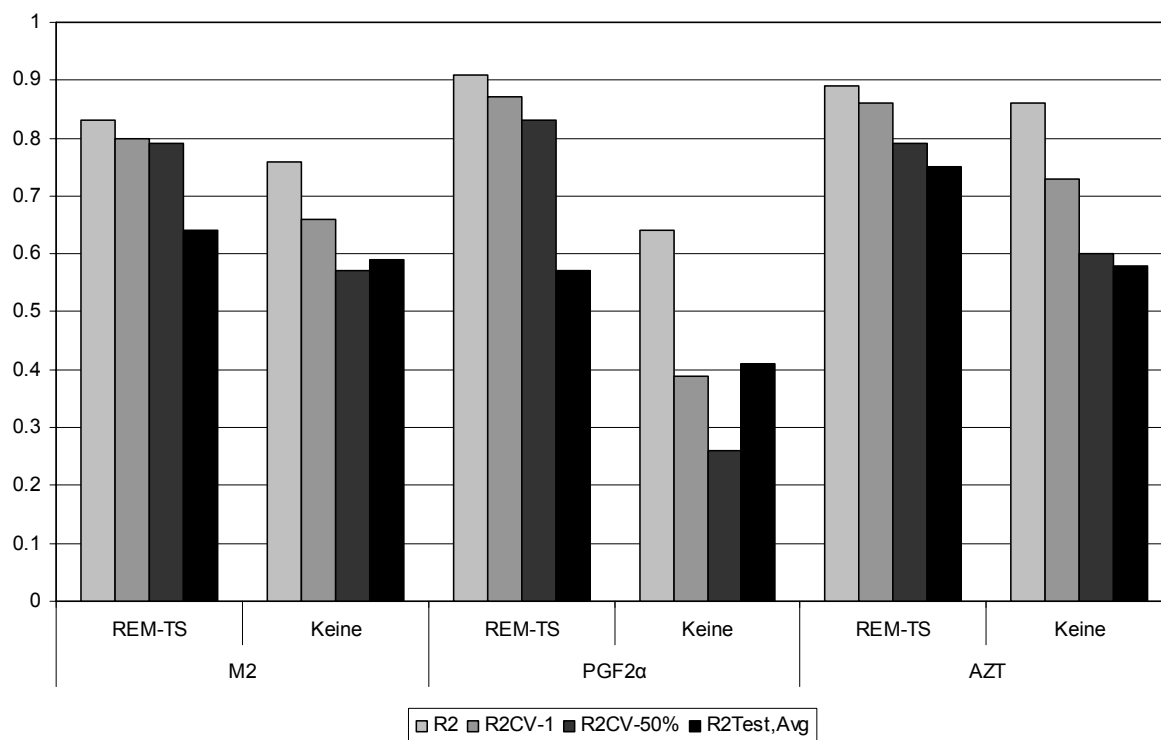


Abbildung 41: Balkendiagramm zur Illustration des Einflusses einer Variablenselektion auf die statistische Qualität der damit erzeugten xMaP-Modelle (Die verwendeten Werte sind in Tabelle 8 gezeigt, eine Legende findet sich bei Abbildung 34).

3.8.7. Ausgewählte Standardparameter

Für zukünftige Anwendungen der xMaP-Technik wird folgendes Vorgehen empfohlen: Zunächst muss anhand des Schemas in Abbildung 42 bestimmt werden, wie das Konformerensemble für jedes einzelne Molekül bestimmt wird. In vielen Fällen wird Information über die Zielstruktur zugänglich sein. Deshalb kann über eine unscharfe Definition der Bindetasche die konformelle Flexibilität eingeschränkt werden. Das sollte im Anschluss daran eine Modellinterpretation deutlich vereinfachen. Ausgehend von dem generierten Konformerensemble (Docking oder Konformationsuche) wird für jedes Konformer die molekulare Oberfläche mit einem Punktabstand von 0.4 Å mit den entsprechenden Eigenschaften berechnet.

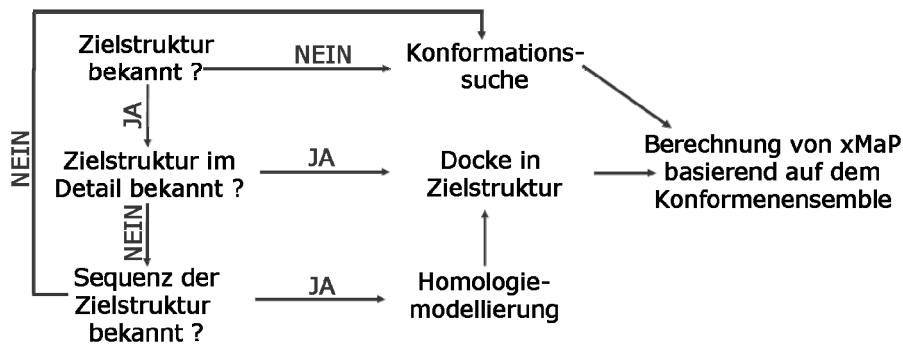


Abbildung 42: Die empfohlene Vorgehensweise, wenn eine xMaP-Analyse durchgeführt werden soll.

Für den Schwellenwert der inneren Energie in der Konformerberechnung kann problemlos ein ziemlich hoher Wert genommen werden. Ein Anhaltspunkt von bis zu 20 kcal/mol über der inneren Energie des energieärmsten Konformers hat sich als sinnvoll herauskristallisiert. Dies wird auch durch Studien anderer Gruppen untermauert [156]. Bei einer dockingbasierten Konformerberechnung können als Anhaltspunkt die 50 besten Dockinglösungen genommen werden. Nach der Berechnung der einzelnen Oberfläche und der darauf basierenden Deskriptoren können die Gesamtdeskriptoren über eine einfache Durchschnittsbildung ermittelt werden. Eine Wichtung mit der inneren Energie reduziert das Ensemble faktisch auf wenige Konformere, wodurch die externe Vorhersagekraft der Modelle leidet. Außerdem kann eine beliebige Regressionstechnik angewendet werden. Die Variablenselektion bietet vor allem für die Interpretierbarkeit der Modelle Vorteile und sollte deswegen verwendet werden. Dabei müssen strenge Kriterien an die Validierung angelegt werden, um Zufallskorrelationen möglichst zu vermeiden.

Im Überblick gesehen hat sich ein klares Bild der zu verwendenden Parameter herausgebildet, das sich an einer Vielzahl verschiedener Datensätze bewährt hat. Details zu den entsprechenden Datensätzen sind im nächsten Kapitel beschrieben.

3.9. Ergebnisse der Modellbildung

Der folgende Teil der Arbeit zeigt die mit xMaP erreichten Ergebnisse für die untersuchten Datensätze. Entsprechend der jeweiligen Zielsetzung können diese klassifiziert werden. Zunächst ist das Ziel, xMaP mit etablierten Techniken zu vergleichen. Hierfür dient der Datensatz mit Acetylcholinesterase-Inhibitoren (AZT). Dieser ermöglicht einen Vergleich zu GRID/GOLPE, MaP, CoMFA und CoMSIA. Überdies sind von der Acetylcholinesterase bereits viele Kristallstrukturen bestimmt worden, so dass zusätzlich ein strukturbasiertes QSAR-Modell gerechnet werden konnte. Als nächstes wird ein Datensatz von Prostaglandin $F_{2\alpha}$ -Analoga ($PGF_{2\alpha}$) bearbeitet. Dies erlaubt einen Vergleich mit der 4D-QSAR von Hopfinger [15]. Danach wird ein Datensatz von M_2 -Modulatoren (M_2) untersucht, der einen weiteren Vergleich mit der MaP-Technik erlaubt. Zusätzlich wird bei dem M_2 -Datensatz gezeigt, dass xMaP mit extrem flexiblen Molekülen umgehen kann. Außerdem wurden Inhibitoren der HIV-1 Reversen Transkriptase (HEPT) untersucht, die wiederum einen Vergleich mit CoMFA und CoMSIA ermöglichen. Auch bei diesem Datensatz gibt es von dem Zielenzym eine Kristallstruktur. Es kann ein strukturbasiertes Modell erzeugt werden. Dabei können die von xMaP identifizierten Parameter in die Bindetasche projiziert werden. Sehr interessant ist der D_1 -Datensatz von Dopamin-Antagonisten (D_1). Es war bisher mit keiner QSAR-Methode möglich, dafür ein Modell zu finden. Mit xMaP konnte ein Modell erzeugt werden. Für diesen Datensatz konnten prospektive Analysen durchgeführt und Strukturen zur Synthese vorgeschlagen werden. Zusätzlich wurde basierend auf einem Homologiemodell des D_1 -Rezeptors ein strukturbasiertes Modell für die untersuchten Daten erstellt. Die gesamte Analyse führte dann zu einer Theorie, die konformelle Änderungen im Rezeptorprotein über die Interaktion mit den Antagonisten erklären können. Ein weiterer bisher noch nicht beforschter Datensatz sind die Glukokortikoide (GK). Hier wurde neben dem xMaP-Modell auch ein CoMFA- und CoMSIA-Modell erzeugt und so eine weitere Vergleichbarkeit der Methoden erreicht. Zusätzlich konnten weitere Einblicke in den vorliegenden Datensatz gefunden werden. Dann wurden im Rahmen des Sonderforschungsbereiches 630 Naphtylisochinolin-Alkaloide (NIQ) untersucht, für die auch mit MaP ein Modell erstellt wurde. Auch hier konnte gezeigt werden, dass xMaP eine sinnvolle Alternative ist und vergleichbare Ergebnisse liefert. Außerdem wurden zur Validierung der xMaP-Technik noch weitere Standarddatensätze der QSAR untersucht. Dies ist im letzten Unterkapitel beschrieben.

3.9.1. Inhibitoren der Acetylcholinesterase (AZT)

Dieser Datensatz wurde mit xMaP intensiv untersucht, da viele Vergleichswerte vorliegen und auch die Kristallstruktur des Zielenzym bekannt ist. Darüber hinaus sind die hier zu untersuchenden Moleküle sehr flexibel. Jedes Molekül hat bei der Standard-Konformerberechnung mit Catalyst im Durchschnitt (Mittelwert und Median) 129 energiearme Konformere. Dabei liegen der niedrigste Wert bei 51 und der höchste bei 237 Konformeren. Bei der Anwendung von 3D-QSAR-Techniken ist ein starker Nutzereinfluss zu erwarten, da die Auswahl der vermuteten bioaktiven Struktur sehr schwer und daher auch subjektiv ist. Im Hinblick auf das Alignment sind ebenfalls Probleme zu erwarten. Eine Möglichkeit ist es, jeweils die beste Lösung aus einem Docking-Experiment zu nehmen. Das wurde in der ersten Publikation zu diesem Datensatz von Sippl im Sinne der Modellbildung erfolgreich durchgeführt [171]. Dabei bleibt das Problem, die „beste“ Dockinglösung zu finden. Der Grund dafür ist, dass der Vorgang der automatischen Ranglistenbildung für die Dockinglösungen (engl. *Scoring*) nach wie vor stark fehlerbehaftet ist [244,245]. Außerdem wird die QSAR normalerweise zu einem Zeitpunkt eingesetzt, wenn keine Information über das Zielenzym vorhanden ist. Die von Sippl verwendete Strategie ist daher meistens nicht anwendbar.

Unter Anwendung des xMaP-Standardprotokolls war es möglich, ein gutes Modell zu erstellen. Diese Ergebnisse werden im Folgenden diskutiert. Die Trainings- und Testdaten aus der ursprünglichen Publikation wurden verschmolzen, da die verwendete Validierung auf einer großen Anzahl von zufälligen Unterteilungen beruht. Mit allen 49 Strukturen als Grundlage ergab sich folgendes Modell mit insgesamt 5 informativsten Variablen (MIVs):

$$\hat{Y} = 7.3437 \cdot DD_2 + 0.0073039 \cdot HH_4 + 0.0013447 \cdot LwA_6 + 0.00029908 \cdot LsLw_8 - 0.0042139 \cdot LsLs_8 + 5.9878$$

Gleichung 35

$$R^2_{Test,Avg} = 0.75, RMSEP_{Test,Avg} = 0.70$$

$$R^2_{CV-50\%} = 0.84, RMSEP_{CV-50\%} = 0.56, m = 49$$

Dabei ist \hat{Y} die vorhergesagte biologische Aktivität und m die Gesamtzahl von Molekülen im Datensatz. Die Variablenerklärungen sind in Tabelle 1 zu finden. Die Indizes beschreiben die mittlere Distanz zwischen den Schwerpunkten der für diesen potenziellen Zweipunkt-Pharmakophor relevanten Oberflächenareale. Die Variablen DD_2 , HH_4 , LwA_6 und $LsLw_8$ weisen einen positiven Regressionskoeffizienten auf, $LsLs_8$ dagegen einen negativen. Es muss beachtet werden, dass aufgrund der Mittelwertzentrierung der Daten ein positiver Koeffizient

nicht unbedingt einen positiven Beitrag zur biologischen Aktivität beschreibt. Fehlt einem Molekül beispielsweise eine negativ mit der biologischen Aktivität korrelierte Eigenschaft (negatives Vorzeichen beim Regressionskoeffizienten), so ist an der entsprechenden Stelle in der Matrix der Eintrag für die Variable negativ. Dadurch wird eine höhere biologische Aktivität vorhergesagt.

Für den AZT-Datensatz ist Variable HH_4 sehr wichtig. Diese beschreibt zwei hydrophile Oberflächenareale im Abstand von 4 Å. Allein diese Variable kann fast zwei Drittel der Varianz in den Daten erklären ($R^2=0.66$, $R^2_{CV-I}=0.63$, $R^2_{CV-50\%}=0.61$). Eine ausführliche Interpretation dieser Variable zeigt den größten Vorteil von 4D-Techniken gegenüber anderen QSAR-Methoden mit denen dieser Datensatz bisher untersucht wurde. Deswegen wird die Interpretation dieser Variable als erstes gezeigt.

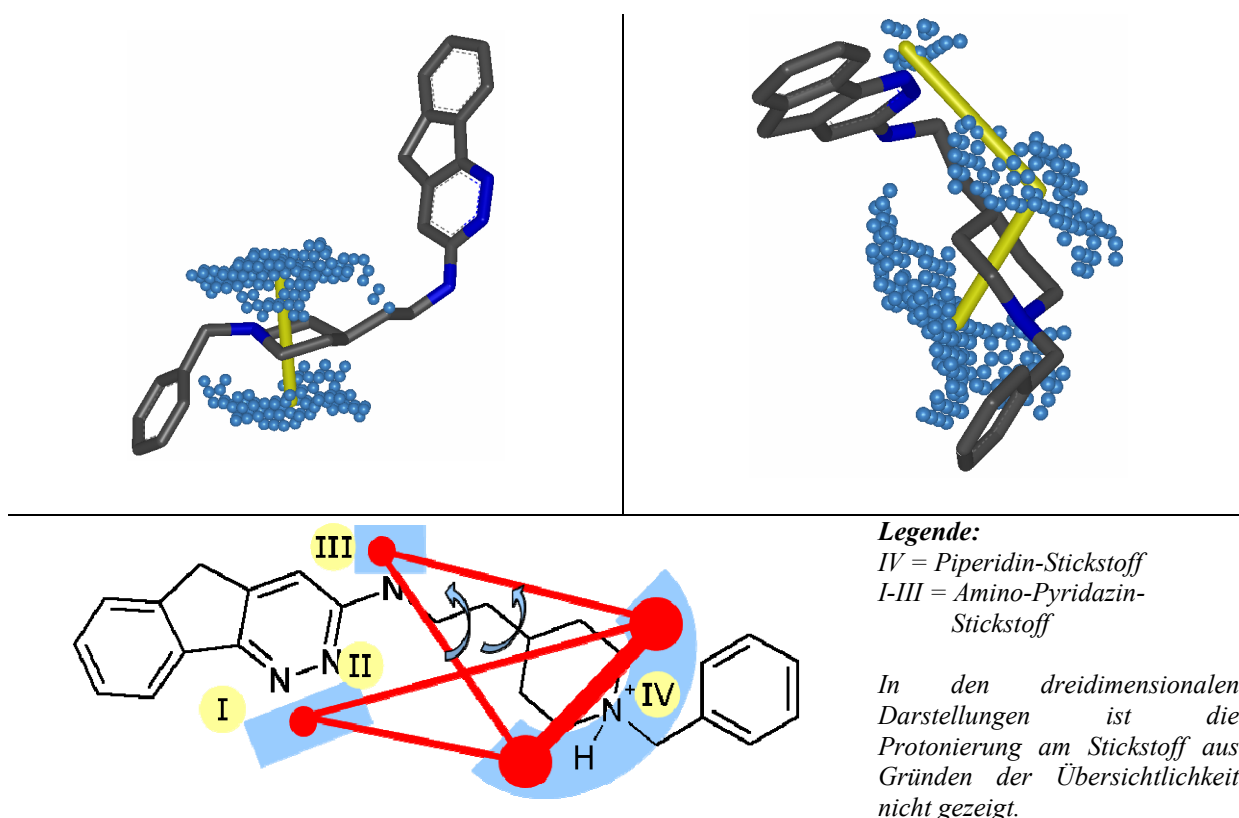


Abbildung 43:

Die Interpretation von Variable HH_4 : Die Wichtigkeit (das Vorhandensein) des protonierten Stickstoffs (IV) wird über die Position zweier definierter hydrophiler Patches in seiner direkten Umgebung kodiert. Das ist hier am Beispiel von Molekül 30 gezeigt. Dies ist jedoch nur ein Aspekt dieser Variable. AChE-Inhibitoren sollten außerdem flexibel sein. Diese Flexibilität wird ebenfalls in Variable HH_4 kodiert. Die Position zweier anderer hydrophiler Areale, die die Position der Amino-Pyridazin-Stickstoffe I bis III beschreiben, geht ebenfalls in HH_4 ein. Diese Areale können so zu den Arealen am Piperidin-Stickstoff IV positioniert sein, dass über die unscharfe Zählweise (der Hauptanteil wird in HH_5 registriert) auch HH_4 inkrementiert wird. Das gilt für etwa die Hälfte der Konformere von flexiblen Molekülen. In der 3D-Darstellung auf der rechten Seite ist ein Beispiel dafür gezeigt: Je höher der pIC_{50} -Wert ist desto höher ist auch das Inkrement, das über derartige flexible Konformere zu Variable HH_4 addiert wird, wodurch die Flexibilität kodiert wird. Durch HH_4 wird beispielsweise die Anzahl der rotierbaren Bindungen oder die Position des Stickstoffes im Ring zwischen den hydrophilen Patches und damit die erforderliche Flexibilität beschrieben. In Abbildung 44 sind die Moleküle 10 und 13 gezeigt, bei denen die Distanz zwischen den angesprochenen Bereichen kürzer (veränderte Position des protonierten Stickstoffs, niedrigere Flexibilität) und somit der vorhergesagte pIC_{50} niedriger ist.

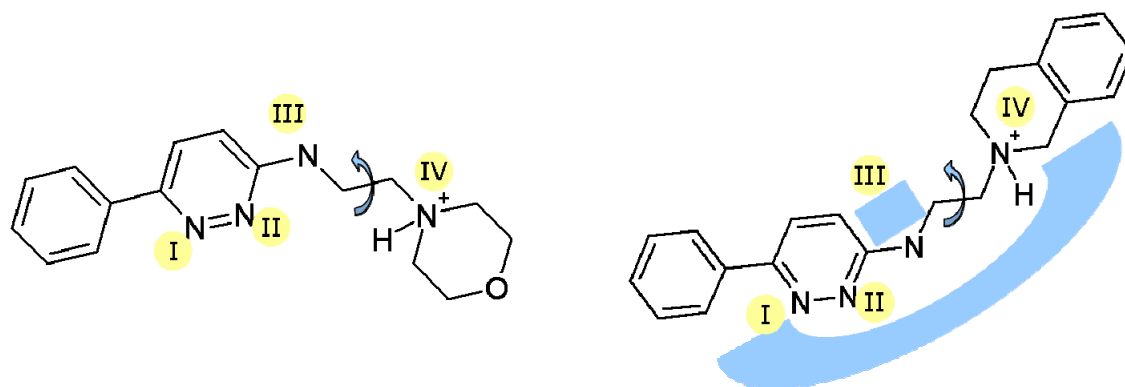
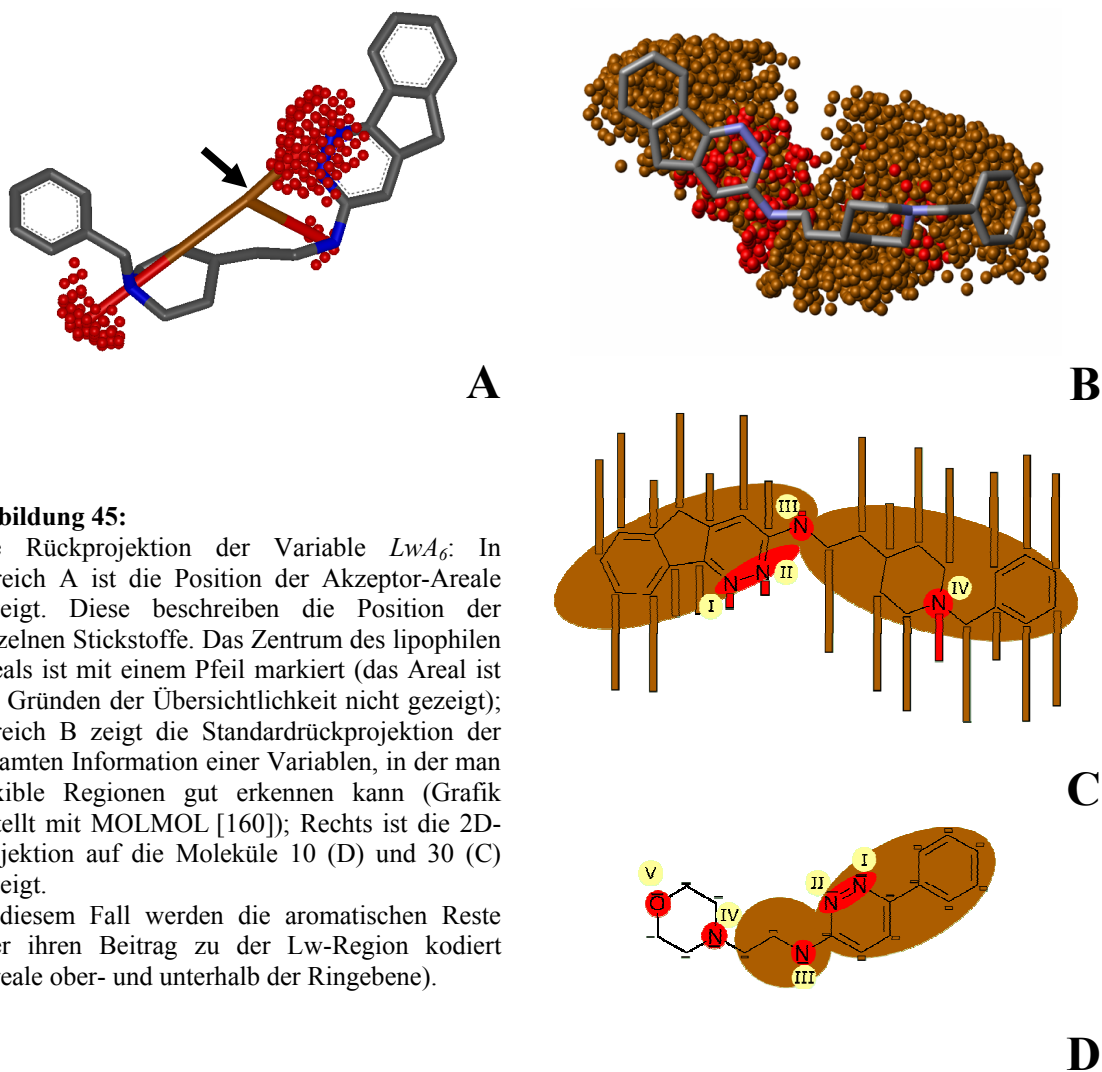


Abbildung 44: Rückprojektion von Variable HH_4 auf die Moleküle 10 (links) und 13 (rechts). Diese Darstellung verdeutlicht, dass bei Molekülen, die keinen ausreichenden Abstand zwischen dem protonierten Stickstoff IV und den Aminopyridazin-Stickstoffen (I - III) aufweisen, die Variable nur in geringerem Ausmaß inkrementiert wird. Damit haben sie eine niedrigere vorhergesagte (und gemessene) Aktivität.

HH_4 beschreibt das Vorhandensein eines protonierten und damit quartären Stickstoffs und seiner Umgebung. Das ist in den Abbildungen 43 und 44 zu sehen. Der entsprechende Piperidin-/Morpholin-Stickstoff ist als Nr. IV markiert. Zwei diesen Stickstoff umgebende hydrophile Areale führen zu hohen Werten für Variable HH_4 . Zusätzlich beschreibt HH_4 auch die Flexibilität, die zwischen dem Piperidin-Stickstoff IV und dem Aminopyridazin (Stickstoffe I-III) notwendig ist. Es ist bekannt, dass diese Flexibilität bei AChE-Inhibitoren essenziell ist. Die Bindetasche ist eine tiefe Kavität, die flexible Liganden bei Molekülen dieses Typs erfordert. Bei der Durchführung einer QSAR-Analyse liegt üblicherweise keine Kristallstruktur der Zielstruktur vor. Sofern eine QSAR-Analyse Informationen über nötige Ligand-Flexibilität trotzdem liefert, so unterstreicht dies die gute Eignung dieser Technik für das vorliegende Problem. Bei xMaP ist das für den AZT-Datensatz gegeben. Bei Molekül 30, welches sehr aktiv ist ($pIC_{50}=8$), erfüllen 84% der Konformere die Bedingung, Variable HH_4 zu inkrementieren. Es sind drei rotierbare Bindungen zwischen dem protonierten Stickstoff IV und dem Aminopyridazin zu finden. Der Piperidin-Ring ist über die 3-Position mit der Seitenkette verknüpft, an der das Aminopyridazinfragment zu finden ist. Auch liegt der protonierte Stickstoff im weiter vom Aminopyridazin entfernten Bereich des Ringes. Die beiden großen hydrophilen Oberflächenareale um Stickstoff IV können daher folgende Position einnehmen: Durch die Faltung/Winkelung der Moleküle wird die Distanz zu den hydrophilen Oberflächenarealen der Aminopyridazin-Stickstoffe I-III klein genug, um Variable HH_4 zu inkrementieren. Dies liegt in der Flexibilität begründet und ist in Abbildung 43 oben rechts gezeigt. Durch diesen zusätzlichen Beitrag zu Variable HH_4 wird ein höherer pIC_{50} -Wert vorhergesagt. Moleküle, denen eine der Bindungen zwischen dem protonierten

Stickstoff und den Aminopyridazin-Stickstoffen fehlt oder bei denen die Position des Stickstoffes im Ring anders ist, weisen einen geringeren Beitrag für diese Variable auf. Beispiele dafür sind die Moleküle 10 und 13. Diese sind in Abbildung 44 gezeigt. Bei diesen Molekülen ist die Flexibilität zwischen dem Aminopyridazin und dem quartären Stickstoff geringer. Dies ist der Fall, da aufgrund der kleineren Distanz zwischen den funktionellen Gruppen die Positionen des protonierten Stickstoffs im Raum stärker eingeschränkt sind. Diese Moleküle haben außerdem weniger energiearme Konformere, sie sind deutlich starrer. Wegen dieser veränderten Position des protonierten Stickstoffs im Raum können die Areale in der Umgebung von Stickstoff IV und den Stickstoffen I - III verschmelzen. Der Abstand zwischen den hydrophilen Arealen um Stickstoff IV und den Arealen um das Aminopyridazin wird somit zu klein. Das hat zur Folge, dass HH_4 aufgrund der veränderten Position des Stickstoffs und somit als direkte Folge der niedrigeren Flexibilität zwischen den hydrophilen Regionen nicht mehr inkrementiert wird.



Somit wird der vorhergesagte pIC_{50} -Wert deutlich niedriger. Um auf das Beispiel in Abbildung 44 zurückzukommen: Bei Molekül 10, das einen pIC_{50} -Wert von 3.1 hat, erfüllt kein einziges Konformer die Bedingung, HH_4 zu erhöhen. Die Position des Stickstoffs IV im Morpholinrest führt zu einer Verschmelzung der beiden Patches aufgrund der hydrophilen Umgebung. Bei Molekül 13 sind es immerhin 21 von 58 Konformeren, also 36%, durch die HH_4 inkrementiert wird. Der pIC_{50} liegt bei 4.08. Hier ist der gesamte Beitrag der Variable HH_4 auf die Beschreibung des Vorhandenseins von Stickstoff IV zurückzuführen. Um den genauen Einfluss der Distanz zwischen diesen Molekülbereichen zu verstehen, muss zusätzlich Variable LwA_6 berücksichtigt werden. Diese beschreibt ein schwach lipophiles Areal im Abstand von 6 Å zu einem Wasserstoffbrücken-Akzeptor.

Variable LwA_6 hat ihren ersten Startpunkt im geometrischen Schwerpunkt eines großen lipophilen Oberflächenareals. Dies ist in Abbildung 45 oben links gezeigt. Dieses Areal umspannt bei den hochaktiven Substanzen aus dem Datensatz fast das ganze Molekül. In schwächer aktiven Molekülen ist es deutlich kleiner und sein Zentrum verschoben. Schwach aktive Moleküle haben an einem Terminus einen großen hydrophilen Rest. Dies ist für Verbindung 10 in Abbildung 45D gezeigt. Der Morpholinrest ist die Basis für ein hydrophiles Areal weshalb der lipophile Bereich kleiner wird. Die Akzeptorendpunkte beschreiben die Positionen der einzelnen Stickstoffe. Der Grund dafür ist die doppelte Eigenschaftszuweisung im Oberflächenalgorithmus. Im Gegensatz zu Variable HH_4 , bei der die Zuweisung „hydrophil“ genutzt wird, um die Stickstoffe zu charakterisieren, ist bei LwA_6 die Zuweisung „Akzeptor“ relevant. Diese trifft für alle Stickstoffe in gleichem Maße zu. Die Akzeptorzuzuweisung ist deutlich schärfer abgegrenzt. Deshalb können einzelne Stickstoffe unterschieden werden, was bei HH_4 nicht der Fall ist. Als Konsequenz kann LwA_6 die genaue Position der Stickstoffe beschreiben und liefert wertvolle Zusatzinformation zu HH_4 . Informationen dazu finden sich auch in Abbildung 46. Untersucht man die Unterschiede in Variable LwA_6 an verschiedenen Molekülen genauer, fällt folgendes auf: Alle Konformere von Verbindung 10 tragen zum Gesamtbeitrag dieser Variable bei, obwohl dies die Verbindung mit der niedrigsten Aktivität im Datensatz ist. Der Beitrag pro Konformer ist jedoch geringer als bei Verbindung 30, bei der 123 der 133 Konformere, also 92.5 %, zu dieser Variable beitragen. In absoluten Zahlen ausgedrückt ist der Gesamtbeitrag pro Konformer bei Verbindung 30 rund 3.5-mal so hoch (1400.8 zu 393.58). Betrachtet man die Daten genauer, wird der Grund dafür deutlich: Bei Verbindung 10 definieren verschiedenste Atome (die Stickstoffe I-IV und Sauerstoff V, siehe Abbildung 45 rechts unten) die Akzeptorbereiche, die für LwA_6 wichtig sind. Deshalb gibt es immer irgendeine Möglichkeit, dass das zentrale lipophile Areal in passender Entfernung zu einem Akzeptor liegt. Dann kann

die dazwischen liegende Distanz einen gerundeten Wert von 6 Å einnehmen oder im Bereich von 4-5 Å bzw. 7-8 Å liegen. Durch die unscharfe Zählweise wird dann Variable LwA_6 inkrementiert. Der andere Grund für den niedrigeren Beitrag zur Variablen LwA_6 ist der Morpholinrest, der den Schwerpunkt des lipophilen Areal durch eine Größenreduktion verschiebt. Ein kleineres lipophiles Areal führt zu einem niedrigeren Inkrement. Im Gegensatz dazu sind in Verbindung 30 die Akzeptorpositionen durch die Stickstoffe I bis IV sehr gut im Zentrum des Moleküls definiert. Ein Beispiel dazu ist in Abbildung 45 zu sehen. Zusätzlich liegt der Schwerpunkt des großen lipophilen Areal praktisch in der Mitte des Moleküls. Dieses große lipophile Areal umspannt fast das komplette Molekül. Das ist durch die terminalen Aromaten und somit lipophilen Termini in Molekül 30 bedingt. Insgesamt erklärt Variable LwA_6 etwa die Hälfte der Varianz der Daten im internen Modell ($R^2=0.52$, $R^2_{CV-I}=0.48$, $R^2_{CV-50\%}=0.46$).

Zusammen mit Variable HH_4 wird beschrieben, dass der flexible Bereich zwischen den einzelnen Stickstoffen gestreckt sein muss, um die optimale Aktivität zu erreichen. Dies trifft analog für Variable LwA_5 zu, die ähnlich viel Varianz erklärt.

Der Schwerpunkt des zentralen lipophilen Areal ist auch der Startpunkt für Variable $LwLs_8$. Diese beschreibt einen schwach lipophilen Bereich in einer Distanz von 8 Å zu einem stark lipophilen Bereich. Diese Variable erklärt ebenfalls sehr viel Varianz: $R^2=0.64$, $R^2_{CV-I}=0.62$, $R^2_{CV-50\%}=0.60$. Die Endpunkte der Variable $LwLs_8$ sind die terminalen Aromaten der stark aktiven Substanzen. Die Schwerpunkte dieser stark lipophilen Areale müssen in einem Abstand von 8 Å zum Mittelpunkt des zentralen schwach lipophilen Bereichs liegen. Damit wird die optimale Molekülgröße kodiert. Ein Überblick über die Interpretation aller Variablen ist in Abbildung 46 gezeigt.

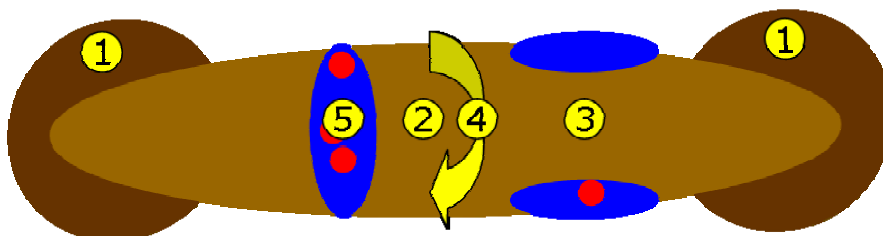


Abbildung 46: Das komplette Pharmakophormodell für den AZT-Datensatz, wie es durch das xMaP-Modell definiert wird.

Zusammenfassend sind also für eine möglichst gute Inhibition der AChE folgende Faktoren nötig:

- (1) Zwei terminale aromatische Reste deren Lage und Abstand durch Variable $LsLw_8$ definiert wird. Dies wird durch Kristallstrukturen, beispielsweise PDB:1ODC bestätigt. Auf beiden Seiten der Bindetasche werden π - π - π -Sandwiches ausgebildet, die Termini der Liganden lagern sich zwischen jeweils zwei aromatische Reste im Protein ein.
- (2) Der zentrale schwach lipophile Bereich, der fast das ganze Molekül umspannt. Sein Zentrum ist sowohl für $LsLw_8$ als auch für LwA_6 die Grundlage. LwA_6 beschreibt die Position des essenziellen protonierten Stickstoffes und des basischen tertiären Amins relativ zum großen lipophilen Bereich, der Größe und Form der Moleküle beschreibt.
- (3) Das Vorhandensein des protonierten quartären Stickstoffes, der in Variable HH_4 durch einen darüber- und einen darunterliegenden hydrophilen Bereich definiert ist. Dieser Stickstoff bildet in der Bindetasche π -Kation-Wechselwirkungen mit dem Phenylalanin 330 und dem Tyrosin 334 [171].
- (4) Die flexible Verbindung zwischen dem protonierten quartären Stickstoff und dem Aminopyridazin. Diese ist für die Interaktion mit dem Rezeptor eminent wichtig und wird in Variable HH_4 kodiert. Durch gewinkelte Konformere besonders flexibler Verbindungen kann HH_4 zusätzlich inkrementiert werden.
- (5) Das Aminopyridazin und die heterozyklischen Stickstoffe durch die es definiert wird, die als Ganzes einen der Startpunkte von HH_4 definieren beziehungsweise einzeln die Akzeptoren für LwA_6 darstellen.

Abbildung 47 zeigt die Qualität der Testdatenvorhersage für den AZT-Datensatz. Bei den Vorhersagen für die Testobjekte handelt es sich um die Mittelwerte der Vorhersage aller Modelle in denen das entsprechende Objekt nicht Teil des Trainingsdatensatzes war (Sub-Bagging, siehe Kapitel 2.2.4).

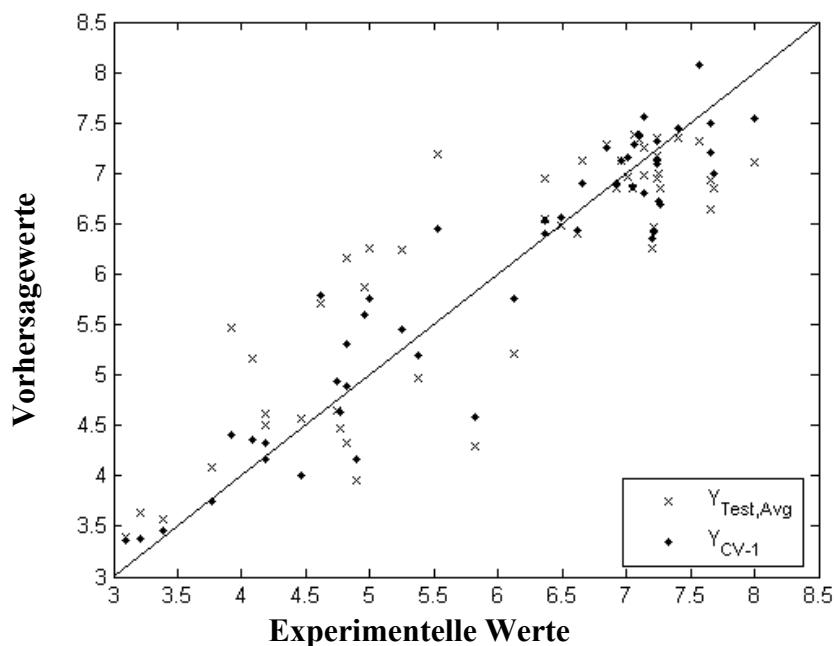


Abbildung 47: Der Vergleich der experimentellen mit den vorhergesagten Werten für den AZT-Datensatz bezogen auf Trainings- und Testdatenvorhersage. Jedes Molekül war sowohl Teil vom Trainings- als auch vom Testdatensatz ist ($R^2_{CV-1}=0.86$, $R^2_{CV-50\%} = 0.84$, $R^2_{Test,Avg} = 0.75$). $Y_{Test,Avg}$ steht dabei für Mittelwerte der Testdatenvorhersage und Y_{CV-1} für die Vorhersagewerte aus einer Leave-one-out-Kreuzvalidierung.

Die bis hierhin gezeigten Untersuchungen für den AZT-Datensatz wurden mit einer schwach hydrophilen Parametrisierung von protonierten Stickstoffen durchgeführt. Einem protonierten Stickstoffatom wurde dabei nach Ghose ein Hydrophiliewert von -0.5427 zugewiesen [158]. Dabei werden quartäre und tertiäre Stickstoff-Atome als identisch angesehen. Damit ist eine Modellinterpretation gut möglich. In einem weiteren Ansatz wurde in der Oberflächenberechnung den protonierten Stickstoffen ein stärkerer Wert (-1.4439) für die Hydrophilie zugewiesen. Dies war der einzige Unterschied in der Berechnung, der Einfluss auf das Modell ist wie folgt:

$$\hat{Y} = 0.0075 \cdot LwA_{14} + 0.0022 \cdot LwD_3 - 0.0123 \cdot LsD_5 + 0.0007 \cdot LsH_6 + 0.0008 \cdot LsLw_{12} + 5.9878$$

Gleichung 36

$$R^2_{Test,Avg} = 0.77, RMSEP_{Test,Avg} = 0.67$$

$$R^2_{CV-50\%} = 0.83, RMSEP_{CV-50\%} = 0.58, m = 49$$

In seiner statistischen Qualität ist dieses in der Stickstoffbeschreibung modifizierte Modell mit dem Vorherigen nahezu gleich. Die ausgewählten Variablen sind jedoch andere. Besonders die Variablen LwA_{14} , LsH_6 und $LsLw_{12}$ erklären mit jeweils rund 60% sehr viel der Variabilität der Daten. $LsLw_{12}$ wird außerdem in den Trainings-/Testdatensatz-Unterteilungen am häufigsten selektiert. Variable HH_4 , die im ersten Modell sehr wichtig ist

kann hier nur mehr rund 45% der Variabilität erklären und wird in den Unterteilungen nur ein einziges Mal selektiert. Dieser Unterschied soll im Folgenden erläutert werden.

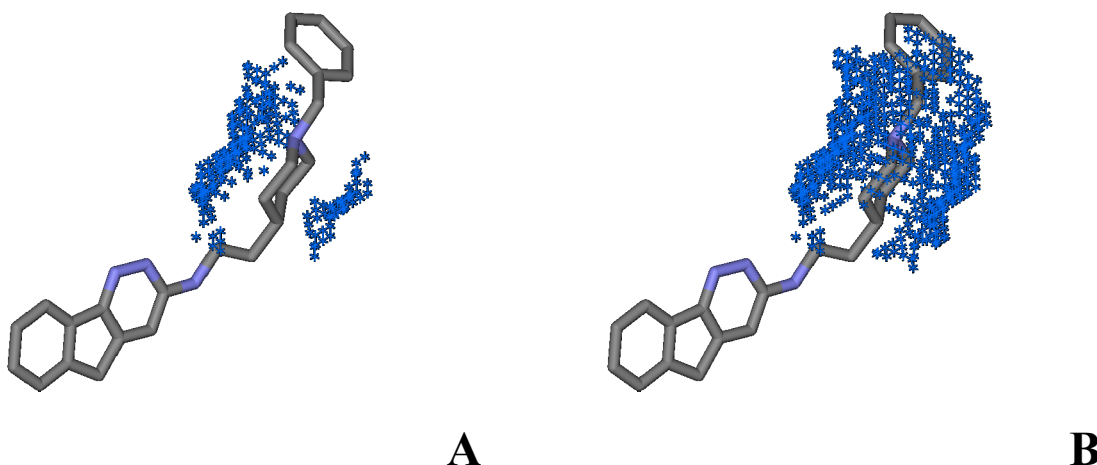


Abbildung 48: Die hydrophile Oberflächenbeschreibung um den protonierten Stickstoff in Struktur 30 aus dem AZT-Datensatz. A: Mit der schwach hydrophilen Parametrisierung der Hydrophilie um den protonierten Stickstoff. B: Mit der stark hydrophilen Parametrisierung der Hydrophilie

Abbildung 48 zeigt den Unterschied in der Oberflächenberechnung um den protonierten Stickstoff. An Stelle zweier hydrophiler Bereiche zur Beschreibung der hydrophilen Umgebung des Stickstoffs gibt es einen einzigen Bereich, der das Molekül im Bereich des protonierten Stickstoffes vollständig umschließt. Diese Veränderung beeinflusst die Interpretation des Modells wie folgt. In Analogie zu Abbildung 46 ergibt sich folgendes Bild für einen hochaktiven AchE-Inhibitor (siehe Abbildung 49):

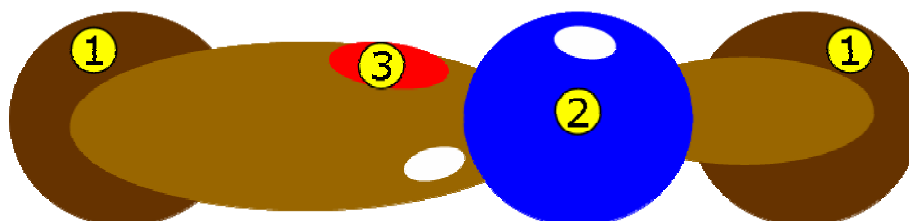


Abbildung 49: Das komplette Pharmakophormodell für den AZT-Datensatz, wie es durch das für die Beschreibung von protonierten Stickstoffen modifizierte xMaP-Modell definiert wird.

- (1) Zwei terminale aromatische Reste deren Lage und Abstand durch Variable $LsLw_{12}$ definiert wird. Im Gegensatz zur $LsLw_8$ -Variable im ersten Modell werden hier beide schwach lipophile Bereiche als Ausgangspunkte verwendet. Es wird die Position des jeweils gegenüberliegenden stark lipophilen Bereichs kodiert.
- (2) Das Vorhandensein des protonierten und somit quartären Stickstoffes: Die optimale Position wird relativ zum näher liegenden stark lipophilen Areal (der dunkelbraune Bereiche in Abb. 49 rechts von dem blauen Bereich) in der Variablen LsH_6 kodiert. Die negativ korrelierte Variable LsD_5 ist komplementär dazu, hier dient die Zuweisung Donor zum protonierten Stickstoff als Ausgangspunkt für die Variable.

Ein zu kurzer Abstand zwischen dem endständigen Aromaten und dem Stickstoff wirkt negativ auf die Aktivität.

- (3) Die Position des Aminopyridazins: Seine Position wird über die Distanz zwischen dem Schwerpunkt des kleineren schwach lipophilen Bereichs (der hellbraune Bereich in Abb. 49 rechts von dem blauen Bereich) zu dem Akzeptorbereich um die aromatischen Stickstoffe in Variable LwA_{14} beschrieben. Diese Variable liefert eine sehr ähnliche Aussage wie LwA_6 im ersten Modell.

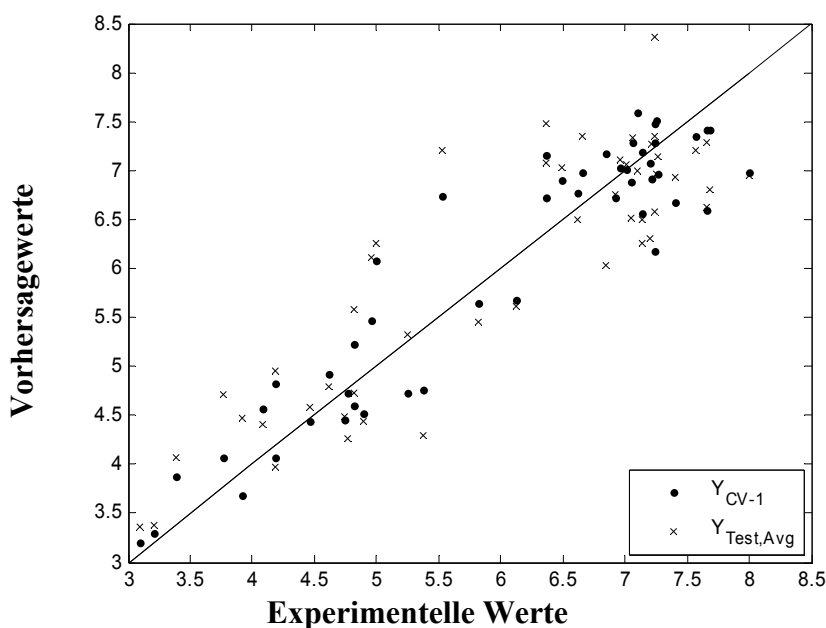


Abbildung 50: Der Vergleich der experimentellen mit den vorhergesagten Werten für den AZT-Datensatz bezogen auf Trainings- und Testdatenvorhersage mit dem in der Stickstoff-Beschreibung modifizierten Modell ($R^2_{CV-1}=0.85$, $R^2_{CV-50\%}=0.83$, $R^2_{Test,Avg}=0.77$).

Dieses Modell hat somit eine ähnliche Aussage wie das erste Modell, kann aber die Information über die nötige Flexibilität nicht kodieren. Die statistische Modellqualität ist sehr ähnlich (siehe Abbildung 50).

Lässt man die beiden bisher beschriebenen Modelle Revue passieren, so zeigt sich folgendes: Eine schwächere Parametrisierung der Hydrophilie um den protonierten Stickstoff wie im ersten Modell ist für die Interpretation von Vorteil. Neben der reinen Grundstruktur des Moleküls, die in beiden Modellen gleich gut beschrieben wird, beschreibt das erste Modell zusätzlich die nötige Flexibilität der untersuchten Moleküle. Das Gesamtbild ist also detaillierter, da die Position des protonierten Stickstoffs besser aufgeschlüsselt werden kann. Es wird daher empfohlen, den Einfluss der Parametrisierung der Hydrophilie an protonierten Stickstoffen zu untersuchen und die für die Interpretation sinnvollere Variante zu wählen. Zudem könnte analog zu MaP versucht werden, die Hydrophilie in starke und schwache

Bereiche zu gliedern und somit das große Areal um den Stickstoff in mehrere kleine zu unterteilen.

Zusammenfassend führt die Untersuchung des AZT-Datensatzes zu folgendem Ergebnis: Es war möglich, ein Modell zu erstellen, das gute statistische Güteparameter zeigt und dessen statistische Qualität vergleichbar mit der von verschiedenen 3D-QSAR-Techniken ist (siehe Tabelle 9). Außerdem ist das gefundene Modell gut interpretierbar.

Tabelle 9: Überblick über die statistischen Güteparameter von Modellen für den AZT-Datensatz, die mit 3D-QSAR-Techniken erzeugt wurden. Der jeweils gezeigte R^2_{Test} -Wert beruht im Gegensatz zu xMaP auf nur einem einzigen Testdatensatz und somit nur einer einzigen Vorhersage. Es sind Daten des jeweiligen manuellen Alignments [19,80] gezeigt. Der R^2_{Test} -Wert für GRID/GOLPE ist in der Publikation [171] nicht angegeben. Er wurde anhand der dort angegebenen Werte errechnet.

QSAR-Methode	R^2	R^2_{CV-1}	$R^2_{CV-50\%}$	R^2_{Test}
MaP [19,80]	0.87	0.85	0.83	0.78
GRID/GOLPE [171]	0.99	0.94	0.91	0.04
CoMFA [19,80]	0.97	0.81	-	0.60
CoMSIA [19,80]	0.93	0.83	-	0.69
xMaP	0.89	0.86	0.84	0.75

Zusätzlich wird im xMaP-Modell mit den Standardparametern die wertvolle Information der nötigen Flexibilität mit kodiert. Dies stellt eine wichtige Zusatzinformation zu den 3D-Techniken dar.

Bei diesem Datensatz wurde zusätzlich ein strukturbasiertes Modell berechnet. Dieses basiert darauf, dass Konformere durch ein FlexX-Docking in die Bindetasche erzeugt wurden. Hierbei ergab sich nach der Deskriptorberechnung folgendes Modell, das sich in seiner Qualität von dem ligandbasierten nicht unterscheidet:

$$\hat{Y} = 0.0176 \cdot HH_3 + 0.1584 \cdot HH_{23} + 0.1333 \cdot HH_{24} \\ + 0.1341 \cdot LsH_{31} + 0.2737 \cdot LsH_{32} + 0.0045 \cdot LsLw_{20} \\ + 0.0029 \cdot LsLs_{16} + 5.9878$$

Gleichung 37

$$R^2_{Test,Avg} = 0.71, RMSEP_{Test,Avg} = 1.39$$

$$R^2_{CV-50\%} = 0.83, RMSEP_{CV-50\%} = 0.58, m = 49$$

Hier wurden insgesamt 7 MIVs identifiziert, also 2 mehr als im ligandbasierten Modell. Besonders auffällig ist, dass eine sehr ähnliche Variable wie im ligandbasierten Modell die zentrale Position einnimmt: Statt der HH_4 beschreibt die HH_3 einen großen Teil der Variabilität in den Daten ($R^2=0.60$, $R^2_{CV-1}=0.57$, $R^2_{CV-50\%}=0.54$). Auch sie ist positiv mit der biologischen Aktivität korreliert. Betrachtet man die Rückprojektion der gewählten Variablen genauer, so wird der Unterschied zwischen beiden Modellen deutlich. Im ersten Modell

kodiert die HH_4 zusätzlich zum Vorhandensein des quartären Stickstoffs die nötige Flexibilität zwischen dem aromatischen Amin und diesem Stickstoff. Die HH_3 dagegen kodiert ausschließlich jeweils die Anwesenheit dieses Atoms und die des aromatischenamins. Die dazwischen liegende Region wird nicht berücksichtigt. Deswegen erklärt HH_3 auch etwas weniger von der Gesamtvariabilität. In diesem Falle trägt die Variable also weniger Information. Abbildung 51 stellt die kodierte Information nochmals dar.

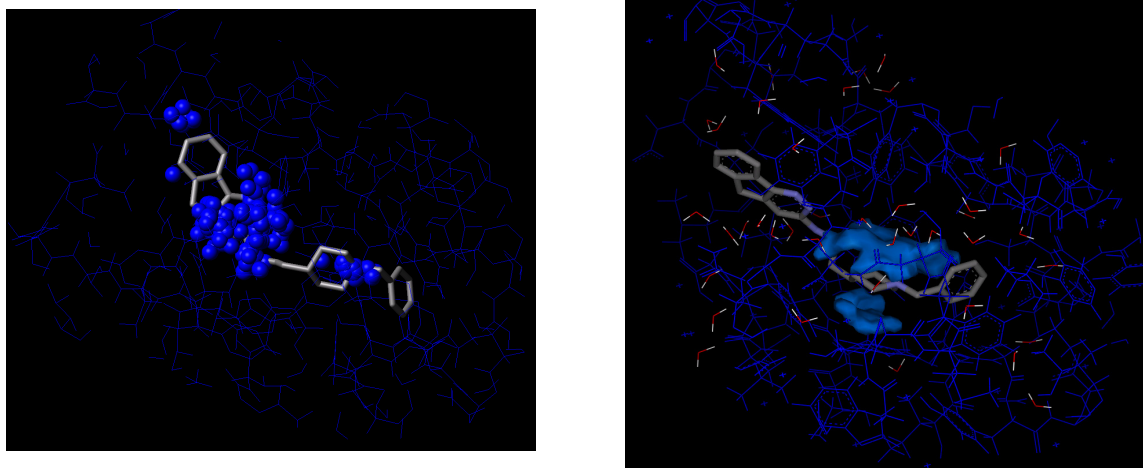


Abbildung 51: Die Rückprojektion von Variable HH_3 im strukturbasierten Modell in die Bindetasche der Acetylcholinesterase. Die Position der verschiedenen Stickstoffe wird kodiert. Information über die nötige Flexibilität fließt nicht in das Modell ein, da die Bindetasche eine zu starke Winkelung der Moleküle verhindert und die Moleküle starr in Deskriptoren (die Flexibilität hingegen ist vor allem während des Assoziationsvorgangs wichtig). Rechts ist ein repräsentatives Konformer gezeigt.

Die weiteren Variablen im Modell beschreiben sehr langreichweitige Informationen, die im Wesentlichen die Optimallänge für Inhibitoren kodieren. Dies trifft beispielsweise für die Variable $LsLs_{16}$ zu. Diese erklärt gut 20% der Variabilität in den Daten. Diese kann analog zur Variable $LwLs_8$ im ligandbasierten Modell interpretiert werden. Dabei wird der zentrale lipophile Bereich nicht mehr als Ausgangspunkt zur Definition der Position verwendet. Es wird direkt der Abstand zwischen den Aromaten an den Molekültermini kodiert. Der Molekülaufbau wird durch die Beschreibung von hydrophil kodierten Substituenten der lipophilen Molekül-Termini kodiert. Das trifft auf die langreichweitigen LsH - und HH -Variablen zu. Diese langreichweitigen Variablen erklären allein jeweils nur einen sehr geringen Teil der Variabilität in den Daten.

In den verschiedenen Bagging-Läufen werden die bereits erwähnten Variablen HH_3 und $LsLs_{16}$ am häufigsten ausgewählt. Das zeigt, dass sie für die Modellierung sehr wichtig sind und bedeutende Merkmale beschreiben, die von der Datensatzunterteilung nicht abhängen und somit sehr generell sind.

Als Fazit lässt sich für das strukturbasierte Modell ziehen, dass hier deutlich langreichweitigere Muster kodiert werden als dies im ligandbasierten Modell der Fall ist. Diese langreichweitigen Einflüsse erklären immer nur einen kleinen Teil der Variabilität. Information über die nötige Flexibilität wird nicht kodiert. Vielmehr wird die Gesamtgestalt des Liganden beschrieben. Dies ist nachvollziehbar, da in der Bindetasche die Moleküle ausgestreckt vorliegen und starr in Deskriptoren umgesetzt werden. Genau dieses Ausstrecken definieren die langreichweitigen Variablen. Das Gesamtbild ist schärfer als im ersten Modell. Die entscheidenden Informationen werden im Modell durch Nutzung in sich ähnlicherer Konformeren kodiert. Eine eindeutige Zuweisung der relevanten Interaktionen im Protein wird dadurch möglich. Die Interpretation ist präziser als im ligandbasierten Modell und kann zusätzlich die entscheidenden Aminosäuren identifizieren. Das Zusammenspiel von QSAR-Modell und strukturbasierter Information verschafft einen zusätzlichen Vorteil vor allem in der Interpretier- und Kommunizierbarkeit der Ergebnisse.

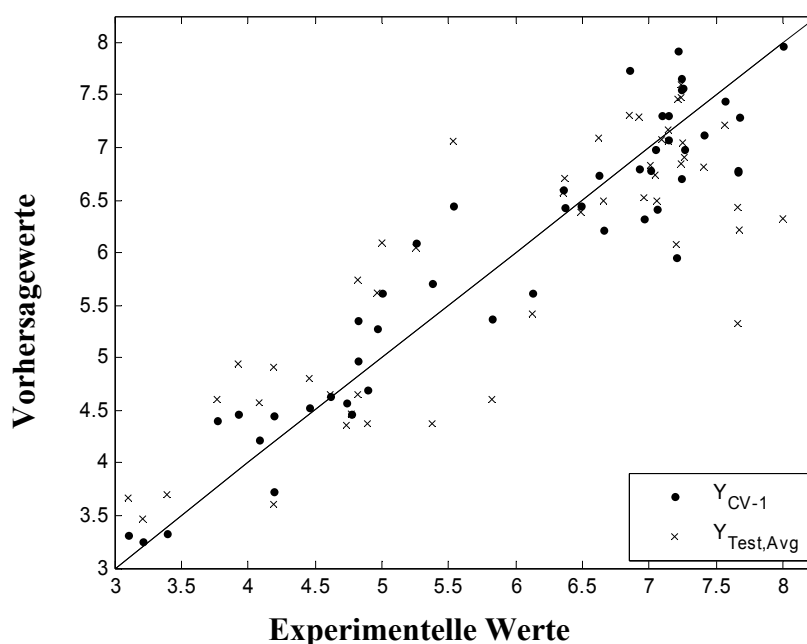


Abbildung 52: Der Vergleich der experimentellen mit den vorhergesagten Werten für das strukturbasierte Modell des AZT-Datensatzes bezogen auf Trainings- und Testdatenvorhersage ($R^2_{CV-1}=0.70$, $R^2_{Test,Avg} = 0.64$).

Die Abbildungen 47 und 52 untermauern noch einmal die gute Vorhersagekraft der Modelle und zeigen den Vergleich der vorhergesagten und gemessenen Werte im Detail.

3.9.2. Prostaglandin F_{2α}-Analoga (PGF_{2α})

Dieser Datensatz war einer der ersten, der von Hopfinger mit seiner 4D-QSAR-Technik untersucht wurde [15]. Das Hauptproblem war dabei, ein sinnvolles Alignment aller Substanzen zu finden. Dieser Schritt hatte extremen Einfluss auf das Ergebnis. Der kreuzvalidierte quadrierte Korrelationskoeffizient R^2_{CV-1} variierte bei verschiedenen Alignments im Bereich von 0.4 bis 0.8. In einem weiteren Ansatz wurde von Martinek und Mitarbeitern mittels 3D-SOMFA ein Modell für diesen Datensatz erstellt [190]. Auch dieses Modell ist extrem anfällig im Hinblick auf die Auswahl des vermuteten biologisch aktiven Konformers und die räumliche Überlagerung der Moleküle. Eine zweistufige Prozedur war nötig, um gute Ergebnisse zu erhalten: Zunächst wurden sinnvolle Konformere aus einem zuvor errechneten Konformerensemble ausgewählt und anschließend überlagert. Da keinerlei Information über den Rezeptor vorhanden ist, beruht jegliche Konformerenauswahl und jeder Überlagerungsschritt auf der Optimierung der statistischen Modellqualität. Das Evaluieren vieler Modelle zur Bestimmung methodenrelevanter Parameter (Auswahl der Konformere, Art der Überlagerung) erhöht die Wahrscheinlichkeit für Zufallskorrelationen und führt zu überoptimistischen internen statistischen Güteparametern [146,148]. Derartige Auswahlsschritte sind im Falle von xMaP bedingt durch das Design des Deskriptors unnötig. Im ersten Ansatz wurde auf Basis der Catalyst-Konformer ein Modell erstellt. Dabei hatte jedes Molekül im Schnitt 211 (Median: 225) energiearme Konformere (min. 83; max. 252). Es resultierte folgendes Modell:

$$\begin{aligned} \log(ED_{50}) = & 0.065465 \cdot AA_4 + 0.070762 \cdot DD_4 \\ & - 0.88088 \cdot DD_{19} - 0.0027916 \cdot HD_4 + 0.25736 \cdot LwD_{22} \\ & - 0.00013058 \cdot LwH_5 + 1.0852 \end{aligned}$$

Gleichung 38

$$R^2_{Test,Avg}=0.41, RMSEP_{Test,Avg}=0.72$$

$$R^2_{CV-50\%}=0.83, RMSEP_{CV-50\%}=0.38, m=38$$

Dabei ist $\log(ED_{50})$ die vorhergesagte biologische Aktivität, m die Gesamtzahl der untersuchten Moleküle. Dieses Modell zeigt gute interne statistische Güteparameter, die externe Vorhersagekraft ist jedoch relativ schlecht. Moleküle deren vorhergesagter $\log(ED_{50})$ -Wert um mehr als das dreifache der Standardabweichung der Vorhersagen abwich, wurden in einem zweiten Schritt als so genannte Ausreißer (engl. *Outlier*) aus der Modellberechnung herausgenommen. In diesem Falle waren es insgesamt 4: Die Moleküle 5, 7, 8 und 27. Nach einer erneuten Modellberechnung kam es zu folgendem Ergebnis:

$$\begin{aligned} \log(ED_{50}) = & 0.49228 \cdot AA_1 + 0.024094 \cdot HH_5 - 0.0039948 \cdot HH_{14} \\ & - 0.00038858 \cdot LwH_5 + 0.0055544 \cdot LsH_4 \\ & + 0.0027614 \cdot LsH_{14} + 1.2164 \end{aligned}$$

Gleichung 39

$$R^2_{Test,Avg}=0.57, RMSEP_{Test,Avg}=0.54$$

$$R^2_{CV-50\%}=0.81, RMSEP_{CV-50\%}=0.38, m=34$$

Die externe Vorhersagekraft des Modells steigt deutlich an, wenn Ausreißer entfernt werden. Dies ist analog zur Originalpublikation von Hopfinger, bei der ebenfalls 4 Ausreißer entfernt werden mussten [15]. Zusätzlich wurde ein Modell mit Omega-Konformeren als Startpunkt gerechnet. Auch hier wurde zunächst der komplette Datensatz modelliert:

$$\begin{aligned} \log(ED_{50}) = & -0.060123 \cdot HH_3 - 0.0007616 \cdot HH_{14} - 0.0036338 \cdot LwD_{10} \\ & + 0.0056814 \cdot LwD_{12} + 0.0051652 \cdot LsD_8 - 0.0030602 \cdot LsD_9 \\ & + 1.0852 \end{aligned}$$

Gleichung 40

$$R^2_{Test,Avg}=0.26, RMSEP_{Test,Avg}=0.79$$

$$R^2_{CV-50\%}=0.59, RMSEP_{CV-50\%}=0.59, m=38$$

Hier liegt ebenfalls eine schlechte externe Vorhersagekraft des Modells vor. Deswegen wurden mit Hilfe der oben genannten Regel zwei Ausreißer entfernt. In diesem Falle waren es die Moleküle 19 und 36. Eine erneute Modellberechnung führte zu folgendem Ergebnis:

$$\begin{aligned} \log(ED_{50}) = & 0.00086358 \cdot HA_5 - 0.00078287 \cdot HD_6 - 0.00065189 \cdot HH_{13} \\ & - 0.0016266 \cdot LwD_8 - 0.00087549 \cdot LwD_{10} \\ & - 0.00037165 \cdot LwLw_7 + 1.1261 \end{aligned}$$

Gleichung 41

$$R^2_{Test,Avg}=0.52, RMSEP_{Test,Avg}=0.66$$

$$R^2_{CV-50\%}=0.70, RMSEP_{CV-50\%}=0.51, m=36$$

Dieses Modell ist relativ gut. Im Folgenden wird das Omega-Modell im Detail diskutiert, da hier nur zwei Ausreißer entfernt werden mussten. Dieses Modell kann folglich einen größeren Bereich der Daten beschreiben als das Catalyst-basierte Modell.

Die Interpretation ist hier einfach und offensichtlich, obwohl einzelne Variablen nie mehr als 20% der Variabilität der Daten erklären können. Variable HH_{14} ist Bestandteil von drei der hier vorgestellten Modelle. HH_{14} und LsD_8 werden im Detail interpretiert, da sie den für die Aktivität entscheidenden Rest des Moleküls beschreiben. Variable HH_{14} wird inkrementiert, wenn die ω -Kette keinen stark lipophilen Terminus hat. Das hat zur Folge dass der Schwerpunkt des hydrophilen Patches Richtung Ende der Seitenkette verschoben wird. Dieser

Unterschied ist in Abbildung 53 in der oberen Hälfte zu sehen (Teil A versus B). Hohe Werte für HH_{14} führen zu einer niedrigeren vorhergesagten biologischen Aktivität, da der Regressionskoeffizient ein negatives Vorzeichen zeigt. Daher kodiert diese Variable indirekt die Notwendigkeit eines starken lipophilen Terminus wie beispielsweise einen aromatischen Rest an der ω -Kette des Moleküls. Das wird deutlich, wenn man eine stark aktive Verbindung im Vergleich mit einer schwach aktiven ohne den aromatischen Rest betrachtet (siehe Abbildung 53, Bereich A versus B).

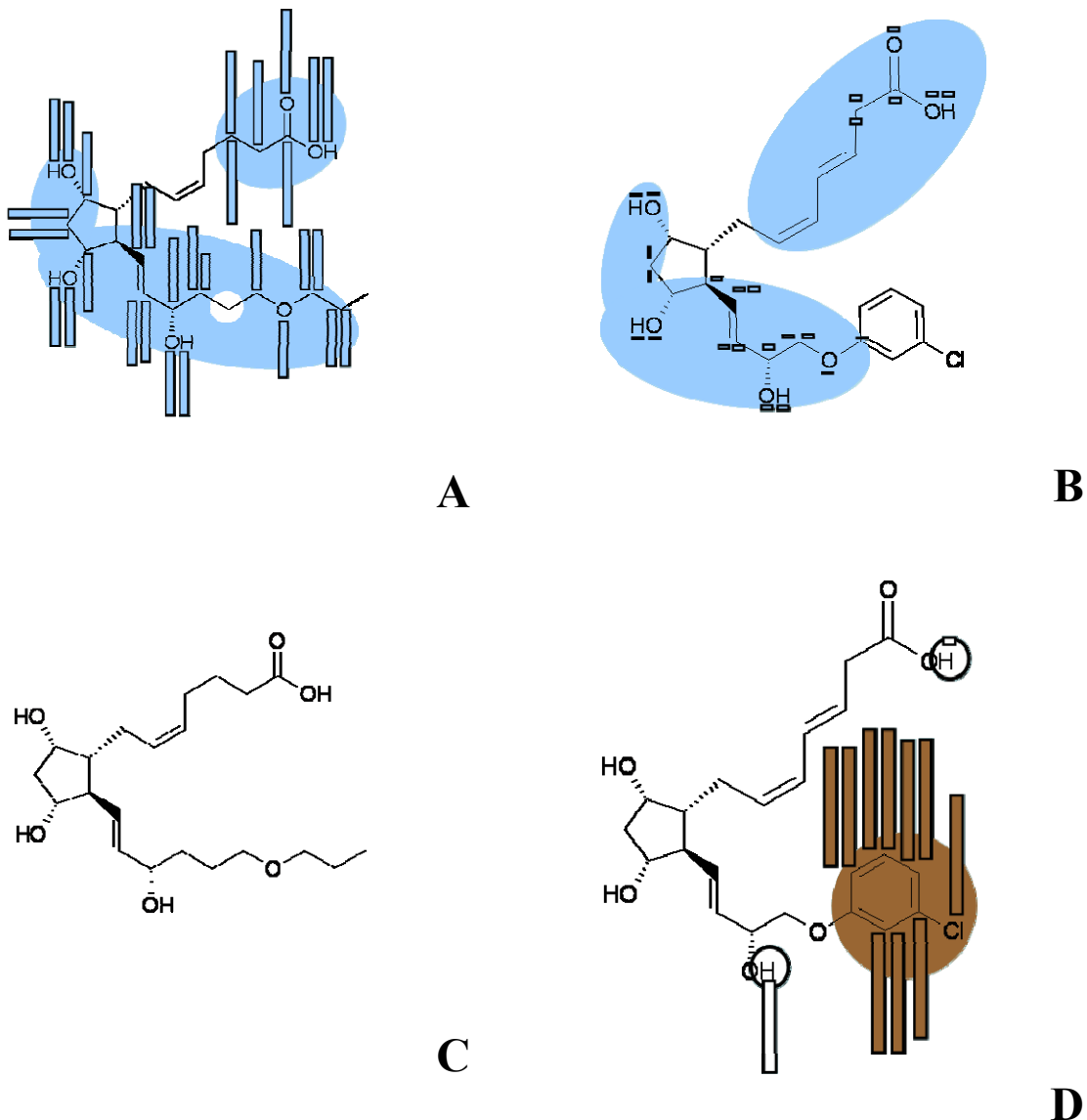


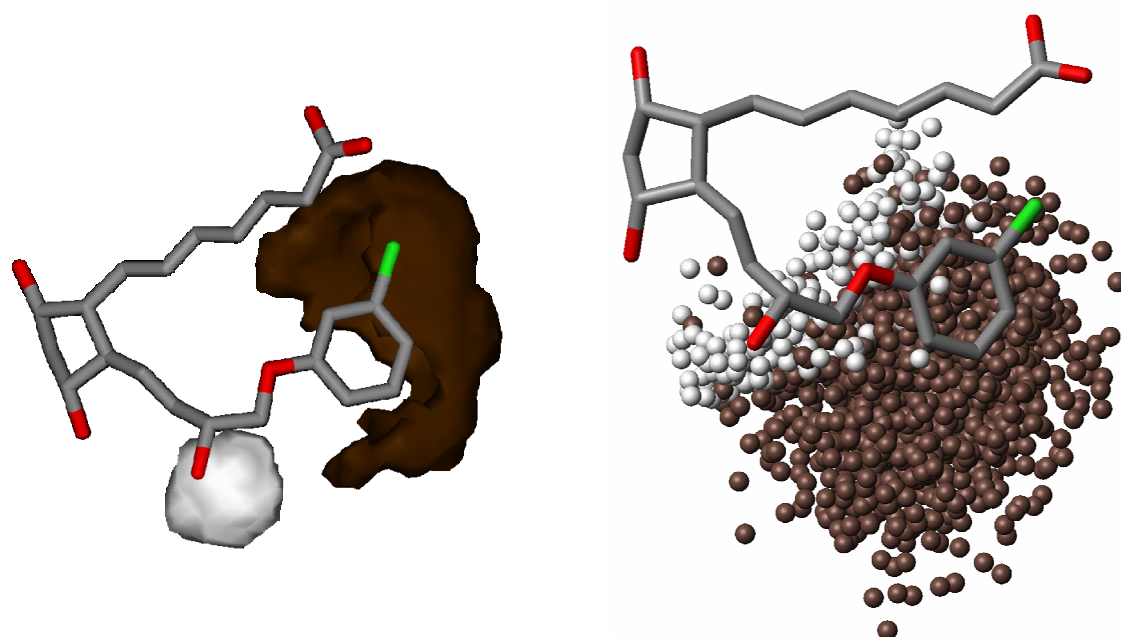
Abbildung 53:

Die 2D-Rückprojektion der Variablen HH_{14} und LSD_8 auf zwei Referenzmoleküle des $PGF_{2\alpha}$ -Datensatzes. Links ist eine schwach aktive Verbindung (A und C, Nr. 14) zu sehen, die in der ω -Kette einen großen hydrophilen Patch ausbildet. Dies führt dazu, dass die negativ mit der Aktivität korrelierte Variable HH_{14} große Werte annimmt (hohe blaue Balken auf den betroffenen Atomen). Die rechts gezeigte Verbindung (B und D, Nr. 24) dagegen hat einen kleineren hydrophilen Patch, da sie an der ω -Kette einen stark lipophilen Rest aufweist. Demzufolge hat für dieses Molekül HH_{14} niedrige Werte, LSD_8 dagegen sehr hohe (D).

Dagegen ist Variable LsD_8 positiv mit der biologischen Aktivität korreliert. Auch LsD_8 beschreibt die Wichtigkeit für bestimmte strukturelle Gegebenheiten in der ω -Kette. Eine Rückprojektion dieser Variable auf ein repräsentatives Konformer ist in Abbildung 54 gezeigt. LsD_8 beschreibt die Distanz von der Hydroxyl-Funktion in der unteren Seitenkette zum Terminus dieser Kette. Die Hydroxyl-Funktion wird dabei als Wasserstoffbrücken-Donor kodiert, der aromatische Terminus als stark lipophiler Bereich. Die Distanz zwischen den Schwerpunkten des Donor-Areals und des stark lipophilen Areals ist im Optimalfall etwa 8 Å. Eine Änderung dieser Distanz (also eine Verlängerung oder Verkürzung) führt zu einer niedrigeren vorhergesagten Aktivität. Damit wird implizit kodiert, wie flexibel die ω -Kette sein muss, um eine optimale Aktivität zu erreichen. Diese Vorhersagen sind analog zu den tatsächlichen Gegebenheiten bei den biologischen Aktivitäten. Die negativ korrelierte Variable LwD_{10} bestätigt die Interpretation der Variablen LsD_8 . Sie hat als Ausgangspunkt den schwach lipophilen Bereich des Aromaten und als Endpunkt die Hydroxyl-Funktion in der unteren Seitenkette. Die Variable LwD_{10} erhält einen höheren Eintrag, wenn die ω -Kette lang ist. Beide Variablen beschreiben die optimale Position des terminalen Aromaten. Offensichtlich ist genau der terminale Aromat entscheidend für eine gute biologische Aktivität. Die Variable LsD_8 wird nur von Molekülen erfüllt, die einen entsprechenden Rest haben.

Interessanterweise kodiert das Catalyst-Modell die gleiche Information auf andere Art und Weise. Die Variable LsH_{14} hat ihren Ursprung im stark lipophilen Bereich und reicht zum hydrophil kodierten Grundgerüst. Die Aussage bleibt die Gleiche. Dies wird durch die Redundanz des Deskriptors ermöglicht. Somit wird auch im Catalyst-Modell die Position des Substituenten an der ω -Kette kodiert.

Die negativ korrelierten LwD -Variablen im Omega-Modell beschreiben einerseits die optimale Länge der α -Kette der Moleküle sowie andererseits auch den Substituenten an dem angesprochenen aromatischen Rest. Da eine Veränderung an der α -Kette bei den untersuchten Molekülen nur geringen Einfluss auf die biologische Aktivität hat, wird auf diese beiden Variablen nicht genauer eingegangen.



A

B

Abbildung 54: Links: 3D-Rückprojektion der Variable LsD_8 auf ein repräsentatives Konformer der Verbindung 24 des $PGF_{2\alpha}$ -Datensatzes; Rechts: Dreidimensionale Rückprojektion von LsD_8 die die Position über alle Konformere aller für die Patchbildung relevanten Atome beschreibt.

Abbildung 55 zeigt die kreuzvalidierten internen Vorhersagen und die Testdatensatzvorhersagen für die Daten des Omega-Modells. Hier sind die Testdatensatzvorhersagen wieder die Durchschnittswerte über alle Vorhersagen, bei denen das betreffende Molekül aus dem Trainingsdatensatz entfernt wurde. Die Abweichungen in diesen Testdatenvorhersagen sind manchmal enorm, was sich in dem relativ niedrigen $R^2_{Test,Avg}$ -Wert widerspiegelt. Berücksichtigt man die starke Flexibilität der Moleküle, so ist dieses Ergebnis trotzdem zufrieden stellend.

Tabelle 10: Überblick über die statistischen Güteparameter von Modellen für den $PGF_{2\alpha}$ -Datensatz, die mit anderen Techniken erzeugt wurden. Es sind die jeweils besten Daten aus den entsprechenden Publikationen gezeigt [15,190].

QSAR-Methode	R^2	R^2_{CV-1}	$R^2_{CV-50\%}$	R^2_{Test}
Hopfinger [15]	0.86	0.74	-	-
Martinek [190]	0.58	0.54	0.42	-
xMaP	0.91	0.87	0.83	0.57

Um das xMaP-Modell mit den besten Modellen von Hopfinger und Martinek zu vergleichen, werden die dazugehörigen statistischen Güteparameter mit angegeben: Hopfinger's 4D-QSAR liefert mit Variablenselektion und PLS einen R^2 -Wert von 0.86 sowie einen R^2_{CV-1} -Wert von 0.74 [15]. Weitere Parameter wurden nicht bestimmt. Martinek mit der mehrstufigen Selektionsprozedur und anschließender Variablenselektion mit PCR als

Regressionstechnik kommt ebenfalls zu schlechteren Werten. Diese sind im Überblick in Tabelle 10 gezeigt. Es muss betont werden, dass sowohl Hopfinger als auch Martinek viele Modelle von deutlich schlechterer statistischer Qualität generiert haben. Aufgrund der verschiedenen verwendeten Selektionstechniken sind die statistischen Güteparameter nicht direkt vergleichbar. Das xMaP-Modell hat eine höhere externe Vorhersagekraft als Martinek's interne Vorhersagekraft. Der $R^2_{CV-50\%}$ ist mit Hopfinger's R^2_{CV-1} zu vergleichen.

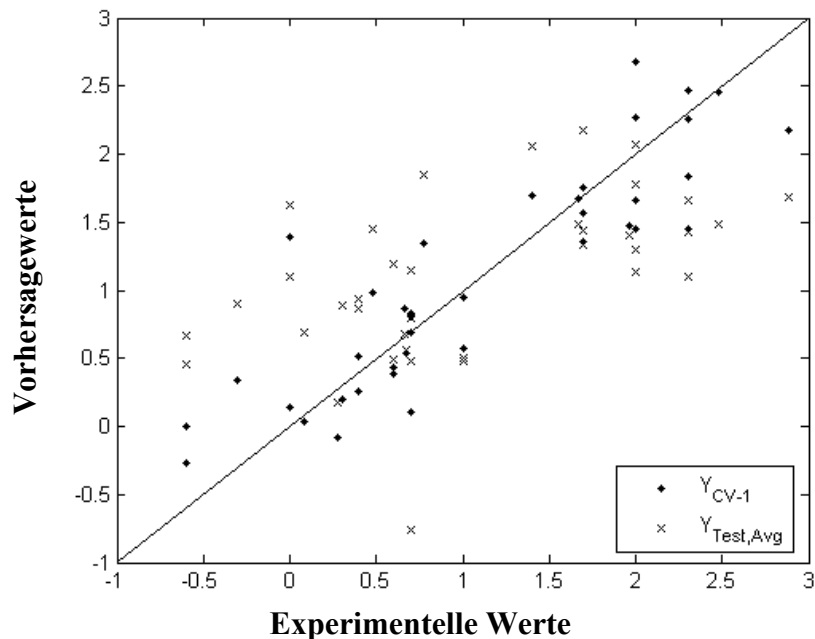


Abbildung 55: Der Vergleich der experimentellen Werte mit den vorhergesagten Werten für den $PGF_{2\alpha}$ -Datensatz bezogen auf Trainings- und Testdatenvorhersage ($R^2_{CV-1}=0.68$, $R^2_{Test,Avg} = 0.52$).

Zudem ist die Interpretation des Modells für den $PGF_{2\alpha}$ -Datensatz sehr geradlinig und liefert einen guten Einblick in die Struktur-Wirkungs-Beziehungen.

3.9.3. Modulatoren des muskarinischen M₂-Rezeptors (M₂)

Dieser Datensatz wird genutzt, um die Unterschiede zwischen xMaP und seiner Vorgängertechnik MaP zu illustrieren. Aufgrund der hohen Flexibilität der hier untersuchten Verbindungen ist xMaP ideal geeignet. Die Standardprozedur mit Catalyst führte im Mittel zu 230 Konformeren (Median: 252). Das Minimum liegt bei 104 Konformeren für ein Molekül und das Maximum bei 254. Diese Werte unterstreichen die hohe Flexibilität der untersuchten Substanzen. Folgendes Modell wurde an Hand dieser Daten berechnet:

$$\hat{Y} = 0.00011425 \cdot LwH_5 + 0.0088021 \cdot LwLw_{27} + 0.00092686 \cdot HA_6 \\ + 0.00086892 \cdot LsH_{14} + 0.00038871 \cdot LsLw_{16} + 5.8064$$

Gleichung 42

$$R^2_{Test,Avg}=0.64, RMSEP_{Test,Avg}=0.54$$

$$R^2_{CV-50\%}=0.78, RMSEP_{CV-50\%}=0.42, m=44$$

Dabei ist \hat{Y} die biologische Aktivität der untersuchten Moleküle und m deren Anzahl. Dieses Modell hat gute statistische Güteparameter. Auch hier wurde ein Modell mit Omega-Konformeren als Grundlage erstellt. Es liefert eine ähnliche Aussage mit ähnlichen statistischen Güteparametern:

$$\hat{Y} = 0.00043362 \cdot HA_5 + 0.35125 \cdot LwLw_{36} + 0.0012994 \cdot LsH_{14} \\ + 0.0033982 \cdot LsLw_{24} + 0.22716 \cdot LsLw_{37} + 5.8064$$

Gleichung 43

$$R^2_{Test,Avg}=0.69, RMSEP_{Test,Avg}=0.50$$

$$R^2_{CV-50\%}=0.82, RMSEP_{CV-50\%}=0.38, m=44$$

Interpretiert man die hier als wichtig identifizierten Variablen, so zeigt sich, dass zwei zentrale Dinge kodiert werden. Das sind ein lipophiler Terminus und das Substitutionsmuster dieses Terminus. Der hydrophile Bereich um die protonierten Stickstoffatome dient als Ausgangspunkt für die beiden wichtigsten Variablen LsH_{14} und HA_5 . Diese sind damit an dieser Stelle verankert. Die Stickstoffatome sind essenziell, d.h. nur bei ihrem Vorhandensein findet eine Interaktion statt. Die Unterschiede, die eine unterschiedlich starke Interaktion mit dem Rezeptor bedingen finden sich in den Zielbereichen der Variablen. Sie sind in den lipophilen Bereichen, sowie bei den Wasserstoffbrückenakzeptoren zu finden. Die 3D-Rückprojektion zu diesen beiden Variablen ist vereinheitlicht in Abbildung 56 gezeigt. Als Beispiel sind in Abbildung 57 zwei Strukturen gezeigt, deren Hauptunterschied die Existenz eines annelierten Phenylrestes ist. Wenn dieser Rest nicht vorhanden ist sinkt die Aktivität rapide ab (von 7.34 auf 3.82). Für die in Abbildung 57 B gezeigte Struktur erfolgt überhaupt

kein Inkrement in Variable LsH_{14} . Das zeigt, dass der lipophile Rest entscheidend für die Aktivität ist.

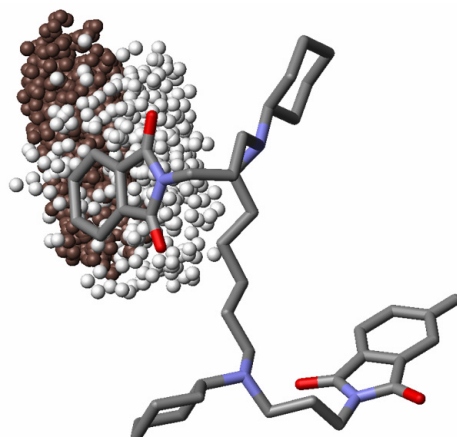
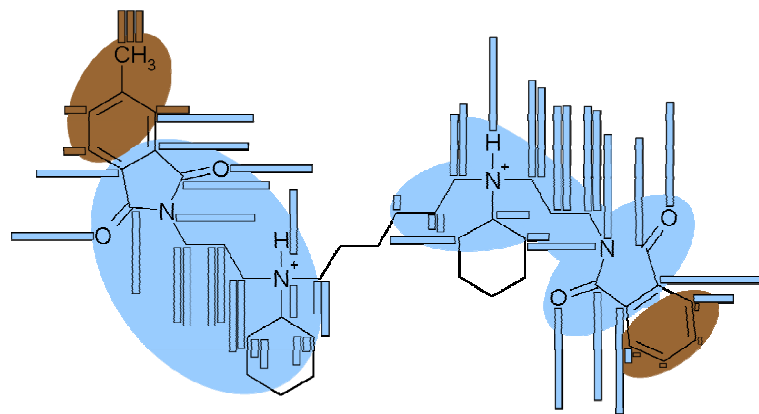


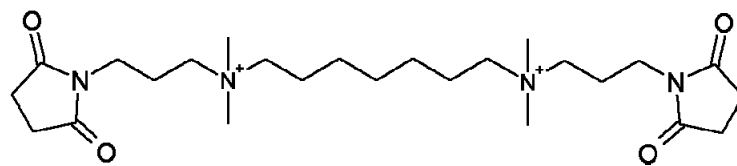
Abbildung 56: Rückprojektion der wichtigen Atome. Hier ist zu berücksichtigen, dass aufgrund der hohen Flexibilität dieser Moleküle der Kabsch-Algorithmus [161] nur auf diese angewendet wurde. Der jeweilige hydrophile Beitrag zu LsH_{14} und $HA_{5/6}$ ist aus Gründen der Übersichtlichkeit nicht gezeigt.

Variable LsH_{14} kann in etwa 60% der Variabilität der Daten erklären: $R^2=0.64$, $R^2_{CV-I}=0.61$, $R^2_{CV-50\%}=0.41$. Zusätzlich sind an jedem Rest zwei Carbonylfunktionen sehr wichtig, die sowohl im Catalyst- als auch im Omega-Modell in den HA-Variablen kodiert werden. Dieses Ergebnis steht in sehr gutem Einklang mit den Ergebnissen der MaP- und GRID/PLS-Analysen[14,19]. Interpretiert man die beiden erwähnten Variablen zusammen, so zeigt sich, dass das Substitutionsmuster der terminalen Reste beschrieben wird. Die zugehörigen Rückprojektionen sind in der dreidimensionalen Version in Abbildung 56 und als zweidimensionale Rückprojektion in Abbildung 57 gezeigt. Abbildung 56 zeigt eine Darstellung ohne den hydrophilen Bereich, der für beide Variablen identisch ist. Es wird ersichtlich, welche Parameter für eine biologische Aktivität wichtig sind. Es ist aus theoretischen Untersuchungen und Mutationsstudien bekannt, dass die Termini mit einem Tryptophan und einem Tyrosin am M_2 -Rezeptor durch π - π -Wechselwirkungen und Wasserstoffbrücken interagieren [246,247]. Für eine optimale Interaktion sind ein aromatischer Bereich und ein Wasserstoffbrückenakzeptor erforderlich. Das Ergebnis der QSAR-Analyse steht im Einklang mit diesen strukturellen Voraussetzungen.

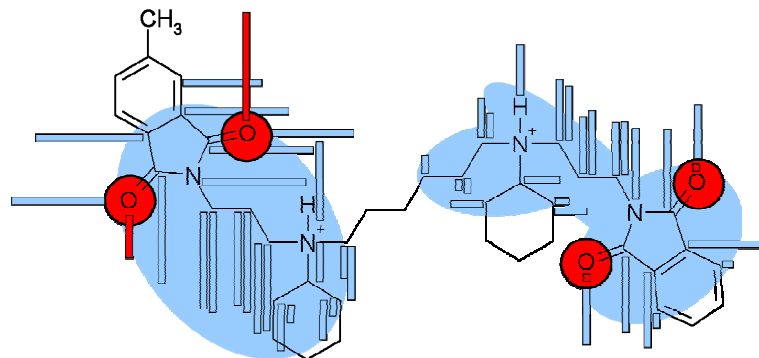
Die übrigen Variablen beschreiben die optimale Distanz zwischen den lipophilen Bereichen und kodieren damit die Kettenlänge des Gesamtmoleküls.



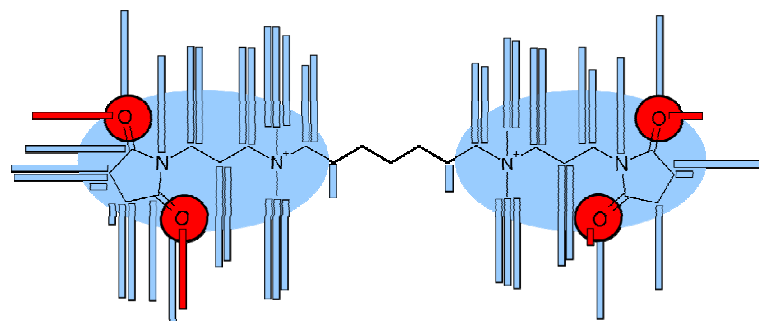
A



B



C



D

Abbildung 57: 2D-Rückprojektion der relevanten Variablen LsH_{14} (oben) und HA_6 (unten) auf die am schwächsten aktive (B und D, Nr. 22) und am stärksten aktive (A und C, Nr. 36) Substanz des M_2 -Datensatzes. Die große Bedeutung des terminalen lipophilen Restes mit seinem Substitutionsmuster wird offensichtlich.

Analog zum AZT-Datensatz wurde ein weiteres Modell mit einer modifizierten Parametrisierung der protonierten tertiären und quartären Stickstoffe erstellt (Werte für die Hydrophilie siehe Kapitel 3.9.1). Dafür wurden die Catalyst-Konformerer als Grundlage verwendet. Die Bedingung, dass sich ein Modell durch die Hereinnahme einer weiteren Variablen um mindestens drei Prozent verbessern muss, wurde hier außer Acht gelassen, da sonst nur zwei Variablen ausgewählt werden und das Gesamtmodell sehr schlecht wird:

$$\hat{Y} = -0.0344 \cdot AA_{24} + 0.0202 \cdot HH_3 + 0.015 \cdot HH_5 \\ + 0.0068 \cdot LwLw_5 - 0.0011 \cdot LsA_8 + 0.001 \cdot LsA_{28} \\ + 0.0001 \cdot LsH_{11} + 5.8064$$

Gleichung 44

$$R^2_{Test,Avg}=0.23, RMSEP_{Test,Avg}=0.78$$

$$R^2_{CV-50\%}=0.46, RMSEP_{CV-50\%}=0.65, m=44$$

Dieses Modell ist in seinen statistischen Güteparametern dem ersten beschriebenen Modell deutlich unterlegen (vgl. Abb. 58 und Formel 42). Die Interpretation beschreibt vorwiegend die Kettenlänge zwischen den protonierten Stickstoffen (vgl. Abb. 57). Die Variablen HH_3 , HH_5 und $LwLw_5$ beschreiben dies über verschiedene Eigenschaftszuweisungen zu Arealen um eben diese Stickstoffe. Die übrigen Variablen beschreiben wie im anfangs beschriebenen Modell das Substitutionsmuster der Molekül-Termini.

Keine der Variablen im Modell erklärt mehr als rund 25% der Gesamtvariabilität in den Daten, am meisten erklären die HH_5 sowie die $LwLw_5$. In den Bagging-Läufen wird HH_5 am häufigsten selektiert. Auch Variable $LwLw_4$ wird ähnlich oft gewählt. Diese ist stark mit $LwLw_5$ korreliert, der Korrelationskoeffizient liegt bei 0.73. Zusammenfassend liefert hier die Interpretation des Gesamtmodells ein Bild von der Struktur eines guten M_2 -Modulators, die statistische Modellqualität ist jedoch schlecht.

Bei dem M_2 -Datensatz ist demzufolge eine veränderte Parametrisierung der Hydrophilie um den quartären Stickstoff in der statistischen Modellqualität von Vorteil. Dies ist ein weiterer Hinweis darauf, dass die veränderte Parametrisierung der Hydrophilie bei der Erstellung von Modellen für Moleküle mit positiv geladenen Stickstoffatomen von Vorteil sein kann und evaluiert werden sollte.

Zusammenfassend betrachtet ist xMaP hier in der Lage, einen tieferen Einblick in die Struktur-Wirkungsbeziehungen und Ideen für mögliche Modifikationen zu liefern. Beispielsweise könnte für Verbindung 22 eine lipophile Substitution am Phtalimidrest angeraten werden.

Tabelle 11: Überblick über die statistischen Güteparameter von den Modellen für den M_2 -Datensatz, die mit 3D-QSAR-Techniken erzeugt wurden. Der hier gezeigte R^2_{Test} -Wert beruht im Gegensatz zu xMaP auf nur einer einzigen Unterteilung in Trainings- und Testdatensatz und somit nur einer einzigen Vorhersage. Es sind Daten des jeweiligen Hand-Alignments gezeigt [19,80]. Bei xMaP sind die Güteparameter des Omega-Modells gezeigt.

QSAR-Methode	R^2	R^2_{CV-1}	$R^2_{CV-50\%}$	R^2_{Test}
MaP [19,80]	0.87	0.85	0.84	0.75
GRID/PLS [19,80]	0.99	0.67	0.48	0.87
CoMFA [19,80]	0.95	0.52	-	0.80
CoMSIA [19,80]	0.96	0.56	-	0.85
xMaP [19,80]	0.89	0.87	0.82	0.69

xMaP liefert ein Modell, das von der Aussage und Interpretierbarkeit her sehr ähnlich zu dem MaP-Modell ist. Variable LsH_{14} ist auch Bestandteil des MaP-Modells. Das xMaP-Modell beruht jedoch auf dem gesamten Konformerensemble für jedes Molekül im Datensatz. Die statistischen Güteparameter liegen in derselben Größenordnung. Der Vergleich mit den Ergebnissen weiterer QSAR-Techniken (Details siehe Tabelle 11) zeigt, dass die Modellqualitäten in allen Bereichen vergleichbar sind. Der R^2_{Test} -Wert für xMaP erscheint niedriger, was aber auf die verwendete, strengere Art der Validierung zurückzuführen ist. Dass er von guter Qualität ist zeigt sich in Abbildung 58.

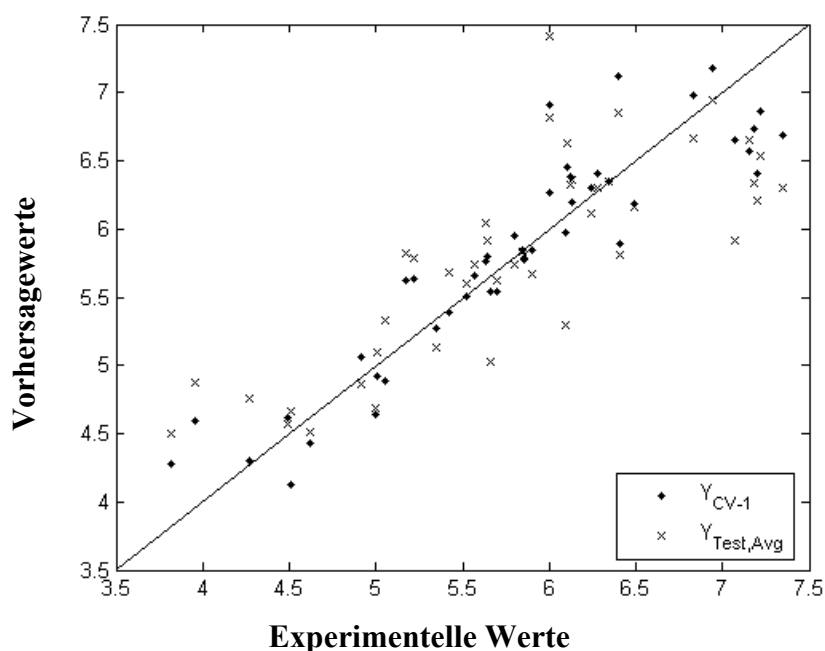


Abbildung 58: Externe vs. interne Vorhersagekraft für den M_2 -Datensatz ($R^2_{CV-1}=0.89$; $R^2_{Test,Avg}=0.69$).

Der zentrale Punkt der QSAR-Analysen ist die Interpretierbarkeit der Modelle. Deshalb kann hier zu Recht behauptet werden, dass xMaP den anderen Techniken deutlich überlegen ist. Die Schritte des Alignments und der Konformerenauswahl entfallen. In Analogie zu MaP

sind die Modelle deutlich einfacher interpretierbar als dies bei CoMFA und CoMSIA der Fall ist. Diese beiden Techniken kodieren zudem den aromatischen Rest nur über seinen sterischen Einfluss, die Position des Carbonyl-Sauerstoffs wird völlig ignoriert.

3.9.4. Inhibitoren der HIV-1 Reversen Transkriptase (HEPT)

Der Datensatz besteht aus 80 HEPT-Derivaten, die die HIV-1 Reverse Transkriptase inhibieren (siehe Kapitel 3.8.4.). Molekül Nr. 34 (siehe Anhang II.4.) ist in vielen Modellen als Ausreißer identifiziert worden. Aus diesem Grund wird dieses Molekül auch hier nicht in die Analyse mit einbezogen [201]. Mit den verbleibenden 79 Strukturen wurde das folgende xMaP-Modell berechnet:

$$\log\left(\frac{I}{C}\right) = -0.0091 \cdot AA_7 - 0.0019 \cdot LwH_{10} + 0.0012 \cdot LsLw_1 + 6.3127$$

Gleichung 45

$$R^2_{Test,Avg} = 0.47, RMSEP_{Test,Avg} = 0.94$$

$$R^2_{CV-50\%} = 0.51, RMSEP_{CV-50\%} = 0.93, m = 79$$

Die statistische Modellqualität ist hier relativ schlecht. Es muss aber betont werden, dass nur 3 Variablen dazu verwendet werden, die Variabilität von allen 79 Substanzen zu erklären. Eine Entfernung von Ausreißern brachte hier keine Verbesserung der Modellgüte. Die negativ korrelierte Variable AA_7 erklärt den größten Teil der Variabilität in den Daten ($R^2 = 0.44$, $R^2_{CV-50\%} = 0.42$, $R^2_{CV-50\%} = 0.40$) und soll deshalb im Folgenden im Detail besprochen werden:

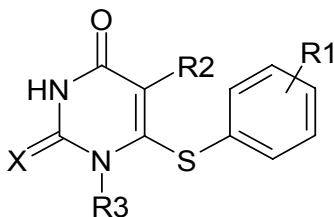


Abbildung 59: Die Grundstruktur der HEPT-Derivate. Die Substitution an R3 hat den stärksten Einfluss auf die biologische Aktivität

Im Wesentlichen werden über Variable AA_7 Moleküle diskriminiert, die einen Rest mit endständiger Akzeptorfunktionalität in Position R3 (siehe Abbildung 59) tragen. Dies ist bei den Verbindungen mit endständiger Hydroxyl-Funktion der Fall (siehe Anhang II.4.). Ein Sauerstoff-Substituent in Position X dient dabei als fester Startpunkt. Auch dieser Rest wird als Wasserstoffbrücken-Akzeptor kodiert. Wenn die Distanz zwischen den Substituenten an R3 und X Variable AA_7 erfüllt, dann wird eine schlechtere biologische Aktivität vorhergesagt. Dies ist beispielsweise bei den Molekülen 2 und 9 der Fall. Diese weisen einen niedrigen Wert für $\log(1/C)$ auf. Diese beiden Moleküle sind im oberen Teil von Abbildung 60 gezeigt.

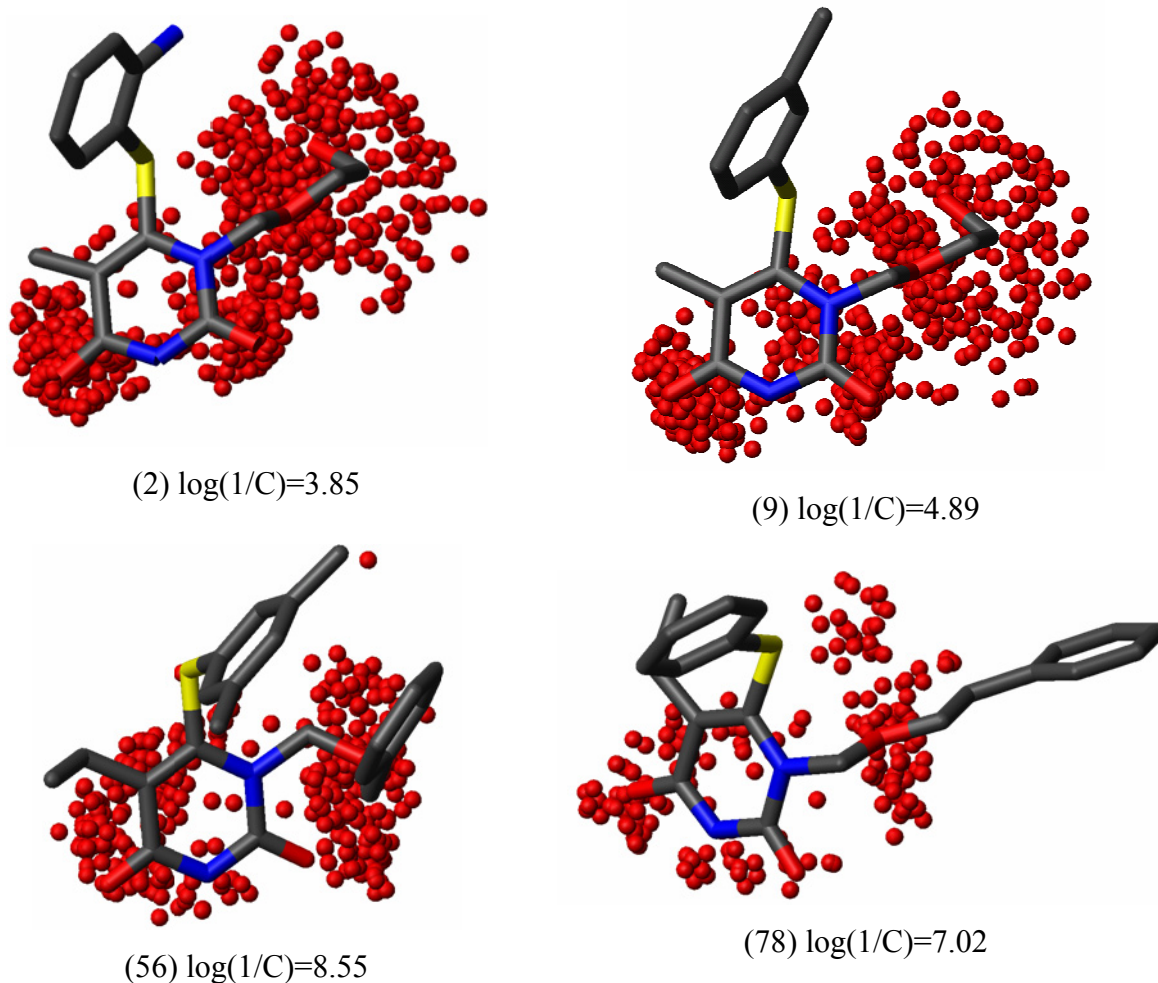


Abbildung 60:

Die Rückprojektion der negativ korrelierten Variable AA_7 auf 4 ausgewählte Strukturen des HEPT-Datensatzes: Die Strukturen mit einem niedrigen pIC_{50} -Wert zeigen diesen potenziellen Zweipunkt-Pharmakophor, so dass ein niedrigerer Wert für die biologische Aktivität vorhergesagt wird. Umgekehrt ist es bei den Strukturen mit hohem pIC_{50} : Diese haben einen lipophilen Rest an der R3-Position. Das hat zur Folge, dass Variable AA_7 weniger stark inkrementiert wird. Der vorhergesagte Wert für die biologische Aktivität wird höher. Ein aromatischer Rest an R3 ist demzufolge von Vorteil für die biologische Aktivität.

Hat dieser R3-Rest einen endständigen lipophilen Terminus, so ist das Inkrement für Variable AA_7 wesentlich geringer. Variable AA_7 beschreibt im Endeffekt die entscheidenden Charakteristiken des Datensatzes: Der Rest in Position R3 sollte keinen Akzeptor mit sich bringen, eine endständige Hydroxyl-Funktion ist schlecht für die biologische Aktivität. Zusätzlich wird AA_7 durch die Etherbindung in Position R3 erfüllt, dies hat aber keinen entscheidenden Einfluss auf die biologische Aktivität. Als größere Reste an R3 werden nur Aromaten toleriert. Dies wird auch direkt durch die Gegebenheiten in der Bindetasche untermauert, was bei dem strukturbasierten Modell detaillierter gezeigt ist. Der entsprechende Teil der Bindetasche ist relativ lipophil. Diese Interpretation wird zusätzlich noch durch Variable $LsLw_1$ unterstützt, die Aromaten in den Positionen R1 und R3 kodiert (siehe Abbildung 61).

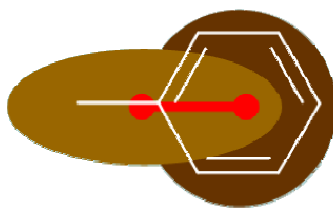


Abbildung 61: Beschreibung der Variable $LsLw_1$: Über ein stark und ein schwach lipophiles Areal wird ein endständiger Aromat kodiert.

Da $LsLw_1$ positiv mit der biologischen Aktivität korreliert, wird ein höherer Wert für $\log(1/C)$ vorhergesagt. Diese Interpretation deckt sich im Wesentlichen mit der von etablierten 3D-QSAR-Techniken [201]. Die Qualität des Modells ist im Vergleich von Trainings- und Testdatenvorhersage in Abbildung 62 herausgestellt. Man erkennt die hohe Streuung der vorhergesagten Werte. Zudem scheint es zu einer systematischen Unterschätzung der sehr aktiven Moleküle zu kommen (nur negative Abweichungen im Bereich $\log(1/C)$: 7-9). Beide Phänomene erklären die geringe statistische Modellqualität.

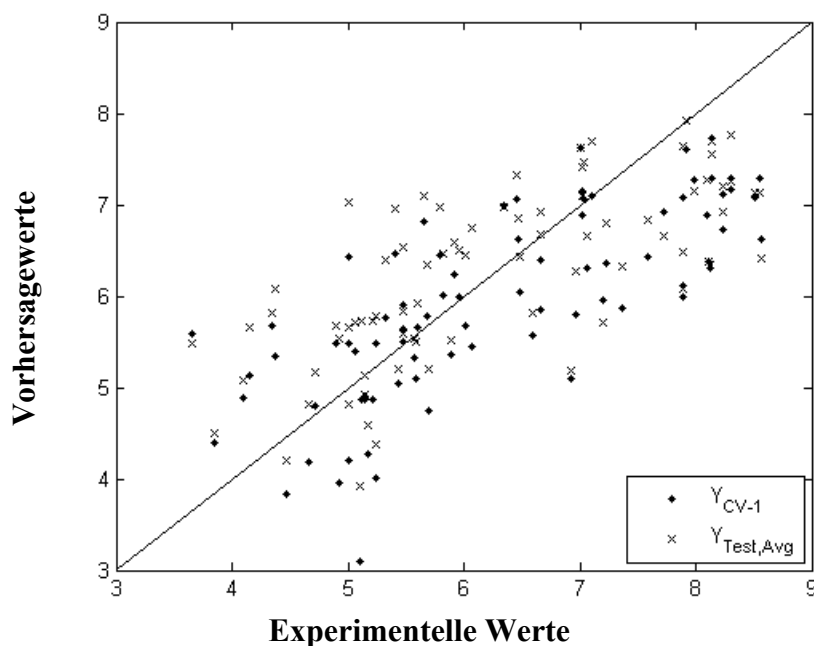


Abbildung 62: Externe vs. interne Vorhersagekraft für das ligandbasierte Modell des HEPT-Datensatzes ($R^2_{CV-1}=0.57$; $R^2_{Test,Avg}=0.47$).

Vergleicht man die xMaP-Modelle mit etablierten Techniken, so zeigt sich, dass diese eine ähnliche statistische Modellqualität (siehe Tabelle 12) erreichen und gut interpretiert werden können.

Tabelle 12: Überblick über den R^2_{CV-1} -Wert (und R^2_{Test} , insofern vorhanden) von Modellen für den HEPT-Datensatz, die mit 3D-QSAR-Techniken erzeugt wurden [201] im Vergleich mit den xMaP-Ergebnissen.

QSAR-Methode	R^2_{CV-1}	$R^2_{Test,Avg}$
CoMFA/std.	0.86	0.73
CoMFA/docking	0.58	-
CoMSIA/std.	0.78	0.65
CoMSIA/docking	0.53	-
xMaP	0.57	0.47
xMaP/Docking	0.75	0.52

Für CoMFA und CoMSIA liegen strukturbasierte Modelle vor [201], da bei diesem Datensatz die Kristallstruktur des Zielenzym bekannt ist. Die Kristallstruktur wurde genutzt, um basierend auf einer mittels Docking erzeugten Strukturfamilie folgendes Modell zu generieren:

$$\begin{aligned} \log\left(\frac{I}{C}\right) = & -0.0098 \cdot AA_6 + 7.4156 \cdot DA_{25} - 0.0015 \cdot HA_2 \\ & + 0.0017 \cdot HA_4 + 0.00033 \cdot LwH_2 + 0.0017 \cdot LsA_7 \\ & + 0.0004 \cdot LsLw_2 + 6.3127 \end{aligned} \quad \text{Gleichung 46}$$

$$R^2_{Test,Avg}=0.52, RMSEP_{Test,Avg}=0.92$$

$$R^2_{CV-50\%}=0.65, RMSEP_{CV-50\%}=0.78, m=79$$

Im Gegensatz zu den 3 MIVs des ligandbasierten Modells werden hier insgesamt 7 MIVs gewählt. Das hat eine höhere interne Modellqualität zur Folge. Die externe Vorhersagekraft bleibt jedoch mehr oder weniger identisch. Auch hier ist es eine AA-Variable, die den größten Teil der Variabilität erklärt. In diesem Falle ist es AA_6 mit $R^2=0.42$, $R^2_{CV-1}=0.40$ und $R^2_{CV-50\%}=0.38$. Auch diese ist negativ zu $\log(I/C)$ korreliert. Die Interpretation ist identisch zum ligandbasierten Modell. Dass hier eine geringere Distanz (AA_6 versus AA_7) ausgewählt wird liegt einerseits an der unscharfen Zählweise, sowie andererseits daran, dass die konformelle Flexibilität der Strukturen in der Bindetasche eingeschränkt ist. Dies betrifft insbesondere die Flexibilität in Position R3. Das strukturbasierte Modell kann also die Aussage des ligandbasierten einerseits verifizieren sowie andererseits verfeinern.

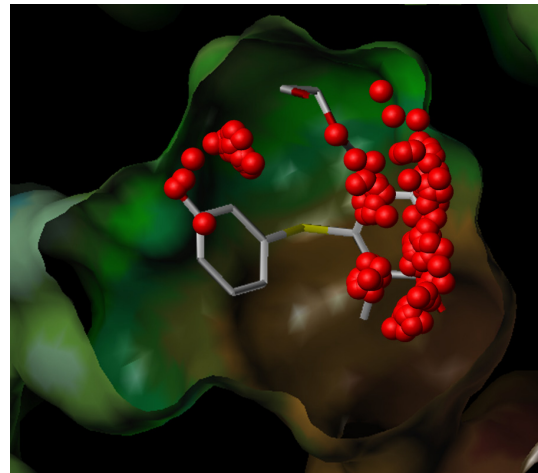
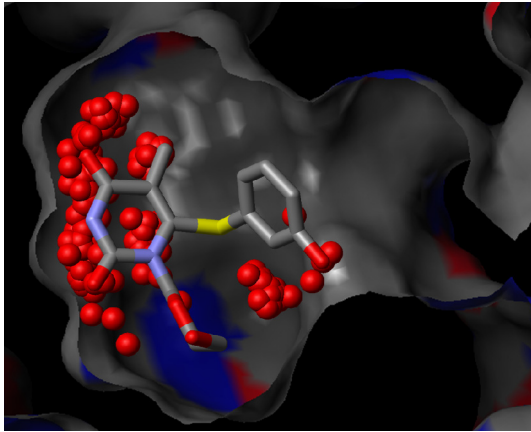


Abbildung 63: Die Rückprojektion der AA_6 -Variable in die Bindetasche (Molekül Nr. 13): Links ist die Wasserstoffbrückenbindungskapazität im Protein gezeigt, rechts die Lipophilie. Die Darstellung erfolgt von unterschiedlichen Seiten, was der Übersichtlichkeit dienen soll. Es zeigt sich, dass die Bereiche, die für eine Erhöhung von Variable AA_6 verantwortlich sind bei dem gezeigten Molekül in Bereichen des Proteins zu liegen kommen, die nicht komplementär zu den Moleküleigenschaften sind, das Molekül hat eine niedrige biologische Aktivität ($pIC_{50}=4.09$).

Die gefundenen Eigenschaften lassen sich in die Bindetasche zurückprojizieren und ergeben ein komplementäres Bild zu den Eigenschaften der Bindetasche: xMaP identifiziert hier die für die Interaktion relevanten Parameter. Details dazu sind in Abbildung 63 für die Variable AA_6 und in Abbildung 64 für die Variable LsA_7 gezeigt. Diese beschreibt den Rest in Position R1 mit den Akzeptoren (Sauerstoff, Rest in Position 1) am zentralen Ring als Startpunkt.

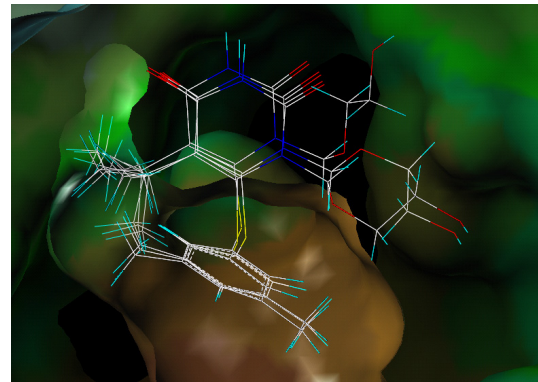
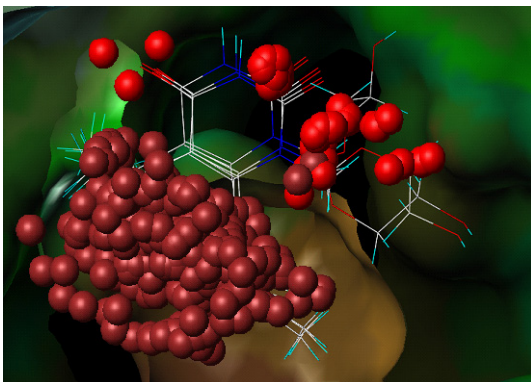


Abbildung 64: Die Rückprojektion der Variable LsA_7 in die Bindetasche der HIV 1-RT. Links: xMaP-Projektion. Rechts: Gedockte Strukturen. Man erkennt die Komplementarität der identifizierten Eigenschaften der gedockten Substanzen zu denen der Bindetasche.

Alle anderen Variablen werden nicht im Detail diskutiert, da sie kaum zusätzliche wertvolle Information liefern. Sie beschreiben Unterschiede zwischen einzelnen Molekülen, sind also für die Feinjustierung des Modells zuständig. Die Darstellung der Testdatenvorhersage in Abbildung 65 zeigt die Güte des strukturbasierten Modells, die ähnlich dem ligandbasierten Modell ist. Erneut werden hoch aktive Substanzen systematisch unterschätzt (siehe Bereich $\log(1/C)$: 7-9)

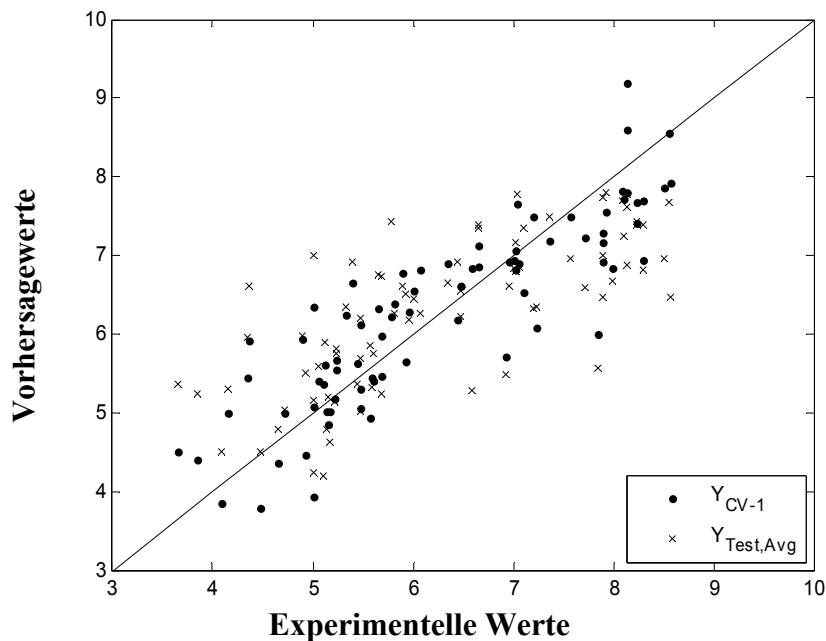


Abbildung 65: Externe vs. interne Vorhersagekraft für das strukturbasierte Modell ($R^2_{CV-1}=0.75$; $R^2_{Test,Avg}=0.52$).

Als Fazit für den HEPT-Datensatz kann gesagt werden, dass eine Kombination von ligand- und strukturbasiertem Modell einen weiteren Hinweis liefert, dass xMaP etablierten Techniken nicht unterlegen ist und wertvolle Einblicke in die Struktur-Wirkungs-Beziehungen erlaubt. Überdies ist festzuhalten, dass die von xMaP identifizierten Strukturparameter komplementär zu den Gegebenheiten in der Bindetasche sind. Daher ist das Modell für den HEPT-Datensatz eine Validierung der xMaP-Technik dahingehend, dass die für die Ligand-Protein-Interaktion relevanten Parameter im Modell wiedergegeben werden. Die Informationen der Proteinstruktur validieren das Ergebnis der QSAR-Analyse.

3.9.5. Dopamin-Antagonisten (D₁)

3.9.5.1 Ein erstes QSAR-Modell

Für diesen Datensatz mit 77 Verbindungen, die antagonistische Aktivität am Dopamin D₁-Rezeptor zeigen, konnte bisher mit keiner Methode ein QSAR-Modell erstellt werden, da ein Alignment der Substanzen nicht möglich war. Erstmals konnte mit xMaP ein QSAR-Modell erstellt werden. Die Informationen dazu werden im Folgenden vorgestellt.

Die Verbindungen hatten im Mittelwert 66 Konformere (Median 51). Die Starrste Struktur zeigte nur 3 Konformere, die Flexibelste 186 Konformere. Die Begrenzung, dass sich in einem Schritt der Variablenselektion das Modell um mindestens 3% verbessern muss, wurde hier außer Acht gelassen. Die anfangs definierten Mindestanforderungen an Modelle wurden alle eingehalten. Mit diesen Voraussetzungen konnte eine gute statistische Qualität des Modells erreicht werden. Es ergab sich folgendes Modell:

$$\begin{aligned} pK_i = & -0.057 \cdot AA_3 - 0.0369 \cdot AA_{10} - 0.0309 \cdot HA_{11} \\ & - 0.0158 \cdot HA_{12} + 0.0196 \cdot HD_8 + 0.1617 \cdot HH_2 \\ & + 0.0034 \cdot LwA_8 - 0.0003 \cdot LwH_4 - 0.0618 \cdot LsA_1 \\ & - 0.0039 \cdot LsD_9 - 0.0017 \cdot LsH_4 + 0.02 \cdot LsLs_4 \\ & - 0.0017 \cdot LsLs_{13} + 7.0187 \end{aligned} \quad \text{Gleichung 47}$$

$$R^2_{Test,Avg}=0.47, RMSEP_{Test,Avg}=0.92$$

$$R^2_{CV-50\%}=0.64, RMSEP_{CV-50\%}=0.76, m=77$$

Diese Werte sind für ein erstes QSAR-Modell zufrieden stellend. Für die Interpretation muss die Entscheidung getroffen werden, welche Variablen dafür am sinnvollsten ausgewählt werden. Es fällt auf, dass keine Variable mehr als knapp 20% der Variabilität erklären kann. Variable HH_2 beschreibt knapp 18%, Variable HD_8 etwa 19%. Alle anderen Variablen liegen deutlich unter diesen Werten. Nimmt man die Ergebnisse der wiederholten Unterteilungen in Trainings- und Testdatensatz mit anschließenden Modellbildungen zusätzlich zur Entscheidungsfindung, so zeigt sich, dass die mit Abstand am häufigsten gewählte Variable HH_2 ist. Sie wird in 69 der 100 Läufe als wichtig eingestuft. Es muss also erklärt werden, welche Information diese Variable trägt. HH_2 ist positiv mit dem negativen dekadischen Logarithmus des K_i -Wertes korreliert. Dieser Wert wird hier als abhängige Variable zur Beschreibung der biologischen Aktivität verwendet. Dies bedeutet, dass je höher der Eintrag in HH_2 wird, desto höher wird der vorhergesagte Wert für die biologische Aktivität. Da ein möglichst hoher Wert für die biologische Aktivität und somit ein niedriger Wert für den K_i angestrebt wird, sind Substanzen mit hohem Gesamtbeitrag in HH_2 interessant.

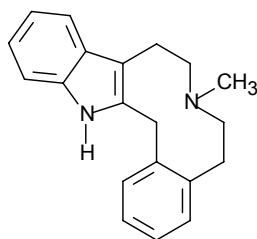


Abbildung 66: Die Leitstruktur der Substanzen im D₁-Datensatz, das LE-300.

Betrachtet man die für HH_2 entscheidenden hydrophilen Areale genauer, so fällt auf, dass diese das hydrophile Amin in LE-300 (siehe Abbildung 66) und allen anderen Strukturen beschreiben. Verschiedene Substituenten an dieser Stelle beeinflussen die Aktivität deutlich. Dies ist aus Untersuchungen der Struktur-Wirkungsbeziehungen bekannt. Variable HH_2 quantifiziert genau diese Substitution. Analog zum AZT-Datensatz wird die Umgebung genau eines Atoms über zwei umgebende Patches kodiert. Abbildung 67 zeigt die genaue Bedeutung anhand von drei sehr ähnlichen Strukturen. Das links gezeigte LE-PM-440 hat einen K_i -Wert von 272 nM, das in der Mitte gezeigte LE-404 0.3 nM und das rechts gezeigte LE-PM-426 einen von 3.8 nM.

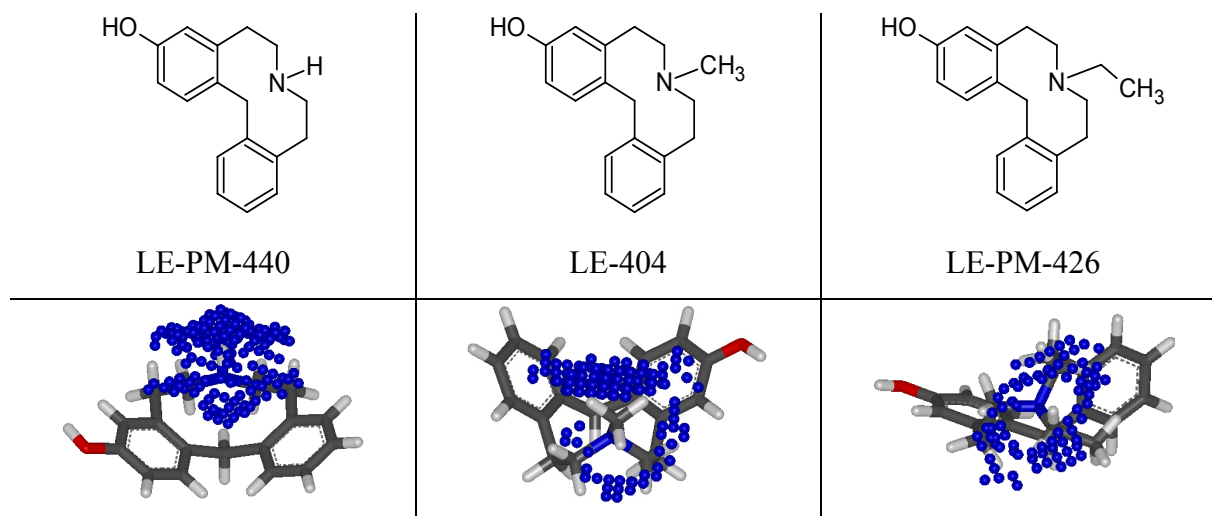


Abbildung 67: Die Rückprojektion der Variable HH_2 auf ausgewählte Strukturen des D₁-Datensatzes: Von links nach rechts sind die Strukturen der Verbindungen LE-PM-440, LE-404 sowie LE-PM-426 gezeigt.

In der linken Struktur gibt es nur einen einzigen hydrophilen Bereich um den Stickstoff. Variable HH_2 wird folglich überhaupt nicht inkrementiert, wodurch eine schlechtere biologische Aktivität vorhergesagt wird. Die mittlere Struktur weist aufgrund des N-Methyl-Restes zwei große hydrophile Areale um den Stickstoff auf. Variable HH_2 wird stark inkrementiert, die vorhergesagte biologische Aktivität wird höher. Wird der Rest am tertiären Amin weiter verlängert, so verringert sich die Hydrophilie des entsprechendenamins. Die hydrophilen Areale werden demzufolge aufgrund des lipophileren Rests kleiner, wodurch

eine schlechtere biologische Aktivität vorhergesagt wird. Im Gegensatz zu Molekül LE-PM-440 zeigt die Variable HH_2 aber einen gewissen Gesamtbeitrag. Dadurch nimmt LE-PM-426 im Hinblick auf die vorhergesagte biologische Aktivität eine Mittelposition zwischen dem unsubstituierten LE-PM-440 und dem optimalen Methylderivat LE-404 ein. Dies entspricht genau den realen Gegebenheiten. Betrachtet man den gesamten Datensatz, so erkennt man, dass HH_2 immer dann ideale Werte annimmt, wenn der aliphatische Stickstoff methyliert ist. Alle weiteren Substituenten führen zu einer schlechteren biologischen Aktivität. Dass HH_2 nicht mehr von der Variabilität erklären kann, liegt daran, dass substituierte Seitenketten oftmals einen zusätzlichen Beitrag in diese Variable bringen und somit das Bild etwas verzerren.

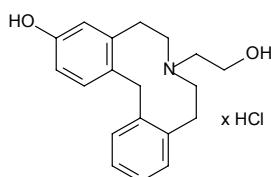


Abbildung 68: Die Struktur von LE-PM-437. Der Rest am tertiären Amin trägt eine OH-Gruppe. Variable HH_2 wird zusätzlich inkrementiert.

Beispielsweise geschieht das, wenn der Substituent am Stickstoff wie im LE-PM-437 (siehe Abb. 68) eine OH-Gruppe trägt. Auch wird ein größerer Substituent (längere Kette) am Stickstoff nicht mehr optimal kodiert. Es sind nach dem Übergang von einem Methyl- auf einen Ethylrest bei weiterer Steigerung der Kettenlänge keine großen Unterschiede in HH_2 mehr feststellbar. Die Beobachtung, dass eine N-Methyl-Funktion optimal ist, wird auch durch die Bildung eines auf einer Indikatorvariablen basierenden Modells untermauert. Dabei wird ein Deskriptor erzeugt, der den Wert 1 annimmt, wenn das untersuchte Molekül eine N-Methyl-Funktion hat und den Wert 0, insofern diese nicht vorhanden ist. Mit diesem Deskriptor wurde ein Modell mit den folgenden Güteparametern erstellt: $R^2=0.33$; $R^2_{CV-I}=0.30$; $R^2_{CV-50\%}=0.28$; Allein das Vorhandensein der N-Methyl-Funktionalität kann also rund 30% der Variabilität der Daten erklären. Damit wird die Aussage von Variable HH_2 weiter unterstützt.

Betrachtet man die Variable HD_8 , die ein hydrophiles Areal in einer Distanz von 8 Å zu einem Wasserstoffbrücken-Donor-Areal beschreibt, dann führt auch das Vorhandensein dieses potenziellen Zweipunkt-Pharmakophors zu einer besseren vorhergesagten biologischen Aktivität. Der Grund dafür ist, dass auch HD_8 einen positiven Regressionskoeffizienten zeigt. Dazu sind in Abbildung 69 die Strukturen der Verbindungen LE-410, LE-CE-521 und LE-403 gezeigt. Diese Verbindungen haben K_i -Werte von 4.5 nM, 10.9 nM und 341 nM. Der Gesamtbeitrag der Variablen HD_8 steigt hier proportional zu den K_i -Werten. Dies steht im Gegensatz zu der eben getroffenen Aussage, dass diese Variable im Gesamtmodell negativ

mit dem K_i -Wert und somit positiv mit der biologischen Aktivität korreliert ist. Folglich muss HD_8 eine andere Bedeutung haben, da sie immerhin 20% der Gesamtvariabilität erklären kann. In fast allen Fällen beschreibt sie die Stellung der Substituenten an den Aromaten vom zentralen aliphatischen Amin aus.

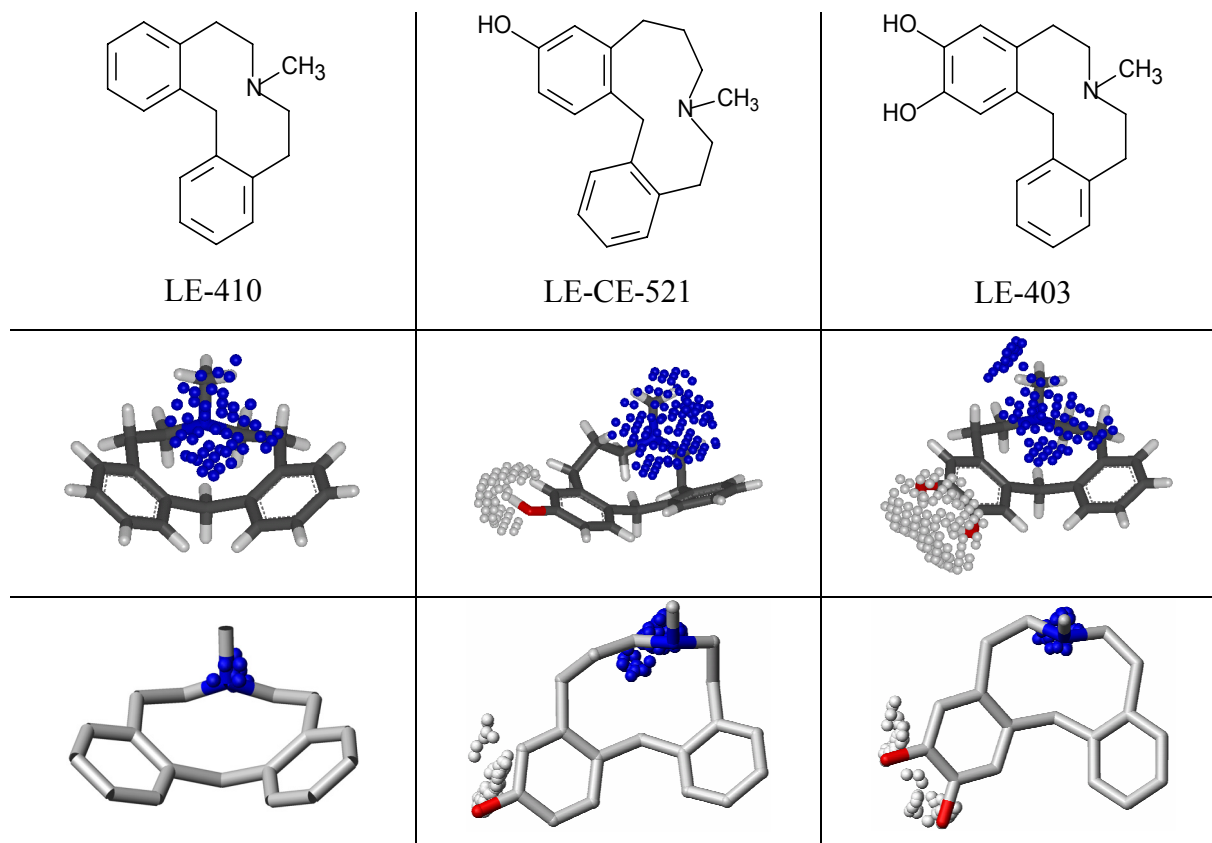


Abbildung 69: Die Rückprojektion der Variable HD_8 auf ausgewählte Strukturen des D₁-Datensatzes: Von links nach rechts die Strukturen der Verbindungen LE-410, LE-CE-521 sowie LE-403 (Mitte: ein repräsentatives Konformer, unten: die vollständige 3D-Rückprojektion mit dem Mittelwertkonformer). LE-410 erfüllt den potenziellen Zweipunkt-Pharmakophor HD_8 überhaupt nicht. Der Gesamtbeitrag der Variablen steigt für die Moleküle von links nach rechts immer weiter an.

Betrachtet man das LE-404 (siehe Abbildung 70), das einen K_i -Wert von 0.4 nM hat, dann kodiert Variable HD_8 , dass das Vorhandensein der zusätzlichen phenolischen Gruppe im Vergleich zum LE-410 (siehe Abbildung 69 links) einen positiven Einfluss hat. Es wird ein niedrigerer K_i -Wert und somit eine bessere biologische Aktivität vorhergesagt. Hier zeigt sich die im Gesamtmodell gefundene positive Korrelation von HD_8 zur biologischen Aktivität.

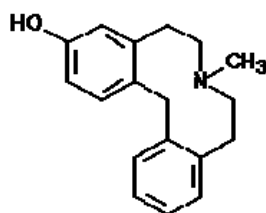


Abbildung 70: Struktur LE-404 passt im Gegensatz zu LE-CE-521 in die in Abbildung 69 gezeigte Reihe, wenn man den Gesamtbeitrag zur Variablen HD_8 zu Grunde legt.

Die Ringerweiterung im LE-CE-521 (der einzige Unterschied zum LE-404), das nicht in die in Abbildung 69 gezeigte Serie passt, muss anderweitig kodiert werden. Damit ist die Aussage von HD_8 bestätigt, es wird die positive Korrelation der phenolischen Substituenten beschrieben.

In Kombination beschreiben die Variablen sehr gut, welche Reste an den Aromaten gut und welche schlecht für einen möglichst niedrigen K_i -Wert sind. Auch die statistische Modellqualität ist gut. Dies wird durch die Darstellung in Abbildung 71 weiter untermauert.

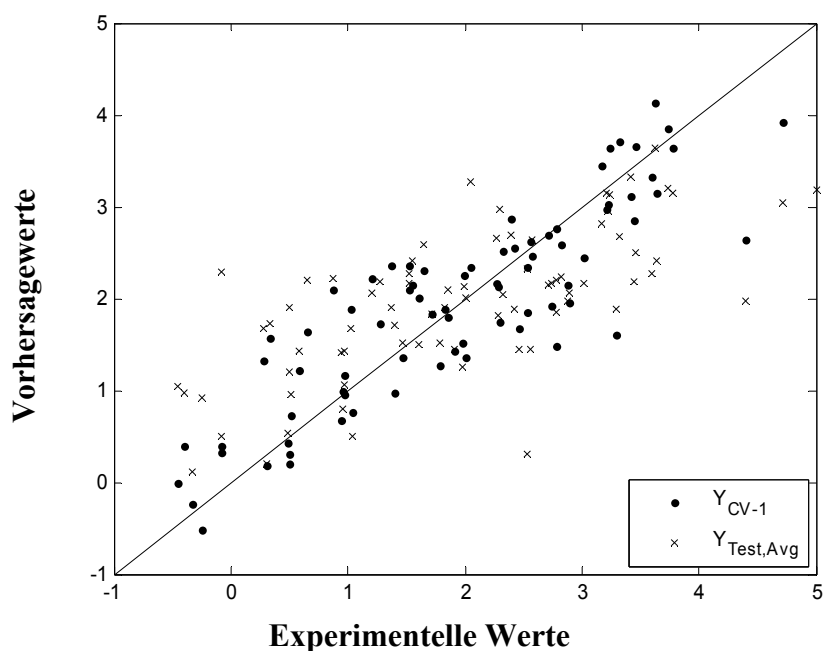


Abbildung 71: Externe vs. interne Vorhersagekraft für das erste Modell für den D_1 -Datensatz ($n=77$; $R^2_{CV,i}=0.71$; $R^2_{Test,Avg}=0.47$).

Die weiteren Variablen im Modell beschreiben die genaue Position der Aromaten und kodieren einzelne Substituenten. Hierbei handelt es sich um eine weitere Anpassung des Modells, der Beitrag zu globalen Modellqualität ist niedrig. Es bleibt festzuhalten, dass die immer vorhandenen aromatischen Ringe essenziell für die biologische Aktivität sind. Für eine Feinanpassung ist exemplarisch folgende Variable zu nennen: Die mit der biologischen Aktivität negativ korrelierte Variable $LsLs_{13}$ diskriminiert die verschiedenen Dimere (siehe

Anhang II.5.), die im Datensatz enthalten sind und alle einen schlechteren K_i -Wert und somit eine schlechtere biologische Aktivität aufweisen. Nur diese sind groß genug, um einen derart langreichweitigen potenziellen Zweipunkt-Pharmakophor zu erfüllen. Da deren Gesamtanzahl aber relativ niedrig ist, kann $LsLs_{13}$ nur knapp 12% der Gesamtvariabilität in den Daten erklären. Das Gesamtmodell wird durch $LsLs_{13}$ aber deutlich verbessert.

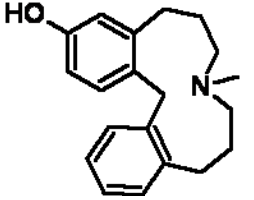
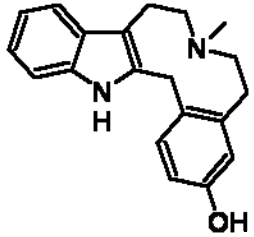
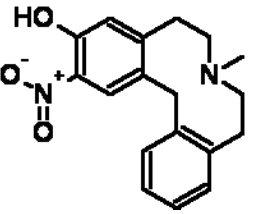
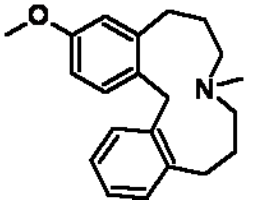
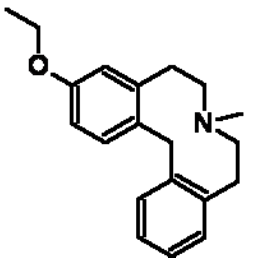
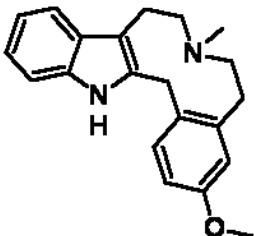
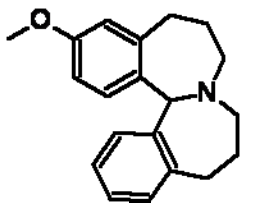
Alle anderen Variablen beschreiben die Stellung verschiedener Reste zueinander sowie die Position der immer vorhandenen Aromaten. Damit wird zum Teil die Größe des zentralen Rings kodiert, die mehrfach variiert wurde (siehe z.B. LE-CE-521 in Abbildung 69). Keine dieser Variablen kann aber einen entscheidenden Beitrag zur Modellinterpretation liefern.

Mit xMaP konnte wie gezeigt erstmals ein zufrieden stellendes QSAR-Modell für diesen Datensatz erstellt werden. Dieses Modell stand für weitere Untersuchungen zur Verfügung. In Kooperation mit der Pharmazeutischen Chemie in Jena wurden auf Basis dieses Modells u.a. eine Prioritätenliste für künftige Synthesen erarbeitet und biologische Aktivitäten für noch nicht getestete Moleküle vorhergesagt. Details dazu sind in den folgenden Kapiteln gezeigt.

3.9.5.2 Vorhersagen bereits synthetisierter Moleküle

Im ersten Ansatz wurde obiges Modell genutzt, um die zu erwartende Aktivität von sieben bereits synthetisierten, aber noch nicht getesteten Substanzen vorherzusagen. Diese Substanzen mit den zugehörigen Vorhersage- und Messwerten sind in Tabelle 13 gezeigt. Auch diese Strukturen wurden von Christoph Enzensperger als maschinenlesbare 2D-Dateien zur Verfügung gestellt. Analog zu der Beschreibung in Kapitel 3.7.5. wurden die Deskriptoren berechnet. Mittels dieser Deskriptoren wurden 100 zufällige Trainings- und Testdatenunterteilungen berechnet und durch anschließende Modellbildung die Vorhersagen für jede Substanz berechnet. Der Medianwert der Vorhersagen wurde als endgültiger Vorhersagewert für das entsprechende Molekül verwendet. Betrachtet man die in Tabelle 13 gezeigten Ergebnisse unter der Maßgabe, dass für Vorhersagen immer die negativen logarithmierten Werte eingesetzt wurden, so muss dieses Ergebnis als sehr gut bezeichnet werden. Der Korrelationskoeffizient zwischen den vorhergesagten und gemessenen Werten liegt bei 0.87. Verbindung I wurde als sehr gut vorhergesagt, was sich im Experiment nur teilweise bestätigte. Die Erklärung hierfür ist leicht zu finden. Im Trainingsdatensatz war kein Zwölfring-Derivat enthalten. Ausgehend vom Modell musste dahin gehend extrapoliert werden. Eine Abweichung von knapp 1.8 log-Einheiten ist also nicht verwunderlich. Bei Substanz II findet sich eine Abweichung von etwa 0.6 log-Einheiten. Ausgehend von der statistischen

Tabelle 13: Übersicht der vorhergesagten und anschließend gemessenen K_i -Werten für die gezeigten Strukturen, die auf dem vorgestellten xMaP-Modell basieren

	Struktur	K_i (Vorhersage)	K_i (Messung)	pK_i (Vorhersage)	pK_i (Messung)
I		2.5	152	8.6	6.82
II		16	3.7	7.8	8.43
III		20	49	7.7	7.31
IV		30	11.8	7.52	7.93
V		89	32.1	7.05	7.49
VI		94	19	7.03	7.72
VII		211	> 10 K	6.68	< 5

Modellgüte ist dieses Ergebnis im erwarteten Rahmen. In diesem Fall musste nicht extrapoliert werden. Gleiches gilt für die Vorhersagen der Strukturen III bis VI. Auch hier liegen die Vorhersagen immer unter einer log-Einheit vom experimentellen Wert entfernt. Anders ist es bei der ringgeschlossenen Verbindung VII, die eine Synthesevorstufe zu Verbindung I ist: Hier ist die Vorhersage völlig falsch. Dies ist nicht ungewöhnlich, da keine Verbindung im Datensatz enthalten war bei der der Stickstoff Teil eines aliphatischen Bicyclus war. Dadurch nimmt die Flexibilität des Moleküls stark ab. Anhand der berechneten Deskriptoren für dieses Molekül wurde jedoch ein guter K_i -Wert vermutet. Die Information, dass die freie N-Methylgruppe nicht nur positiven Einfluss auf die biologische Aktivität hat (wie es in Variable HH_2 kodiert ist), sondern absolut essenziell für eine gute biologische Aktivität ist, war in dem Modell zum Zeitpunkt der Vorhersage nicht kodiert. Eine dementsprechend falsche Vorhersage ist demzufolge zu erwarten. Zusammenfassend kann man zu diesem Punkt sagen, dass sich das xMaP-Modell hier in der Praxis bewährt hat.

3.9.5.3 Strukturvorschläge

Diese Ergebnisse wurden genutzt, um neue Moleküle zur Synthese vorzuschlagen. Im ersten Schritt wurde ein Minimalmolekül vorgeschlagen, das wichtige Aspekte des xMaP-Modells mit Informationen aus den Struktur-Wirkungs-Beziehungen verbindet. Das Ableiten dieses Moleküls aus dem Modell ist in Abbildung 72 gezeigt. Es handelt sich um Di-Phenyl-Methyl-Amin (DPMA). Der pK_a -Wert der N-Methyl-Funktion dieses Moleküls ist deutlich niedriger als dies in den Strukturen des Datensatzes der Fall ist. Dies wird durch die Oberflächenberechnung nur zum Teil widerspiegelt. Trotzdem wurde für dieses Molekül der entsprechende Deskriptor errechnet und ein K_i -Wert von 116 nM vorhergesagt.

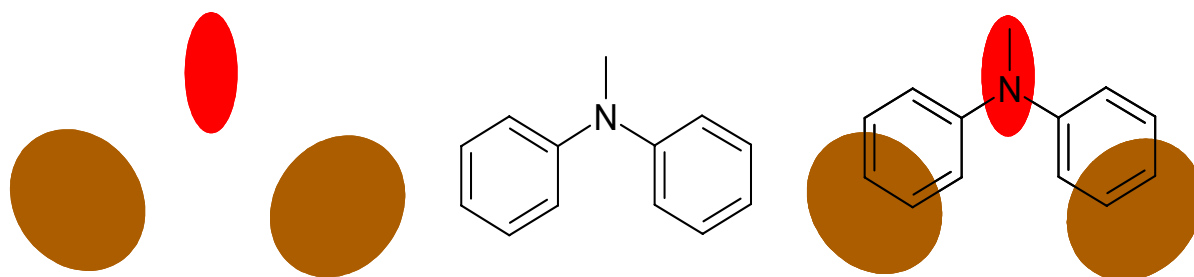
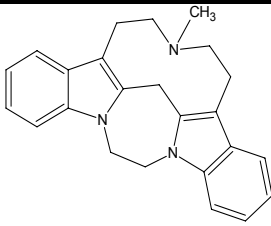
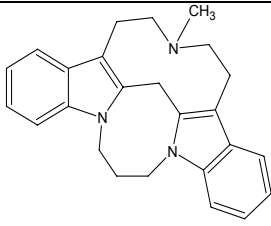
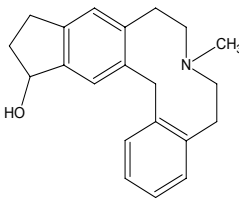
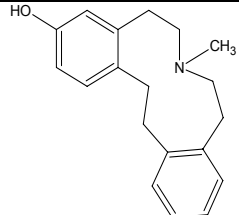
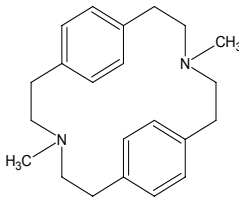
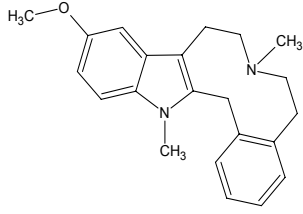
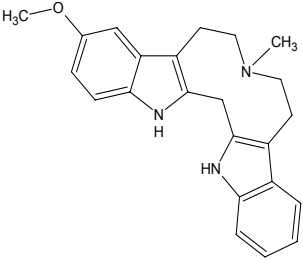


Abbildung 72: Die minimalen Anforderungen an einen Dopamin-D1-Antagonisten gemäß xMaP-Modell: Ein Akzeptor- bzw. hydrophiler Bereich (im Optimalfall ein methylierter Stickstoff) in einer bestimmten Entfernung zu zwei stark lipophilen Resten. Die Wichtigkeit der lipophilen Bereiche ist aus Untersuchungen der Struktur-Wirkungsbeziehungen bekannt und wird im xMaP-Modell u.a. in den LsLs-Variablen kodiert. Das Molekül in der Mitte – DPMA – erfüllt dies fast, wenn auch der Abstand zu klein ist.

Tabelle 14: Übersicht über die basierend auf dem Modell geplanten, vorhergesagten und anschließend zum Teil synthetisierten Moleküle.

Vorgeschlagene Struktur	K_i [nM] (Vorhersage)	K_i [nM] (Experiment)
	29	Synthese fehlgeschlagen
	2.7	noch nicht synthetisiert
	21	noch nicht synthetisiert
	35	noch nicht synthetisiert
	253	noch nicht synthetisiert
	2.2	1.9
	2.2	8.8

DPMA konnte in Jena getestet werden. In einem Calcium-Assay [248] ergab sich tatsächlich ein K_i -Wert von 1706 nM als Dopamin-Antagonist am D1-Rezeptor. Im Radioligandassay [206,249] konnte dieser Wert nicht bestätigt werden ($K_i > 10000$ nM). Diese Struktur ist sehr ähnlich zu der der Wirkstoffgruppe der Phenothiazine, die unter anderem als Antagonisten an Dopamin-Rezeptoren wirken. Bei diesen Verbindungen sind die beiden Ringe zusätzlich durch einen Schwefel verbrückt und der Stickstoff trägt längere Substituenten, von denen bekannt ist, dass sie essenziell für die Wirkung sind.

Das Chlorpromazin als Vertreter der Phenothiazine hat einen K_i -Wert von 76 nM als Dopamin-Antagonist am D1-Rezeptor. Das Modell wurde verwendet, um den K_i -Wert für Chlorpromazin vorherzusagen. Es resultierte eine Vorhersage von 721 nM. Vor dem Hintergrund, dass die Grundstruktur im Vergleich zu den Strukturen des Datensatzes stark unterschiedlich ist, so ist diese Abweichung von knapp einer log-Einheit im Rahmen der Erwartungen. Im nächsten Schritt wurden weitere Strukturen mit Christoph Enzensperger erarbeitet und deren K_i -Werte vorhergesagt. Einige der Strukturen sind in Tabelle 14 mit den vorhergesagten und gemessenen K_i -Werten gezeigt.

Bei den beiden unteren Strukturen sind die vorhergesagten und gemessenen K_i -Werte im Prinzip identisch. Dies ist ein weiterer Hinweis darauf, dass das erstellte xMaP-Modell geeignet ist und sich hier im „Alltagseinsatz“ bewährt hat. Die Extrapolation aus dem Datenraum des Modells heraus war nicht sehr groß. Das Ergebnis der biologischen Testung für die weiteren vorgeschlagenen zu synthetisierenden Moleküle bleibt abzuwarten.

3.9.5.4 Aktualisierung des bestehenden Modells

Nach Abschluss der obigen Untersuchungen lagen insgesamt 84 Strukturen mit antagonistischer Aktivität am D₁-Rezeptor vor. Dies wurde als Grundlage für ein neues Modell genutzt, das mit den Standardparametern erzeugt wurde. Die resultierende Gleichung lautet:

$$\begin{aligned}
 pK_i = & 2.72 \times 10^{-7} \cdot AA_{13} - 1.07 \times 10^{-6} \cdot HD_7 + 3.93 \times 10^{-6} \cdot HH_2 \\
 & + 1.66 \times 10^{-7} \cdot LwA_8 - 3.26 \times 10^{-5} \cdot LwD_3 \\
 & - 1.17 \times 10^{-4} \cdot LwH_3 - 0.0012 \cdot LwLw_1 - 0.0036 \cdot LwLw_3 \\
 & - 0.0327 \cdot LwLw_{12} - 0.03 \cdot LsA_1 - 0.04 \cdot LsD_7 \\
 & - 0.03 \cdot LsD_9 + 0.0476 \cdot LsLs_2 + 6.9885
 \end{aligned}
 \tag{Gleichung 48}$$

$$R^2_{Test,Avg}=0.41, RMSEP_{Test,Avg}=0.95$$

$$R^2_{CV-50\%}=0.52, RMSEP_{CV-50\%}=0.86, m=84$$

Dieses Modell hat etwas schlechtere Güteparameter als das bisherige. Für dieses Modell ist in Abbildung 73 die Qualität der Testdatenvorhersage gezeigt. Es ist auch bei diesem Modell wieder die Variable HH_2 , die die meiste Variabilität der Daten erklärt. In dem jetzigen Falle sind es etwa 19 Prozent. Die Interpretation ist insgesamt sehr ähnlich zum vorherigen Modell, weshalb auf eine erneute Diskussion der Interpretation verzichtet wird.

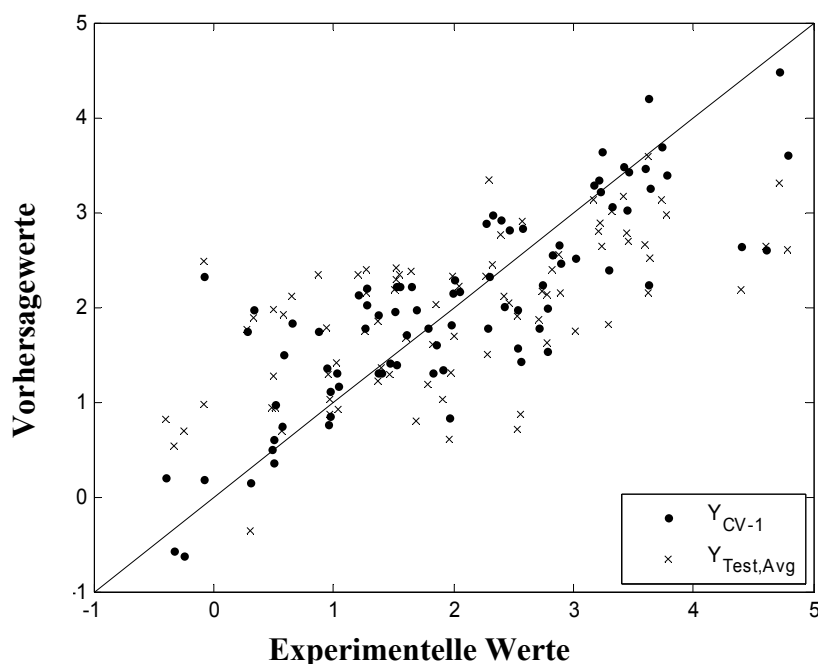


Abbildung 73: Externe vs. interne Vorhersagekraft für das aktualisierte Modell des D_1 -Datensatzes ($m=84$; $R^2_{CV-1}=0.59$; $R^2_{Test,Avg}=0.41$).

Für künftige Vorhersagen kann dieses Modell genutzt werden, das auf einer breiteren Datenbasis steht als das bisher benutzte.

3.9.5.5 Docking in Homologiemodell und strukturbasierte Interpretation

Um weiteren Einblick in die Struktur-Wirkungs-Beziehungen und Rezeptor-Ligand-Interaktionen dieses Datensatzes zu gewinnen wurde strukturbasierte Information in die Interpretation des Modells mit aufgenommen. Für den Dopaminrezeptor als G-Proteingekoppelten Rezeptor gibt es derzeit keine Kristallstruktur. Es muss eine Homologiemodellierung [250] mit der bekannten Struktur des bovinen Rhodopsins [251] als Grundlage erfolgen. Ein solches Homologiemodell wurde unter der Zugriffsnummer Q4QRJ0 der Modbase [162-164] entnommen. Es wird erwartet, dass ein Homologiemodell den Konformationsraum bei der Generierung der Konformere mit ausreichender Qualität einschränken kann. Es wird hier nicht dazu genutzt, um exakte Rezeptor-Ligand-Interaktionen

vorherzusagen oder Bindungsenergien abzuschätzen. Deshalb sollte die Qualität eines Homologiemodells ausreichen.



Abbildung 74: Ein Sequenzalignment der fünf Dopaminrezeptoren (durchgeführt mit ClustalX [252-256]). Die sieben transmembranären Helices sind markiert, ebenso wie die Aminosäuren, die für diese Arbeit wichtig sind (Einbuchstabencode für Aminosäuren, siehe Anhang III).

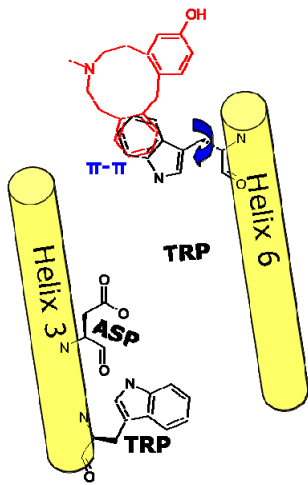
Um eine Analyse der Primärsequenzen der Dopamin-Rezeptoren durchführen zu können wurde mit ClustalX [252-256] ein Sequenzalignment aller Daten durchgeführt. Die dafür nötige Sequenzinformation wurde der UniProt-Datenbank [4-6] entnommen. Das Ergebnis ist in Abbildung 74 zu sehen.

Zur Beschreibung der Position der Aminosäuren wird im Folgenden die Nomenklatur von Ballesteros und Weinstein genutzt [257]. Dabei erhält jeweils die am stärksten konservierte Aminosäure in der Helix die Position 50. Alle Aminosäuren werden mit dem Einbuchstabencode benannt, der in Anhang III gezeigt ist. Eine Aminosäure-Bezeichnung beinhaltet darüber hinaus auch die Nummer der transmembranären Helix vor der Aminosäure-Nummer. Die Aminosäuren werden ausgehend von Aminosäure 50 nummeriert. Aus verschiedenen Publikationen war bekannt, dass Dopamin-Antagonisten und –Agonisten eine Salzbrücke mit dem Aspartat D3.32 (in Abbildung 74 markiert) des Rezeptors ausbilden [258]. Dies entspricht im Rhodopsin der Bindestelle des Retinals. Diese Information wurde dahingehend genutzt, dass für das mit FRED durchgeführte Docking eine Nebenbedingung definiert wurde. Der substituierte Stickstoff des Azecins (großes Ringsystem) musste in der Umgebung dieses Aspartats platziert werden. Dafür wurde eine Beschränkung gesetzt, dass der Stickstoff nicht mehr als 4.5 Å von einem der Sauerstoffe des D3.32 entfernt sein durfte. Eine ähnliche Beschränkung (3.5 Å) wurde von Xhaard und Mitarbeitern bei einem Docking von Catecholaminen in verschiedene G-Protein-gekoppelte Rezeptoren genutzt [259]. Dieses Aspartat ist in den Rezeptoren der D₁-Familie konserviert. Die Bindestelle liegt räumlich gesehen im Zentrum der sieben Helices.

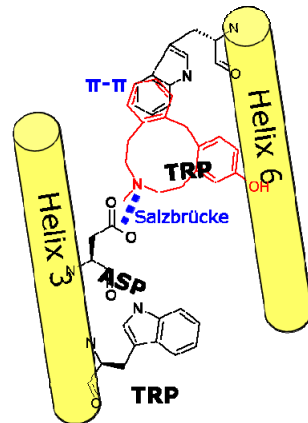
Zudem fällt bei der Sequenzanalyse auf, dass in direkter Nähe zu dem zentralen Aspartat ein Tryptophan (W3.28) zu finden ist, welches in den Dopaminrezeptoren der D₁-Familie konserviert ist. Von diesem Tryptophan ist bekannt, dass eine Mutation zum Tyrosin die Selektivität im Hinblick auf die Affinität zu Rezeptoren der D₁- und D₂-Familie umkehrt. Das wurde mittels Mutationsstudien gezeigt [261]. Dieses Tryptophan muss also in der Interaktion einen entscheidenden Einfluss haben. Ausgehend von dieser Stelle müssen Agonisten weiter entlang dem hydrophilen Bereich an Helix 3 tiefer in den Rezeptor hinein gelangen. Deshalb findet keine Konformationsänderung des Rezeptors bei der Bindung eines Antagonisten statt. Die Signaltransduktion wird unterbrochen [258,260]. Außerdem fällt bei einer detaillierten Analyse der Sequenzen auf, dass es am Eingang zur aminergen Bindestelle an Helix 6 ein weiteres hochkonserviertes Tryptophan gibt. Dieses Tryptophan W 6.48 ist in allen G-Protein gekoppelten Rezeptoren (d.h. auch in allen Dopamin-Rezeptoren) konserviert. Es muss im Rahmen der Ligandbindung eine entscheidende Rolle spielen. Nur so kann die Bindestelle vom extrazellulären Raum aus erreicht werden kann. Für dieses Tryptophan wurde in am β_2 -Adrenorezeptor mit UV-Resonanz eine konformelle Umlagerung aufgezeigt [262]. Diese Umlagerung wird durch dynamische Kristallographie am Meta-Rhodopsin untermauert [263]. Die vorliegende Information wurde in ein Mechanismus-Modell umgesetzt. Die jeweils 30 besten gefundenen Dockinglösungen wurden als Grundlage für ein QSAR-Modell zu den untersuchten Substanzen genutzt. Dieses Modell ist in seiner statistischen Qualität deutlich schlechter als das ligandbasierte und wird daher nicht weiter diskutiert. Fasst man alle bisher gesammelten Informationen zusammen, so kann ein Modell für den Wirkmechanismus der untersuchten Substanzen erstellt werden. Dabei passen die im QSAR-Modell identifizierten wichtigen Struktureigenschaften (N-Methyl-Funktion, lipophile Bereiche) zu den bei den anderen Analysen als wichtig identifizierten Aminosäuren. Dem Modell liegen die beschriebenen Informationen zu Grunde. Details dazu sind in Abbildung 75 gezeigt und werden im Folgenden erläutert.

Das am Eingang zu der Dopamin-Bindestelle liegende hochkonservierte Tryptophan fungiert als „Falle“ für Antagonisten und Agonisten. Es ist zu vermuten, dass dies über die Ausbildung einer π - π -Wechselwirkung zwischen einem der Aromaten aus den untersuchten Antagonisten und dem Indolrest des Tryptophans W6.48 geschieht. Diese π - π -Wechselwirkung in dem Bereich der Helix 6 ist für Agonisten am β_2 -Rezeptor nachgewiesen worden [264]. Das xMaP-Modell kodiert die lipophilen Bereiche in den verschiedenen Lw- und Ls-Variablen.

Interaktion („Einfangen“) des Antagonisten am W6.48 mit anschließendem Tryptophan-Flip



Ausbildung der essenziellen Salzbrücke zum D3.32 im Rezeptor



Zweites Tryptophan W3.28 interagiert mit („fängt“) den Antagonisten, der Rezeptor ist blockiert

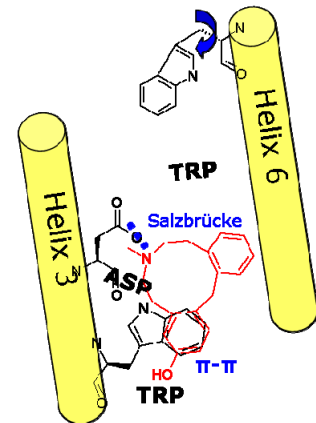


Abbildung 75: Das Modell für den Wirkmechanismus der hier untersuchten Dopamin-Antagonisten.

Ist der Antagonist fixiert, so klappt das Tryptophan in den Freiraum, der von den transmembranären Helices umspannt wird und nimmt dabei den Antagonisten mit. Diese Aminosäure hat also gewissermaßen eine Funktion als Türsteher am Zugang zur Bindestelle. Nur Moleküle, die entsprechend „genehm“ sind, werden in die Bindestasche aufgenommen. Nach dem Umklappen des Tryptophans zusammen mit dem gebundenen Antagonisten wird es für den Stickstoff des aliphatischen Rings möglich, eine Salzbrücke zum Aspartat D3.32 auszubilden. Dies ist in Abbildung 75 in der Mitte gezeigt. Diese Ausbildung der Salzbrücke befördert gleichzeitig das Umklappen des Tryptophans. Der Antagonist wird also weiter in den freien Raum innerhalb des Rezeptors gezogen, die Salzbrücke wird immer stärker. Irgendwann kommt dann das Molekül dem weiteren innenliegenden Tryptophan so nahe, dass mit dem zweiten aromatischen System eine weitere π - π -Wechselwirkung ausgebildet werden kann. Somit wird der Antagonist immer weiter in den Innenraum hineingezogen. Durch die Fixierung an zwei Punkten (Aspartat und Tryptophan) lässt nun das „Türsteher-Tryptophan“ los und klappt wieder in seinen Ursprungszustand zurück. Der Antagonist verharrt in seiner Position mit den beiden hochkonservierten Aminosäuren und verhindert somit, dass eine weitere Interaktion und somit Konformationsänderung tiefer in der Bindestasche erfolgen kann, was für eine Freigabe des G-Proteins notwendig wäre [264]. Das zeigt Abbildung 75 auf der rechten Seite. Der Rezeptor ist blockiert. Weitere Umlagerungen am Rezeptor sind nur möglich, wenn der Antagonist wieder in den extrazellulären Raum freigegeben wird. Anhand dieses Modells können die Struktur-Wirkungsbeziehungen aller untersuchten Moleküle über die Interaktion zu den Tryptophanen erklärt werden. Ferner kann auch die teilweise Selektivitätsumkehr der untersuchten Antagonisten zwischen D_1 - und D_2 -Rezeptor erklärt

werden. Der Grund dafür ist, dass im D₂-Rezeptor das W3.28 ein Tyrosin ist. Dieses Y3.28 des D₂-Rezeptors kann mit den Verbindungen aus dem Datensatz besser über eine π-π-Wechselwirkung interagieren, die einen niedrigeren K_i-Wert am D₂-Rezeptor haben. Umgekehrt können Verbindungen, die besser mit dem D₁-Rezeptor interagieren besser mit einem Tryptophan als einem Tyrosin wechselwirken. Dies passt zu den Ergebnissen der Mutationsstudien [261].

Eine Kombination von ligand- und strukturbasiertem QSAR-Modell wurde mit den Ergebnissen weiterer theoretischer Untersuchungen kombiniert. Die durch die QSAR-Modelle beschriebenen Molekülbestandteile sind komplementär zu den Aminosäuren, die durch die Sequenzanalyse als wichtig identifiziert wurden.

3.9.5.6 xMaP-Modelle für Antagonisten weiterer Dopaminrezeptoren

Für die vorliegenden Moleküle sind auch die K_i-Werte gegenüber weiteren Dopaminrezeptoren ermittelt worden. Es wurde daher auch versucht, für D₂ bis D₅ jeweils ein xMaP-Modell zu erstellen. Die entsprechenden Ergebnisse sind im Folgenden kurz diskutiert. Für die antagonistische Aktivität am D₂-Rezeptor ergab sich folgende Gleichung:

$$\begin{aligned}
 pK_i = & 4.0 \times 10^{-5} \cdot AA_7 + 2.86 \times 10^{-5} \cdot LwA_3 - 9.12 \times 10^{-5} \cdot LwD_4 \\
 & - 9.18 \times 10^{-4} \cdot LwH_3 + 6.73 \times 10^{-4} \cdot LwH_{14} + 0.0027 \cdot LwLw_1 \\
 & + 0.0035 \cdot LwLw_7 + 0.0035 \cdot LsH_1 + 0.0034 \cdot LsH_{12} \\
 & + 0.0034 \cdot LsLs_3 + 0.0034 \cdot LsLs_6 + 6.6648
 \end{aligned}
 \tag{Gleichung 49}$$

$$R^2_{Test,Avg}=0.45, RMSEP_{Test,Avg}=0.69$$

$$R^2_{CV-50\%}=0.62, RMSEP_{CV-50\%}=0.57, m=67$$

Dieses Modell hat auch eine mäßige Qualität. Interessanterweise wird hier oft der zentrale schwach lipophile Bereich als Ausgangspunkt eines potenziellen Zweipunkt-Pharmakophors für die Beschreibung von wichtigen Substituenten an den Aromaten verwendet. Die beiden wichtigsten Variablen sind die LwLw₇, die allein rund 35 % der Variabilität in den Daten beschreibt, sowie die LwH₃, die allein rund 25% erklärt. Bei der zuletzt genannten Variablen ist die Erklärung einfach: Analog zum D₁-Modell wird wieder der Substituent am Stickstoff beschrieben. Am größten sind die ihn umgebenden hydrophilen Patches wieder bei einer N-Methyl-Funktion. Der zweite Bestandteil des potenziellen Zweipunkt-Pharmakophors ist der große das ganze Molekül umfassende lipophile Bereich. Es wird folglich das gleiche wie bei Variable HH₂ im D₁-Modell kodiert.

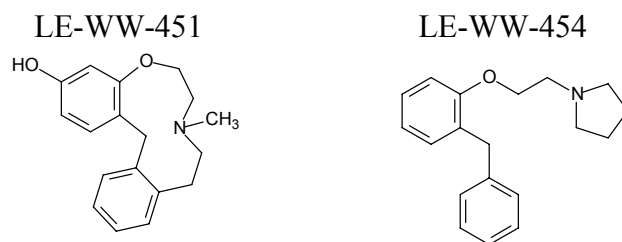


Abbildung 76: Zwei Beispiele von Strukturen aus der LE-WW-Serie.

Variable $LwLw_7$ ist insbesondere bei den LE-WW-Strukturen sehr hoch inkrementiert: Diese Strukturen passen nicht in das allgemeine Grundgerüst sondern sind eher lineare Verbindungen (siehe Abbildung 76 und Anhang II.5). Sie enthalten oft eine Etherfunktionalität und weisen relativ schlechte K_7 -Werte auf. Die Etherfunktionalität hat zur Folge, dass aus dem einen zentralen lipophilen Areal zwei werden. Deshalb kann über eine positiv korrelierte Variable viel von der Variabilität in den Daten beschrieben werden, für Grundstrukturen mit zwei großen, schwach lipophilen Arealen wird eine niedrigere biologische Aktivität vorhergesagt. Diese Strukturklasse hat interessanterweise zum Teil am D_1 -Rezeptor eine sehr gute Aktivität. Deshalb kann die entsprechende Variable dort kaum Variabilität erklären. Auch hier zeigt sich wieder die gute Interpretierbarkeit der xMaP-Modelle.

Tabelle 15: Die errechneten statistischen Güteparameter für die xMaP-Modelle für die weiteren Dopaminrezeptoren.

	D_3	D_4	D_5
n	57	64	57
R^2_{CV-1}	0.27	0.38	0.36
$R^2_{CV-50\%}$	0.18	0.27	0.27
$R^2_{Test,Avg}$	-0.02	0.04	0.00

Bezüglich der anderen Dopaminrezeptoren ergaben sich schlechtere Ergebnisse, diese sind in Tabelle 15 gezeigt. Warum diese Werte vor allem in ihrer externen Vorhersagekraft so deutlich von den Modellen zu D_1 und D_2 abweichen kann nur gemutmaßt werden.

Möglicherweise reichen die jeweils vorliegenden Daten nicht aus, um ein Bild herausarbeiten zu können.

3.9.5.6 Fazit für den Dopamin-Datensatz

Dieser Datensatz zeigt die breite Anwendbarkeit von xMaP im Rahmen der Arzneimittelforschung. Ein Datensatz, der bisher allen Modellierungsversuchen erfolgreich widerstanden hat, konnte hier zumindest mit den Daten für den D_1 - und D_2 -Rezeptor modelliert werden. Weitere Untersuchungen sind Erfolg versprechend. Sollte sich die Datenbasis für die weiteren Rezeptoren noch erweitern, so können auch dort möglicherweise gute Modelle generiert werden.

3.9.6. Glukokortikoide (GK)

3.9.6.1 Der Einfluss der Lipophilie

Es sollte die Korrelation der Lipophilie der untersuchten Substanzen mit der relativen Rezeptoraffinität (RRA) bestimmt werden. Bisher zeigten alle QSAR-Modelle für Glukokortikoide eine Abhängigkeit der relativen Rezeptoraffinität von der Lipophilie und somit vom Oktanol-Wasser-Verteilungskoeffizienten (logP) [265-267]. Um dies für die vorliegenden Substanzen nachzuvollziehen wurden mit VCC-Lab [268] pro Molekül sieben verschiedene Oktanol-Wasser-Verteilungskoeffizienten vorhergesagt. Eine Übersicht dieser Werte ist im Anhang II.6. zu sehen. Tabelle 16 zeigt die statistischen Güteparameter der Modelle, bei denen die einzelnen logP-Werte als unabhängige Variable einer einfachen linearen Regression eingesetzt wurden. Dabei wurde die gleiche Validierung wie bei xMaP verwendet.

Tabelle 16: Die errechneten statistischen Güteparameter für den GK-Datensatz mit auf den vorhergesagten logP-Werten als Grundlage.

	logP (Broto)	virtual logP	milogP	alogPs	IAlogP	clogP	XLOGP
R²	0.12	0.09	0.12	0.14	0.10	0.07	0.09
R²_{CV-1}	-0.04	-0.07	0.00	0.02	-0.04	-0.09	-0.07
R²_{CV-50%}	-0.19	-0.23	-0.11	-0.08	-0.18	-0.24	-0.21
R²_{Test,Avg}	-0.04	-0.07	0.00	0.02	-0.03	-0.10	-0.08

Es zeigt sich, dass sämtliche statistischen Güteparameter für diese Modelle um Null schwanken. Man kann also davon ausgehen, dass hier keine Korrelation zwischen der Lipophilie und der relativen Rezeptoraffinität vorliegt. Das steht im Widerspruch zu bisherigen Untersuchungen an Glukokortikoiden [227]. Diese nicht vorhandene Korrelation wird durch experimentelle Bestimmungen der Lipophilie untermauert [269]. Auch mit experimentell bestimmten Werten kann für alle Substanzen keine Korrelation gefunden werden. Diese Resultate sind in Abbildung 77 gezeigt.

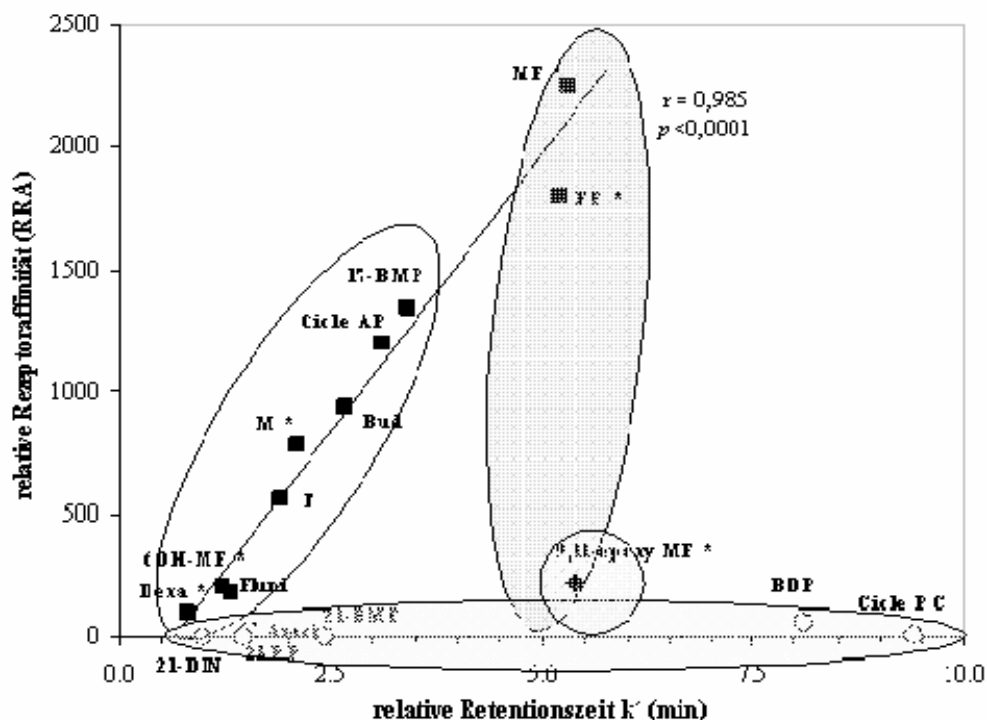


Abbildung 77: Die RRA in Korrelation zur Lipophilie der untersuchten Moleküle. Es fällt auf, dass einige Substanzen keine Korrelation zwischen den beiden Parametern aufweisen.

Da xMaP mehr Informationen als nur die Lipophilie von Molekülen kodiert, wurde versucht, für diese Daten ein Modell zu erstellen.

3.9.6.2 Das xMaP-Modell für den GK-Datensatz

Die mit Catalyst errechneten Konformere für alle Moleküle des Datensatzes wurden genutzt, um mit den Standardparametern ein Modell zu erstellen. Es wurde erwartet, dass die relativ starren Grundstrukturen nur wenige Konformere aufweisen. Somit war es überraschend, dass es im Mittelwert 49 (Median: 42) Konformere pro Molekül waren. Das Starrste Molekül (Nr. 29) hatte 18 energiearme Konformere, das flexibelste Molekül (Nr. 2) wies sogar 106 Konformere auf. Diese Werte sind, wie erwartet, niedriger als bei anderen Datensätzen, zeigen jedoch, dass dreidimensionale Techniken wertvolle Informationen vernachlässigen. Unter Verwendung der berechneten Konformere und der damit berechneten xMaP-Deskriptoren ergab sich folgendes Modell:

$$\log(RRA) = -0.0207 \cdot AA_{13} - 0.0191 \cdot DA_{12} - 0.0172 \cdot DD_4 \\ + 0.0074 \cdot LsD_9 - 0.0011 \cdot LsH_2 + 1.9102$$

$$R^2_{Test,Avg} = 0.37, RMSEP_{Test,Avg} = 1.61$$

Gleichung 50

$$R^2_{CV-50\%} = 0.74, RMSEP_{CV-50\%} = 0.57, m = 30$$

Dieser $R^2_{Test,Avg}$ -Wert ist wegen der relativ niedrigen Flexibilität der Substanzen und trotz der guten internen Güteparameter des Modells nicht zufrieden stellend. Eine der Substanzen (Nr.

29) kann als Ausreißer identifiziert werden, da der Mittelwert der Vorhersage um mehr als die dreifache Standardabweichung von der Abweichung aller Vorhersagen abweicht. Daher wurde ein weiteres Modell mit den verbleibenden 29 Molekülen berechnet:

$$\log(RRA) = -0.0247 \cdot AA_{13} - 0.0191 \cdot LwLw_{13} + 0.0088 \cdot LsA_6 \\ + 0.0014 \cdot LsD_{10} - 0.0021 \cdot LsH_2 + 1.9746$$

Gleichung 51

$$R^2_{Test,Avg}=0.58, RMSEP_{Test,Avg}=1.08$$

$$R^2_{CV-50\%}=0.80, RMSEP_{CV-50\%}=0.48, m=29$$

Der $R^2_{Test,Avg}$ ist hier deutlich höher, so dass von einer zufrieden stellenden Vorhersagekraft des Modells gesprochen werden kann. Vergleicht man die Modelle, so fällt auf, dass Variable AA_{13} in beiden eine zentrale Position einnimmt. Diese Variable erklärt in beiden Modellen ungefähr 40% der Varianz der Daten (Modell 2: $R^2_{CV-1}=0.42$; $R^2_{CV-50\%}=0.38$). Darüber hinaus ist ein Zusammenhang von zwei Wasserstoffbrückenakzeptoren für den Bereich der Glukokortikoide sehr interessant, da wie erwähnt bisher immer eine Korrelation mit der Lipophilie festgestellt wurde. Diese Variable wurde in den 100 Modellberechnungen basierend auf den verschiedenen Trainings- und Testdatensatzunterteilungen am zweithäufigsten gewählt. Daher muss ihr Einfluss diskutiert werden. Eine weitere in beiden Modellen enthaltene Variable ist die LsH_2 , welche auch negativ mit der relativen Rezeptoraffinität korreliert ist. Sie ist in den verschiedenen Unterteilungen die am Häufigsten gewählte Variable, obwohl sie nur rund 12% der Variabilität erklärt. Betrachtet man eine Verbindung mit sehr niedriger relativer Rezeptoraffinität wie beispielsweise Substanz 11, so zeigen sich für Variable AA_{13} mehrere Strukturvarianten, die diesen potenziellen Zweipunkt-Pharmakophor enthalten. Diese sind in Abbildung 78 oben zu sehen.

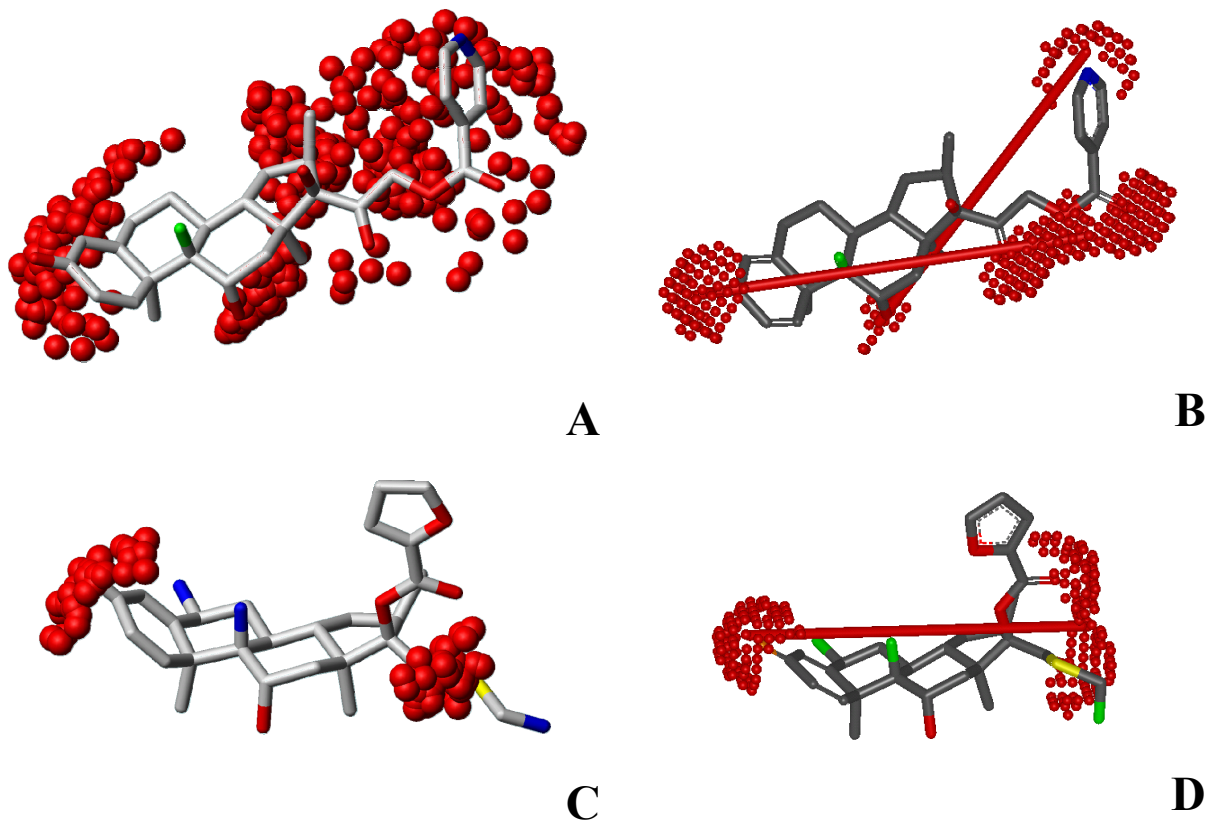


Abbildung 78:

Die Rückprojektion von Variable AA_{13} auf Molekül Nr. 11; A: Die Gesamtprojektion, dabei zeigt sich die Flexibilität in der Seitenkette; B: Die Projektion der Akzeptorpatches auf ein repräsentatives Konformer, es wird deutlich welche beiden Oberflächenkombinationen einen Beitrag zu AA_{13} leisten (der zusätzliche Akzeptorbereich an Sauerstoff an C-11 ist aus Übersichtlichkeitsgründen nicht gezeigt)

Unten ist die Rückprojektion auf Molekül 16 gezeigt. Hier ist aufgrund des veränderten Substitutionsmuster nur eine Möglichkeit vorhanden, AA_{13} zu erfüllen. Dieses Substitutionsmuster führt dazu, dass die vorhergesagte RRA deutlich höher wird. Dies korreliert gut mit den experimentellen Daten.

Über den kompletten Datensatz betrachtet gibt es zwei Strukturvarianten, die dazu führen, dass AA_{13} inkrementiert wird:

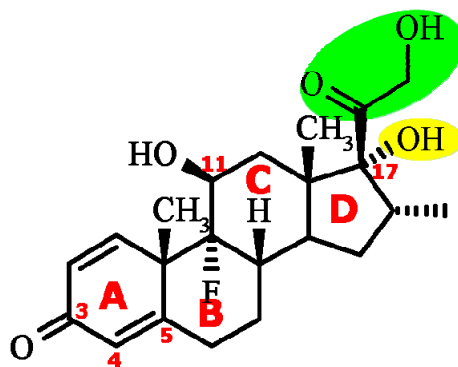


Abbildung 79: Die Struktur des Dexamethasons zur Illustration der Nummerierung der Kohlenstoff-Atome im Glukokortikoid-Datensatz. In grün ist die β -Position für Substituenten markiert, in gelb die α -Position.

Im einen Falle wird die für die Wirkung essenzielle Carbonylfunktion an C3 (Nummerierung siehe Abbildung 79) als Ausgangspunkt für Variable AA_{13} genommen und eine weitere Akzeptorfunktionalität an C17 als Endpunkt beschrieben. Diese Akzeptorfunktionalität liegt in fast allen Molekülen des Datensatzes vor (siehe Abbildung 78). Es sind verschiedene funktionelle Gruppen wie Ester oder Ether, die die Verbrückung zwischen C17 und dem entsprechenden Substituenten darstellen. Dies können Substituenten sowohl in der β - als auch in der α -Position sein (siehe Abbildung 78). Es muss betont werden, dass hier nicht die Substituenten, sondern die Verbindung zu ihnen kodiert wird. Diese ist faktisch in allen Molekülen des Datensatzes ähnlich. Dieser potenziellen Zweipunkt-Pharmakophor führt zu einem konstanten Teilbeitrag zu Variable AA_{13} und trägt deshalb nicht zur Unterscheidung von Molekülen bei.

Interessant für die Unterscheidung der einzelnen Moleküle ist dagegen der zweite potenzielle Zweipunkt-Pharmakophor, der zu AA_{13} beiträgt. Dieser hat seinen Ausgangspunkt an der OH-Funktionalität an C11. Dies ist in der Rückprojektion für Molekül 11 in Abbildung 78 B gezeigt. Trägt das C17 in der β -Position einen entsprechend großen Rest wie z.B. bei den Molekülen 1, 8, 11 und 28, so erfolgt ein zusätzlicher Beitrag zu Variable AA_{13} . Für das Molekül wird eine niedrigere Rezeptoraffinität vorhergesagt. Ein entsprechend großer Rest an der α -Position wird also als schlecht klassifiziert. Eine Substitution an der α -Position dagegen liegt zu nahe an C11 als dass der potenzielle Zweipunkt-Pharmakophor für AA_{13} erfüllt wird. Insgesamt gesehen kodiert also die Variable die Position der Substituenten und liefert gleichzeitig die Information welche Art von Substituent (Akzeptor) an der β -Position besonders schlecht ist. Der Akzeptor an C11 ist zusätzlich der Ausgangspunkt für die positiv korrelierte Variable LsA_6 , ist also einmal Bestandteil eines potenziellen Zweipunkt-Pharmakophors mit einem positiven Regressionskoeffizienten und einmal Bestandteil eines Zweipunkt-Pharmakophors mit negativem Regressionskoeffizienten. Das veranschaulicht, welche Art von Substituent an C17 β unerwünscht ist, nämlich ein entsprechend langer, der einen Akzeptor trägt.

Im Ensemble mit Variable LsH_2 ergibt sich ein Bild davon welche Art von Substituenten an welcher Stelle erwünscht ist. Dabei beschreibt diese Variable vor allem lokale Unterschiede. In vielen Fällen wird dabei ein Chlor- oder Fluor-Substituent beschrieben, womit jeweils zwei Moleküle unterschieden werden können. Interessanterweise ist es dann oft so, dass das Molekül mit der niedrigeren RRA den höheren Gesamtbeitrag zu dem entsprechenden potenziellen Zweipunkt-Pharmakophor beisteuert, obwohl die Variable global negativ korreliert ist (Beispiele: $RRA\ 9 < RRA\ 10$, aber $Inc\ 9 > Inc\ 10$; oder analog bei 23+24 oder

27+28). Eine durchgehende Interpretation für diese Variable zu finden ist schwierig. Den besten Einblick in den Datensatz liefert die AA₁₃.

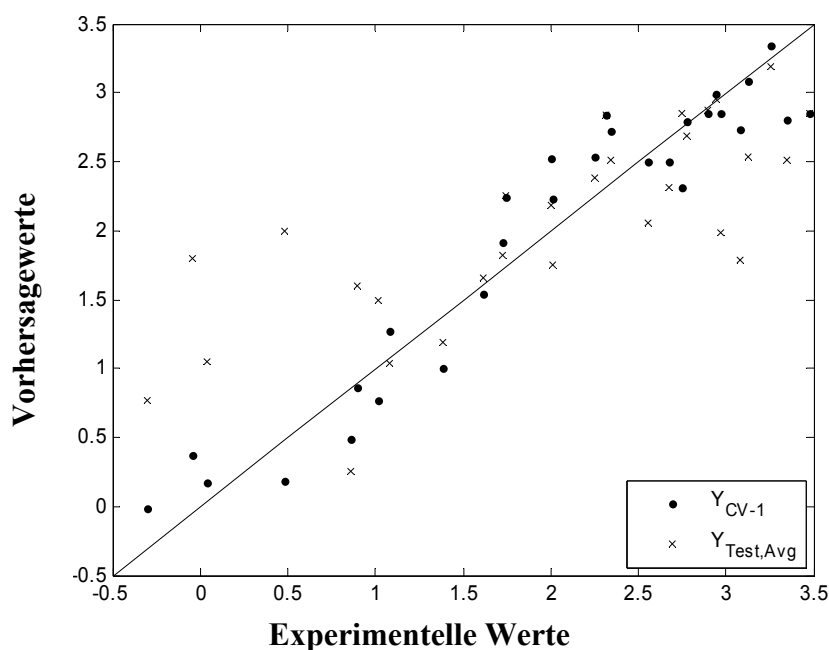


Abbildung 80: Externe vs. Interne Vorhersagekraft für xMaP-Modell zum GK-Datensatz ($R^2_{CV-1}=0.84$; $R^2_{Test,Avg}=0.58$)

Als ein Fazit kann hier gezogen werden, dass xMaP auch bei Molekülen mit starren Grundgerüsten sinnvolle Ergebnisse liefert, die wertvolle Beiträge bei einer Weiterentwicklung leisten können.

3.9.6.3 CoMFA- und CoMSIA-Modelle für den GK-Datensatz

Um weitere Einblicke in die Struktur-Wirkungs-Beziehungen der untersuchten Substanzen zu bekommen wurden CoMFA- und CoMSIA-Analysen durchgeführt. Die entsprechenden statistischen Güteparameter der einzelnen Modelle sind in Tabelle 17 gezeigt. Es zeigt sich, dass diese Werte (die interne Güteparameter der Modelle darstellen) gut mit den von xMaP erzielten Werten vergleichbar sind.

Tabelle 17: Ergebnisse der alignment-abhängigen 3D-QSAR-Techniken CoMFA und CoMSIA für den GK-Datensatz. In Klammern jeweils die Art des Alignments, H bedeutet Hand und F, dass die Software FlexS [231,232] verwendet wurde. SEP steht für den „Standard error of prediction“, q für die Anzahl der ausgewählten Komponenten.

	R^2	R^2_{CV-1}	SEP	q
CoMFA (H)	0.89	0.65	0.711	3
CoMFA (F)	0.97	-0.42	1.379	1
CoMSIA (H)	0.88	0.71	0.688	4
CoMSIA (F)	0.76	-0.28	1.333	2

Die Rückprojektion der einzelnen Felder liefert im Hinblick auf die Positionierung einzelner Reste das gleiche Ergebnis wie die xMaP-Analyse: An der α -Position sind sterisch anspruchsvolle Substituenten gut und erhöhen die RRA. An der β -Position dagegen haben sie negativen Einfluss auf die RRA. Obwohl beide Modelle gleich gut in ihrer statistischen Qualität sind liefert die Darstellung von CoMSIA-Feldern (siehe Abbildung 81 rechts) eine bessere, da eindeutiger graphische Interpretation

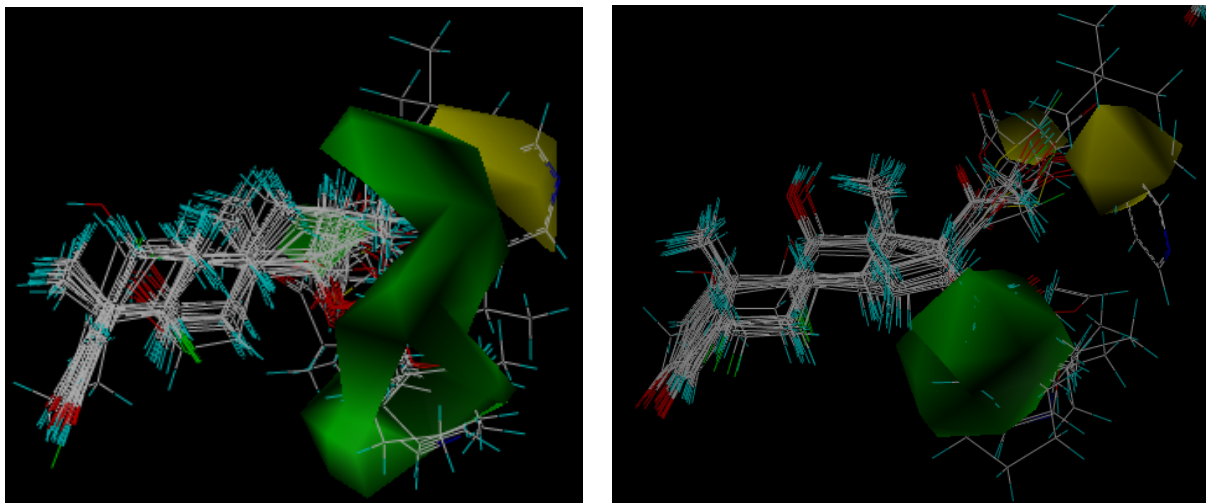


Abbildung 81: Die Rückprojektion der sterischen Felder von CoMFA (links) und CoMSIA (rechts); grün bedeutet dabei dass ein sterischer Einfluss an dieser Stelle (die α -Position für Substituenten, siehe Abbildung 31) positiv ist, gelb (also die β -Position) negativ. Gezeigt ist das Ergebnis des Hand-Alignments.

Zusätzlich zur Aussage über einen sterischen Einfluss kann mittels CoMFA- und CoMSIA-Analysen auch Information über vorteilhafte elektrostatische Eigenschaften der Moleküle ermittelt werden. Damit wird u.a. die Fähigkeit zur Wasserstoffbrückenbildung charakterisiert.

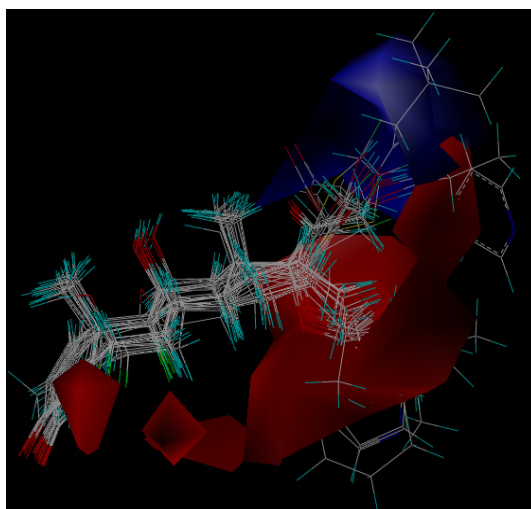


Abbildung 82: Die Rückprojektion des von CoMFA identifizierten Einflusses der Elektrostatik; blau bedeutet dabei dass ein Einfluss an dieser Stelle (die β -Position für Substituenten, siehe Abbildung 31) negativ ist, rot (also die α -Position) positiv.

Betrachtet man die elektrostatischen Felder, findet man die komplementäre Aussage zur Variablen AA_{13} des xMaP-Modells: Demzufolge ist beispielsweise ein Wasserstoffbrücken-Akzeptor an Position β negativ für die RRA. An Position α dagegen ist ein Wasserstoffbrücken-Akzeptor positiv zu sehen. Kombiniert man diese Aussagen mit denen des sterischen Feldes ergibt sich ein klares Bild. An Position β wird ein kleiner lipophiler Rest bevorzugt, Position α soll durch einen sterisch anspruchsvollen, nicht lipophilen Rest substituiert sein.

Diese Aussagen sind analog zu denen des xMaP-Modells. Dieses kodiert in der Variable AA_{13} neben der Wasserstoffbrückenbildungskapazität implizit durch die Distanz auch die Größe des Substituenten. Der Vorteil der CoMFA- und CoMSIA-Interpretation ist, dass alle Moleküle in einer einzigen Darstellung begutachtet werden können. Dies wird durch den Alignment-Schritt ermöglicht. Dieser entfällt bei xMaP, weswegen bei der Interpretation des xMaP-Modells alle Moleküle einzeln untersucht werden müssen.

3.9.7. Naphtylisochinolin-Alkaloide (NIQ)

Diese Strukturen werden im Arbeitskreis von Herrn Professor Bringmann an der Universität Würzburg beforscht. Sie zeigen u.a. Aktivität gegen den Malaria-Erreger *Plasmodium falciparum*. Da das Target dieser Substanzen nicht bekannt ist erscheint eine Verwendung von QSAR-Techniken vorteilhaft. Es konnten bereits mehrere QSAR-Modelle erstellt werden [81,233].

Nach der Konformerberechnung wurden auf Grundlage von NMR-Daten verschiedene Rotamere die *in-vitro* nicht vorliegen von der Analyse ausgeschlossen. Mit den verbleibenden Konformeren konnte mit den Standardparametern folgendes Modell generiert werden:

$$-\log(EC_{50}) = 0.0034 \cdot DA_{11} - 0.0026 \cdot HA_{13} + 0.0035 \cdot HD_4 \\ + 0.0012 \cdot LsH_8 + 0.1726$$

Gleichung 52

$$R^2_{Test,Avg} = 0.38, RMSEP_{Test,Avg} = 0.55$$

$$R^2_{CV-50\%} = 0.56, RMSEP_{CV-50\%} = 0.47, m = 45$$

Dieses Modell ist in seiner Aussagekraft mäßig. Dioncophyllinol B (siehe Anhang II.7) wurde daher analog zu der früher verwendeten Regel als Ausreißer aus der Analyse genommen. Es ergab sich folgendes Modell:

$$-\log(EC_{50}) = 0.009 \cdot DA_{13} + 0.0037 \cdot HD_4 + 0.0007 \cdot LsH_8 \\ + 0.144$$

Gleichung 53

$$R^2_{Test,Avg} = 0.46, RMSEP_{Test,Avg} = 0.50$$

$$R^2_{CV-50\%} = 0.66, RMSEP_{CV-50\%} = 0.40, m = 44$$

Wenn man die Beiträge der einzelnen Variablen zum Modell betrachtet, so fällt auf, dass hier insbesondere die HD_4 wichtig ist. Sie erklärt rund 35% der Variabilität in den Daten. Außerdem wird sie in den verschiedenen Trainings-/Testdatenunterteilungen mit Abstand am häufigsten als wichtig ausgewählt (in 50 von 100 Fällen). Am nächsthäufigsten wird die im Gesamtmodell ohne Ausreißer vorhandene DA_{11} gewählt, auch die im Modell nach Ausreißereliminierung vorhandene DA_{13} wird häufig gewählt. Darauf soll im Folgenden kurz eingegangen werden. Interessanterweise sind sowohl HD_4 als auch $DA_{11/13}$ Variablen, die identisch zu den Variablen sind, die in verschiedenen MaP-Modellen wichtig waren [19,81]. Beim MaP-Modell ohne Donor-Differenzierung war die DH_6 wichtig, die verschiedene Strukturparameter beschreibt, worauf hier nicht im Detail eingegangen werden soll. Dies entspricht der Variablen HD_4 im xMaP-Modell. Die niedrigere Distanz ist durch die folgenden Unterschiede zwischen MaP und xMaP bedingt: Aufgrund der doppelten

Eigenschaftszuweisung auf der Oberfläche werden die hydrophilen Bereiche größer und rücken somit in die Nähe des Donor-Areals, die Distanz wird kleiner. Der zweite Faktor ist die Flexibilität der Substanzen, auch hier passiert es, dass Patches näher zusammenrücken. Darüber hinaus beeinflussen die unscharfe Zählweise und die Verwendung von Patch-Zentroiden das Ergebnis der Modellbildung. Alles in allem ergibt sich eine sehr ähnliche Aussage bei MaP und xMaP. Auf die bei MaP erfolgte Donor-Differenzierung wurde hier verzichtet, die Variable $DA_{11/13}$ kann als identisch zur $AD_{S_{10}}$ bei MaP angenommen werden, der Unterschied um 1 in den Kategorien lässt sich durch die unscharfe Zählweise sowie die Flexibilität erklären. Da somit die Interpretation identisch zu der von MaP ist, wird auf eine ausführliche Darstellung verzichtet.

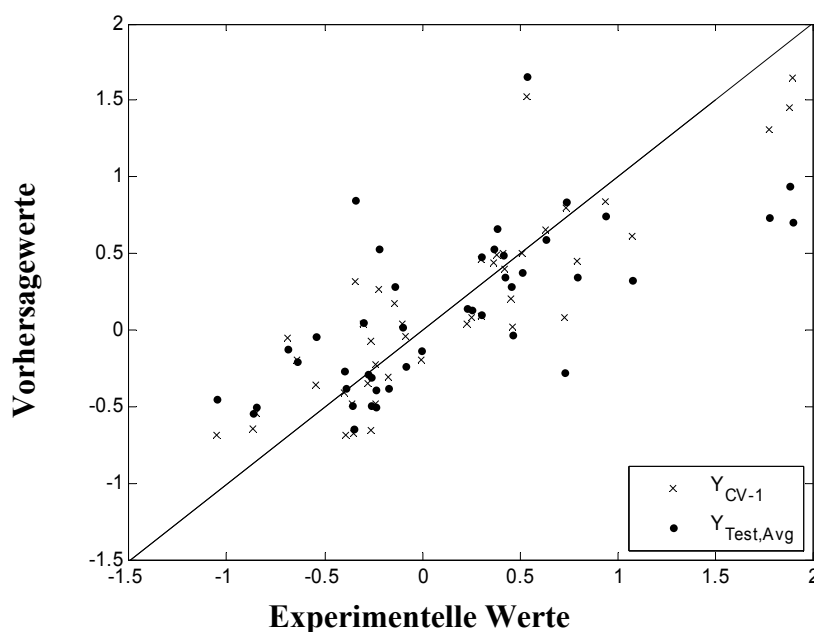


Abbildung 83: Die Qualität der Testdatenvorhersage für den NIQ-Datensatz ($R^2_{CV-1}=0.84; R^2_{Test,Avg}=0.46$).

Das hier erstellte Modell wurde auch zur Vorhersage von Aktivitäten von Molekülen, die sich derzeit in der Synthese befinden, verwendet. Da es hier derzeit noch keine Ergebnisse gibt, wird auf eine detaillierte Darstellung verzichtet. Ebenso befinden sich Modelle zur Beschreibung der NIQ-Aktivität beispielsweise gegen *Leishmania major* derzeit noch in der Anfangsphase.

3.9.8. Weitere Datensätze

Die folgenden Ergebnisse werden nur im Überblick der statistischen Güteparameter und ohne detaillierte Interpretation gezeigt. Diese Ergebnisse zeigen die breite Anwendbarkeit von xMaP für eine Vielzahl unterschiedlicher Datensätze. In allen Fällen war es ohne weiteres mit den Standardparametern von xMaP möglich, Modelle mit sehr ordentlicher Qualität zu erzeugen. Auch die entsprechende externe Vorhersagekraft ist jeweils zufrieden stellend. Abbildung 84 und Tabelle 18 zeigen die entsprechenden Werte.

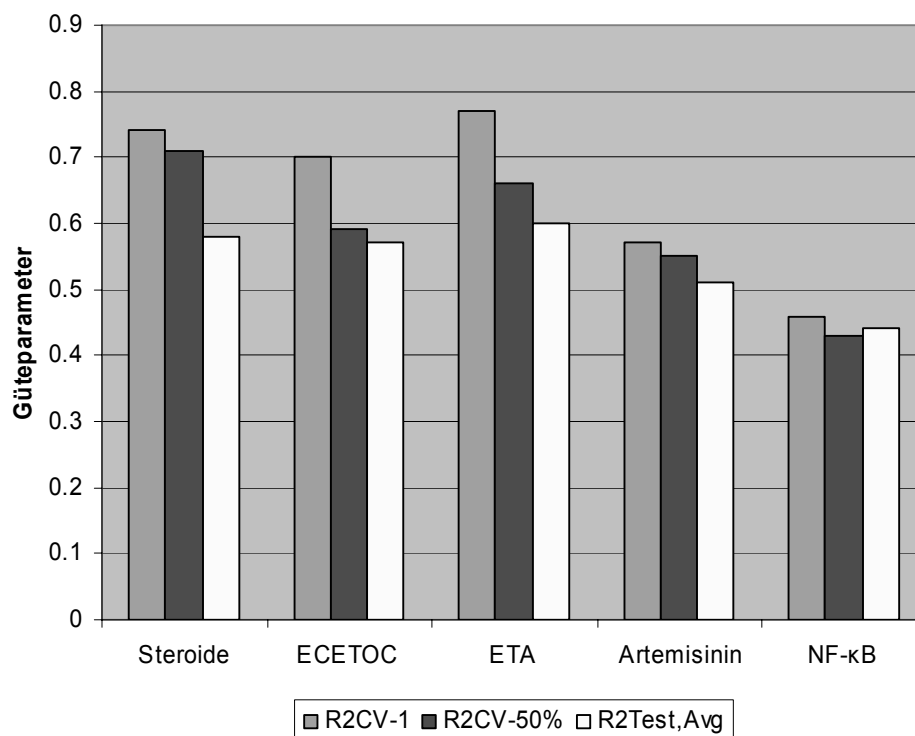


Abbildung 84: Die statistischen Parameter der weiteren untersuchten Datensätze im Überblick.

Bei dem Steroid-Datensatz [34,234], dem Standarddatensatz der QSAR-Methodenvalidierung [270], war beispielsweise eine Modellierung ohne die Herausnahme von Ausreißern möglich und das Modell für die sehr starren Substanzen ist gut. Dieser Datensatz besteht aus 31 Steroiden von denen die Bindungsaffinität zum Kortikosteroid-Bindenden-Globulin (Transcortin) bestimmt wurde.

Der ECETOC-Datensatz [235] besteht aus 38 Molekülen mit sehr diversen Grundstrukturen, die ein Alignment unmöglich machen. Hier wurde das augenirritierende Potenzial mittels des Draize-*in-vivo*-Augenirritationstests [271] bestimmt. Für diesen Datensatz konnte ein Modell in der gleichen Qualität wie bei MaP erzeugt werden.

Ähnliches gilt für den ET_A -Datensatz [236-238], der 36 Arylsulfonamiden besteht, die antagonistische Aktivität am Endothelin-Rezeptorsubtyp A (ET_A) aufweisen. Der ET_A -Datensatz war der erste Datensatz, der erfolgreich mit xMaP modelliert wurde.

Tabelle 18: Übersicht über die xMaP-Ergebnisse für die verschiedenen vorgestellten Datensätze.

	R^2_{CV-1}	$R^2_{CV-50\%}$	$R^2_{Test,Avg}$	# MIV	m
Steroide [34,234]	0.74	0.71	0.58	3	31
ECETOC [235]	0.70	0.59	0.57	2	37
ET _A [236-238]	0.77	0.66	0.60	4	36
Artemisinin [239]	0.57	0.55	0.51	8	204
NF- κ B [240,241]	0.46	0.43	0.44	14	112

Der Artemisinin-Datensatz mit 204 Analoga des Antimalaria-Wirkstoffes Artemisinin als ein sehr großer Datensatz mit relativ diversen Grundstrukturen diene als weiterer Benchmarkdatensatz. Hier waren sehr interessante Interpretationen möglich, die bereits bekannte Informationen sehr gut kodieren (nicht gezeigt) [239].

Außerdem wurde ein Datensatz von 112 Inhibitoren des Transkriptionsfaktors NF- κ B untersucht. Bei diesen NF- κ B-Daten [240,241] sind weitere Analysen erforderlich, die ersten Ergebnisse sind jedoch sehr viel versprechend.

4. Zusammenfassung und Ausblick

Die vorliegende Arbeit beschreibt die Entwicklung, Validierung und erfolgreiche Anwendung der interpretierbaren 4D-QSAR Methodik xMaP. Die neue Methode benötigt weder die Auswahl des vermuteten bioaktiven Konformers noch eine Überlagerung der Moleküle im Raum, sie ist also alignment-frei. xMaP ist invariant gegenüber Rotation, Translation und kodiert die Flexibilität der Moleküle. Dadurch wird der Einfluss durch den Benutzer praktisch ausgeschaltet.

Im Folgenden sollen die wichtigsten methodischen Elemente zusammen mit den für verschiedene Datensätze erzielten Ergebnissen noch einmal im Überblick dargestellt werden:

Wie bei den Vorgängertechniken bilden Radialverteilungsfunktionen die theoretische Grundlage für den xMaP-Deskriptor. Diese selektiven Distanz-Zählstatistiken können in einem einzigen Vektor die Verteilung einzelner Oberflächeneigenschaften zueinander, die Form der Moleküle und die Flexibilität der Moleküle beschreiben. Die xMaP-Variablen kodieren dabei die Größe und die Orientierung der verschiedenen Oberflächeneigenschaften als potenzielle Zweipunkt-Pharmakophore. Dabei wird die Information des gesamten Konformerensembles abgebildet.

Zunächst wird für jedes Molekül ein Konformerensembles berechnet, welches die Grundlage für alle weiteren Berechnungen bildet. Damit erfolgt der Schritt von der dritten in die vierte Dimension. Diese Konformerberechnung kann gewöhnlich ohne Information über die biologische Zielstruktur erfolgen. Wenn jedoch Information über die Zielstruktur der Substanzen vorhanden ist, kann diese im Rahmen eines Dockings auch in die Berechnung der Konformere einfließen. Die nachfolgend beschriebenen Schritte für die Oberflächenberechnung erfolgen für alle Konformere aller Moleküle:

Zunächst wird eine Oberfläche aus gleichverteilten Punkten berechnet. xMaP setzt hier einen Punkt-Punkt-Abstand von 0.4 \AA ein. Dadurch wird der Diskretisierungsfehler, der bei der Beschreibung einer kontinuierlichen Oberfläche durch verteilte Oberflächenpunkte entsteht, vernachlässigbar. Die Oberflächenpunkte werden in fünf verschiedene Eigenschaftsklassen klassifiziert: Es wird die Hydrophilie respektive Hydrophobie beschrieben, die anhand der Eigenschaften des nächstliegenden Atoms zugewiesen wird. Regionen, die Wasserstoffbrückenbindungen ausbilden können tragen zusätzlich die Information darüber. Es erfolgt also teilweise eine doppelte Eigenschaftszuweisung.

Um die Rohdaten der Oberflächenbeschreibung in den Vektor überführen zu können sind mehrere Schritte erforderlich. Ausgehend von den gleichverteilten Oberflächenpunkten mit verschiedenen Eigenschaften werden Regionen mit identischen Eigenschaften durch einen

rekursiven Suchalgorithmus verschmolzen. Damit werden Oberflächenareale, die so genannten Patches, erhalten. Eine Moleküloberfläche kann durch die Verteilung dieser Eigenschaftsareale charakterisiert werden. Auf Basis des geometrischen Schwerpunkts und der Größe der Areale werden potenzielle Zweipunkt-Pharmakophore gebildet und die Information darüber in den Zählvektor überführt. Dabei wird eine erweiterte unscharfe Zählweise verwendet, um die Größe der Patches und die Flexibilität besser zu beschreiben. Das Ergebnis ist jeweils ein Vektor mit verschiedenen zuvor festgelegten potenziellen Zweipunkt-Pharmakophoren, die sich durch Eigenschaften und Abstände der eigenschaftstragenden Oberflächenareale unterscheiden. Danach wird durch eine Mittelwertbildung der Vektoren aller Konformere ein Vektor berechnet, der genau ein Molekül beschreibt, so dass nach Kodierung jeden Moleküls jeweils ein Vektor pro Molekül aus dem untersuchten Datensatz vorliegt. Dieser Vektor trägt sämtliche Information über Oberflächeneigenschaften und deren Verteilung im gesamten Konformerensemble.

Zur Modellbildung werden die informativsten Variablen, die jeweils einem potenziellen Zweipunkt-Pharmakophor entsprechen, mit Hilfe der „Reverse-Elimination-Method“-Tabu-Suche identifiziert. Diese Variablen können in den ursprünglichen Datenraum der Moleküle zurückprojiziert werden. Dadurch können die Modelle gut interpretiert und einem Medizinischen Chemiker kommuniziert werden. Eine derartige Darstellung ist sowohl im 2D- als auch im 3D-Raum gut machbar, wobei die Information aus vier Dimensionen beschrieben wird. Da es bei dieser Variablenselektion zu Zufallskorrelationen kommen kann, werden die erhaltenen Modelle sehr streng validiert. Pro Modell werden viele Unterteilungen (Standard: 100) der Daten in Trainings- und Testdaten vorgenommen. Auf Grundlage der Trainingsdaten wird dabei mit der „Leave-multiple-out“-Kreuzvalidierung als Gütefunktion bei der Variablenselektion ein Modell erstellt und danach die Testdaten vorhergesagt. Daraus kann für jede Vorhersage ein so genannter Ensemble-Mittelwert berechnet werden. Diese Werte dienen dann zur Berechnung eines Korrelationskoeffizienten, der auf 100 Modellberechnungen basiert. Wenn ein Modell all diesen Schritten standhält, so kann eine Zufallskorrelation ausgeschlossen werden.

Um die Anwendbarkeit und Qualität des xMaP-Deskriptors zu überprüfen wurden verschiedene Datensätze detailliert untersucht. Zunächst wurde für verschiedene Parameter wie Konformerenauswahl und Obergrenzen für die innere Energie von Konformeren eine ausführliche Validierung der Technik durchgeführt. Damit wurden die optimalen Standardparameter bestimmt, die in späteren Analysen unverändert angewendet wurden. Die untersuchten Datensätze können aufgrund der jeweiligen Zielsetzung in unterschiedliche Bereiche aufgeteilt werden. Die ersten drei Datensätze wurden als Benchmark zum Vergleich

mit etablierten QSAR-Techniken genutzt: Für die Acetylcholinesterase-Inhibitoren (AZT) liegen Untersuchungen mit vielen verschiedenen dreidimensionalen QSAR-Techniken vor, außerdem sind diese Moleküle sehr flexibel. Zusätzlich ist hier die Kristallstruktur des Zielenzym bekannt, so dass auch ein strukturbasiertes Modell erstellt werden konnte. Der PGF_{2α}-Datensatz besteht aus sehr flexiblen Molekülen die einen Vergleich mit der alignment-abhängigen – und bisher einzigen existierenden – 4D-QSAR-Technik von Hopfinger ermöglichen. Die allosteren Modulatoren des muskarinischen M₂-Rezeptors zeigen, dass xMaP bei extrem flexiblen Substanzen eine identische Modellqualität im Sinne von Statistik und Interpretierbarkeit wie seine Vorgängertechnik erreicht, obwohl die Information über die Flexibilität mit in der Analyse berücksichtigt wird. Für Inhibitoren der HIV1-Reversen Transkriptase konnte wiederum ein ligand- und ein strukturbasiertes Modell erstellt werden. Damit konnte ein Einblick in die Komplementarität der xMaP-Eigenschaften mit den Gegebenheiten in der Bindetasche gefunden werden, was die gute Anwendbarkeit weiter untermauert. Besonders interessant sind die Dopamin-D₁-Antagonisten: Für diese Daten war es bisher nicht möglich, ein QSAR-Modell aufzustellen. Mit xMaP konnte problemlos ein Modell etabliert und interpretiert werden. Dieses Modell wurde genutzt, um die Aktivität noch nicht getesteter Substanzen vorherzusagen. Dabei stellte sich heraus, dass das Modell in der Realität gut anwendbar ist. Außerdem konnten neue Substanzen zur Synthese vorgeschlagen werden. Die experimentellen Ergebnisse der zu synthetisierenden Moleküle stehen jedoch größtenteils noch aus. Zusätzlich wurde ein Homologiemodell des Dopamin-D₁-Rezeptors genutzt, um die Interpretation des QSAR-Modells zu verbessern. Zusammen mit einer Sequenzanalyse der Dopamin-Rezeptoren konnte ein Mechanismus-Modell für die Wirkweise der untersuchten Substanzen erstellt werden. Dieses Modell beschreibt die Struktur-Wirkungs-Beziehungen sehr gut. Des Weiteren wurde ein Datensatz von Glukokortikoiden untersucht. Auch dafür gab es bisher kein QSAR-Modell, so dass zusätzlich auch CoMFA- und CoMSIA-Modelle erstellt wurden. xMaP konnte hier die wesentlichen Strukturparameter ohne weiteres identifizieren und ein Ergebnis ähnlich zu dem der etablierten Techniken liefern. Die Untersuchung von weiterer Datensätze untermauerte die breite Anwendbarkeit von xMaP in allen Gebieten der Wirkstoffforschung.

Die gefundenen Modelle zeigen, dass die im Rahmen dieser Arbeit entwickelte Technik, sich durch ihre guten Modelle sowie die gute Interpretier- und Kommunizierbarkeit dieser Modelle auszeichnet. Dies ist möglich, obwohl Schritte wie das Alignment und die Auswahl eines vermuteten biologisch aktiven Konformers nicht mehr nötig sind. xMaP ist also eine vollwertige QSAR-Technik aus einer Kombination von Deskriptor, Modellierung, Modellvalidierung und Visualisierung. Dass vom Benutzer keine Prozedur-Parameter außer

der Wahl der Art und Weise der Konformerengeneration festgelegt werden müssen macht xMaP im Hinblick auf die Robustheit einzigartig.

Obwohl xMaP Modelle in einer Qualität und Interpretierbarkeit liefert, die denen etablierter Techniken wie CoMFA in nichts nachstehen, sind aufgrund der einfachen und hocheffizienten mathematischen Grundlagen verschiedene Weiterentwicklungen denkbar:

Momentan muss man bei xMaP wie bei allen translations- und rotationsinvarianten Techniken bei der Interpretation mehrere interessante Moleküle anschauen, um entsprechend einen Einblick in die Struktur-Wirkungs-Beziehungen des Datensatzes zu finden. Dies ist etwas zeitaufwändiger als bei alignment-abhängigen Techniken. Eine Weiterentwicklung zu einer einzigen Darstellung, die analog zu alignment-abhängigen Techniken die vollständige Information des Modells vermittelt, sollte mit den in der Variablenselektion gefundenen Daten als Grundlage möglich sein. Dafür müssten basierend auf einem als essenziell identifizierten potenziellen Zweipunkt-Pharmakophor alle Moleküle anhand ihrer wichtigen Oberflächenareale überlagert werden. Man würde direkt sehen, welche Molekülteile relevant sind. Auch noch nicht synthetisierte oder getestete Moleküle könnten in eine solche Darstellung ohne weiteres eingefügt werden, so dass eine Interpretation analog zu den CoMFA-Feldern möglich wäre.

Derzeit ist der xMaP-Deskriptor auf potenzielle Zweipunkt-Pharmakophore beschränkt. Es werden Verbindungen zwischen jeweils zwei Oberflächenarealen beschrieben. Es könnte dabei passieren, dass Beiträge zu einer Variablen aus verschiedenen Molekülteilen stammen. Dies könnte theoretisch zu Inkonsistenzen bei für die Rezeptorinteraktion relevanten Variablen führen, weswegen eine Erweiterung auf potenzielle Mehrpunkt-Pharmakophore denkbar ist. Wegen der redundanten Informationsbeschreibung in den xMaP-Deskriptoren und der sehr genauen Validierung, ist jedoch zu erwarten, dass dieses Problem nicht gravierend ist. Bei einer Verwendung von potenziellen Mehrpunkt-Pharmakophoren dagegen ist ein starker Anstieg der Wahrscheinlichkeit für Zufallskorrelationen zu erwarten, so dass man hier abwägen muss, welches Ziel wichtiger ist. Für eine sinnvolle Anwendung ist wohl die bisherige Methode optimal, eine Evaluierung von Mehrpunkt-Pharmakophoren sollte aber in Erwägung gezogen werden.

Summary and Outlook

This thesis describes the development, validation and successful application of the interpretable 4D-QSAR technique xMaP. The novel method does neither rely on the selection of a presumed bioactive conformer nor on a spatial superimposition of the molecules which means that it is so-called alignment-free. Put differently, xMaP is invariant to rotation and translation and encodes the flexibility of the molecules under scrutiny. By combining these features a possible user bias is almost completely eliminated.

In the following the most important methodological basics as well as the obtained results will be summarized:

Radial distribution functions form the theoretical basis for the xMaP descriptor. These selective distance count statistics in their employed form encode the distribution of molecular surface properties, the shape of the molecules as well as the molecules' flexibility in a single vector. Thus, the information carried by many low-energy conformers covering the conformational space one molecule can adopt is encoded. To achieve this, xMaP variables store the size and orientation of particular surface properties as potential two-point-pharmacophores for each conformer. Afterwards the information of the conformers is merged to a unitary description of the entire conformer family.

The detailed procedure is as follows: First, for each molecule under scrutiny a conformational ensemble is computed, which is the basis for all further calculations. This step distinguishes 3D-QSAR from 4-D-QSAR methods. The conformer calculation is typically performed without information about the biological target. However, if target structure information is available it can be incorporated in the calculation through docking experiments. The best docking solutions are used as starting point for further investigations.

In the next step a surface of equally distributed points is computed. When calculating the xMaP descriptor a resolution (point-point-distance) of 0.4 Å is used for approximating the molecular surface. This renders the discretization error negligible, which results from the conversion of a contiguous surface into surface points. The surface points are afterwards classified into five property classes. Apart from the H-bonding potential (properties: H-bond donor, H-bond acceptor), the molecular lipophilic potential (properties: hydrophilic, weakly hydrophobic, strongly hydrophobic) is assigned to each point. Hence, a single surface point may be assigned two properties as a result of this calculation.

In order to encode the surface property data in a molecular descriptor, several steps are required. Starting from the equally distributed surface points carrying different properties, regions with identical properties are merged by a recursive search algorithm. As a result,

surface areas, the so-called patches, are obtained. Consequently, the molecular surface can be characterized by the patches' distribution. The geometric centers of mass and the patch size are employed for creating potential two-point-pharmacophores. The information about the occurrence of two-point pharmacophores with different property combinations and different distances between the geometric centers of mass is transferred to the radial distribution function. The value of an element of the function describes a distinct property-property-combination in a defined distance range and is incremented by the product of the patch sizes describing the respective two-point-pharmacophore. Additionally, the neighboring distance bins are incremented with a certain percentage of the main bin's increment. This so-called extended fuzzy increment was used to avoid discretization errors due to the formation of distance bins and to better encode the molecule's flexibility. A single vector for every conformer results from this step. Afterwards a vector describing one molecule with all its conformers is determined by calculating the mean value of the vectors of all conformers. Different techniques to integrate the information of all conformers were evaluated and were found to be inferior. The resulting vector carries the whole information about surface properties and their distribution in the entire conformer ensemble.

During model building the most informative variables, i.e. single two-point-pharmacophores, are identified by employing the "Reverse-Elimination-Method"-Tabu-Search. These most informative variables can easily be backprojected onto the molecules, facilitating the interpretation of the model. This projection is possible into 2D-space as well as into 3D-space, although four-dimensional information is displayed.

Possible chance correlations are major challenges that have to be faced in data modeling. To lower the probability of chance correlations, a stringent validation procedure is applied to xMaP models. During the model building process a large number of splits in training and test set are computed (100 by default). Using the training data of a particular split as a basis a model is established using variable selection with the "leave-multiple-out"-cross-validation as the objective function. This model is employed to predict the respective test data. The prediction values are afterwards used to calculate the ensemble mean value. These values for all molecules are then used to calculate a statistical figure of merit for the model quality based on all 100 splits. If a model withstands all these steps, a chance correlation is unlikely.

To validate the applicability and quality of the xMaP-descriptor and the models obtained with it, a number of datasets was investigated. Initially, for various parameters such as conformer selection and limits for the internal energy of conformers an extensive validation of the technique was performed in order to determine the best standard parameters. The parameters found during this process were used in the same manner in all further analyses. The analyzed

datasets can be classified in different areas of research depending on the particular purpose. The first three datasets were used to benchmark xMaP against established QSAR techniques. They consisted of the following datasets. First, acetylcholinesterase inhibitors (AZT) were chosen because these data have been investigated thoroughly with numerous 3D techniques. In addition, all the molecules are highly flexible and the crystal structure of the target is known. Hence, a structure-based model could be compared to the merely ligand-based procedure. Second, the PGF_{2α} dataset was used because it comprises of highly flexible compounds that permit a comparison with the alignment-dependent 4D-QSAR technique by Hopfinger. Third, allosteric modulators of the muscarinic M₂-receptor were included in the benchmark set to demonstrate that xMaP generates models of identical quality in terms of statistical figures of merit and interpretability as its predecessor technique with the additional advantage that the molecular flexibility is incorporated in the analysis.

Additional validation datasets were modeled to investigate specific characteristics of the novel descriptor. For inhibitors of HIV-1 reverse transcriptase a ligand as well as a structure-based model could be established. The information about relevant parameters was backprojected to the binding pocket. Thus, the complementarity of the xMaP properties with the properties of the binding pocket under scrutiny could be shown. A dataset of dopamine antagonists are of special interest: up to now, no QSAR model could be established for these substances. By employing xMaP this was done with ease (i.e. standard parameters). The model was interpreted and used to predict the activity of compounds that have not yet been tested. It was shown that xMaP can readily be used in everyday applications. Additionally, new substances were proposed for synthesis. The experimental results of molecules to be synthesized are anticipated in the near future. Additionally, a homology model of the dopamine D₁-receptor was used to advance the model interpretation. Combining these results with a sequence analysis of the different dopamine receptors, a model for the mode of action of the scrutinized molecules was established. This combined model reflects the structure-activity relationships very well, too. Moreover, a dataset of glucocorticoids was studied. Again, no previous model existed. Therefore, both CoMFA- and CoMSIA-models were generated in addition. By using xMaP, it was possible to identify the relevant structural parameters and gain information which is comparable to that of the renowned techniques. Investigations of further datasets delivered more evidence of xMaP's broad applicability in all areas of medicinal chemistry research.

The established models confirm that xMaP generates good results with easily interpretable models which can be straightforwardly communicated to medicinal chemists. It should be recalled that an alignment step and the selection of a presumably bioactive conformer is not

necessary to arrive at these results. Hence, xMaP is a full-fledged QSAR technique based on a combination of molecular descriptor, mathematical modeling procedure, model validation and model visualization. The fact that a user does not have to choose procedural parameters beside the selection of the conformer generation method renders xMaP unique in terms of robustness.

Even though xMaP creates models with quality and interpretability that are comparable to well-known techniques such as CoMFA, some extensions are conceivable owing to xMaP's simple but highly efficient fundamental mathematical principles:

As all translationally and rotationally invariant techniques, it is necessary with xMaP to inspect several molecules to be able to interpret the model at hand. This step is slightly more time-consuming as compared to the interpretation of the results of alignment-dependent techniques. Further development towards a graphical display of the model for all molecules at ones should be easy to integrate in the standard workflow based on the most important variables. It should be possible to superimpose all molecules according to particular important two-point-pharmacophores which are represented by surface patches to show which parts of the molecules are important. Even molecules not yet synthesized or tested could be easily integrated in such a representation. Consequently; an interpretation analogous to that of CoMFA could become possible.

At the moment the xMaP-descriptor is restricted to connections between two points; only connections between two patches are described. A numerical artifact of this constraint is that the same variable might be incremented by surface areas of completely different parts of the molecule. Theoretically, if not all identified surface areas encode properties that are relevant for the ligand-receptor interactions this could lead to inconsistencies of the resulting model. Consequently, an extension of the descriptor to more-point-pharmacophores is conceivable. By using more-point-pharmacophores a sharp increase of the number of variable and of noisy variables and thus an increase in the probability of chance correlations can be expected. Therefore, using more-point-pharmacophores does not guarantee an increase in encoded information. Probably, the current way of encoding the properties as two-point pharmacophores is the most suitable compromise of encoded information and the amount of noisy variables.

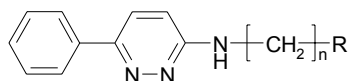
Anhang I. Wichtige Abkürzungen und Symbole

Å	Ångström (10 nm)
2D	zweidimensional
3D	dreidimensional
4D	vierdimensional
Abb.	Abbildung
AChE	Acetylcholinesterase
CATS	Chemically Advanced Template Search
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis
CV	Kreuzvalidierung (engl. <i>Cross-validation</i>)
DiP	Distance Profile
DPMA	Di-Phenyl-Methyl-Amin
J	Joule
kcal	Kilokalorien
LMO	Lass-mehrere-Objekte-heraus (engl. <i>Leave-Multiple-Out</i>)
logP	logarithmierter Oktanol-Wasser-Verteilungskoeffizient
LOO	Lass-ein-Objekt-heraus (engl. <i>Leave-One-Out</i>)
MaP	Mapping Property distributions of molecular surfaces
MIF	Molekulares Interaktionsfeld (engl. <i>Molecular Interaction Field</i>)
MIV	Informativste Variable (engl. <i>most informative variable</i>)
MLR	Multiple lineare Regression
PCR	Hauptkomponentenregression (engl. <i>Principal component regression</i>)
PDB	Protein Data Bank
PLS	Partial Least Squares
QSAR	Quantitative Struktur-Wirkungs-Beziehungen (engl. <i>Quantitative Structure-Activity Relationships</i>)
RCSB	Research Collaboration for Bioinformatics
RMSD	Root Mean Square Deviation
RRA	Relative Rezeptor-Affinität
SDZS	Selektive Distanz-Zählstatistik
SESP	Start-End-Shortest-Path
SVD	Singular Value Decomposition (Singulärwertzerlegung)
TRI	translations- und rotationsinvariant
TS	Tabu-Suche
vgl.	Vergleiche
VS	Variablenselektion

Anhang II. Strukturen und biologische Aktivitäten der untersuchten Datensätze

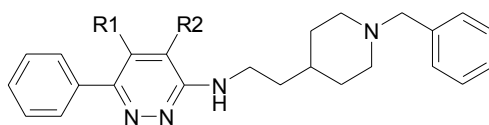
II.1. Inhibitoren der Acetylcholinesterase (AZT)

Tabelle 19: Gruppe 1 des AZT-Datensatzes und deren biologische Aktivität.



Nr.	<i>N</i>	R	pIC ₅₀ (-log ₁₀ [nM])
1	2		3.10
2	3		3.39
3	2		4.19
4	2		4.08
5	3		4.46
6	4		4.82
7	5		5.00
8	2		3.77
9	3		4.89
10	4		4.96
11	5		6.13
12	0		4.19
13	1		5.25
14	2		6.92

Tabelle 20: Gruppe 2 des AZT-Datensatzes und deren biologische Aktivität.



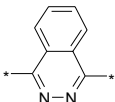
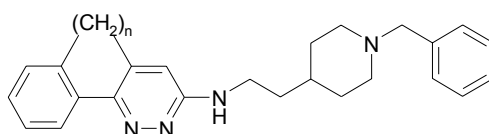
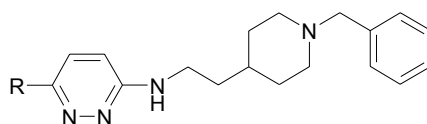
Nr.	R1	R2	pIC ₅₀
15	H	Me	6.49
16	H	<i>i</i> -Pr	6.37
17	Me	H	7.68
18	Et	H	7.57
19	Pr	H	7.21
20		-	6.36

Tabelle 21: Gruppe 3 des AZT-Datensatzes und deren biologische Aktivität.



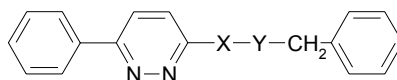
Nr.	<i>n</i>	pIC ₅₀
21	3	7.66

Tabelle 22: Gruppe 4 des AZT-Datensatzes und deren biologische Aktivität.



Nr.	R	pIC ₅₀
22	H	6.62
23	Cl	7.14
24	MeO	6.66
25	2-Me-Ph	7.05
26	2-Et-Ph	7.06
27	2,4,6-(Me) ₃ -Ph	5.53
28	2-MeO-Ph	6.96
29	2-Cl-Ph	7.10
30	3,5-(CF ₃) ₂ -Ph	7.25
31	2-Thiophenyl	7.01
32	3-Pyridinyl	7.24

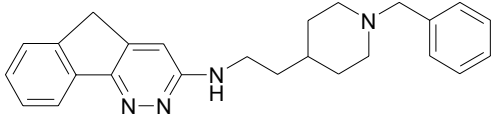
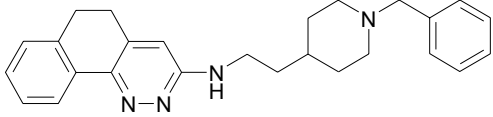
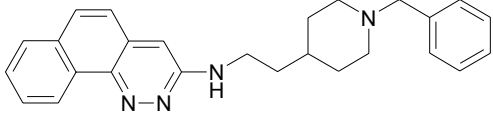
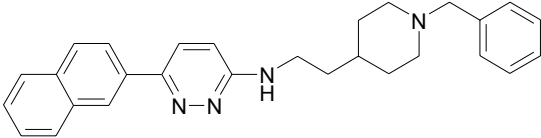
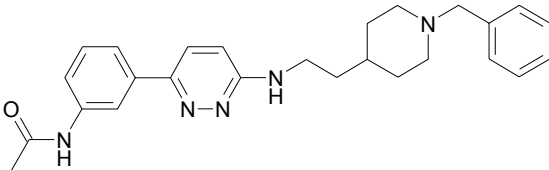
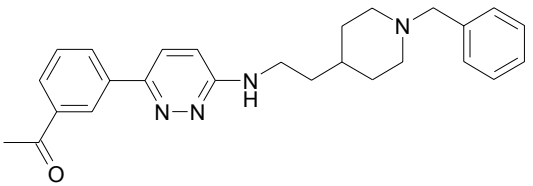
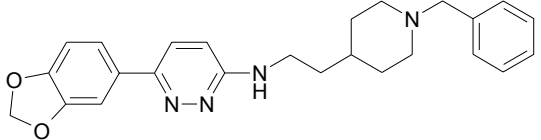
Tabelle 23: Gruppe 5 des AZT-Datensatzes und deren biologische Aktivität.



Nr.	X	Y	pIC ₅₀
33	OCH ₂ CH ₂		6.85
34	SCH ₂ CH ₂		7.20
35	NHCOCH ₂		5.38
36	NHCH ₂ CH ₂		5.82
37	NHCOCH ₂		4.77
38	NHCH ₂ CO		4.82
39	NHCH ₂ CH ₂		4.62
40	NHCOCH ₂		4.74
41	NHCH ₂ CO		3.92
42 ^a	-	-	3.21

^a Minaprin.

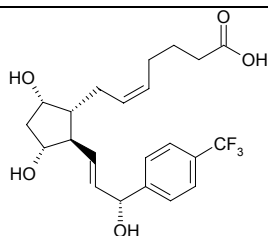
Tabelle 24: Testdaten des AZT-Datensatzes (in der Originalpublikation) und deren biologische Aktivität.

Nr.	Struktur	pIC ₅₀
43		8.00
44		7.41
45		7.66
46		7.24
47		7.24
48		7.27
49		7.14

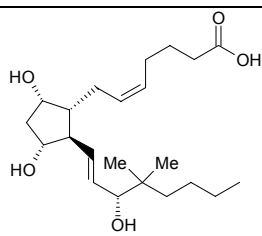
II.2. Prostaglandin F_{2α}-Analoga (PGF_{2α})

Tabelle 25: Die Strukturen des PGF_{2α}-Datensatzes (Nummer 1 bis 12) und deren biologische Aktivität Y

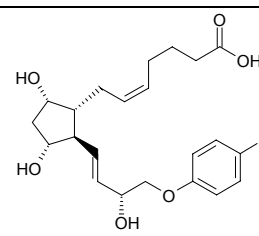
$$(Y = \log_{10} \left(\frac{ED_{50, \text{Substanz}}}{ED_{50, \text{PGF}_{2\alpha}}} \right)).$$



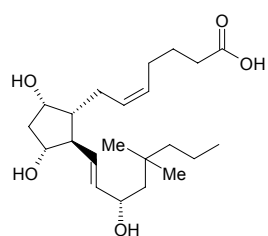
(1) 1.699



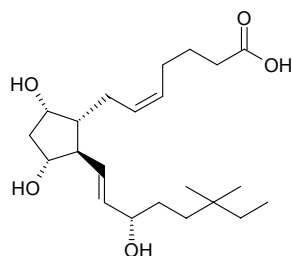
(2) 0.081



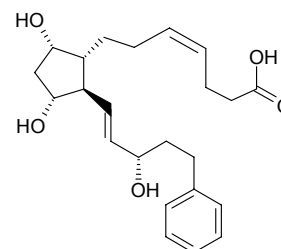
(3) 2.301



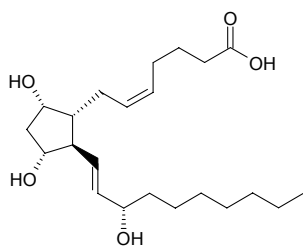
(4) 0.279



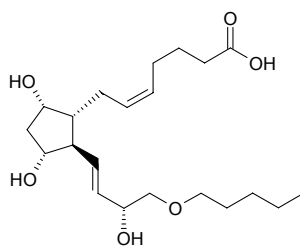
(5) 0.664



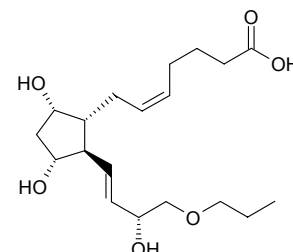
(6) 2.301



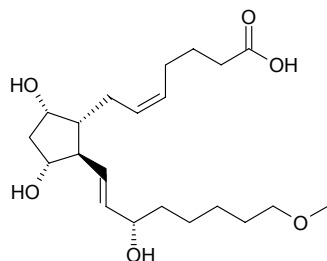
(7) 0.699



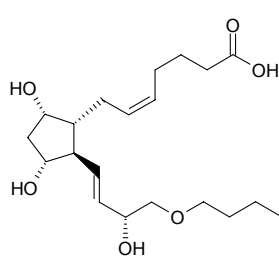
(8) 1.000



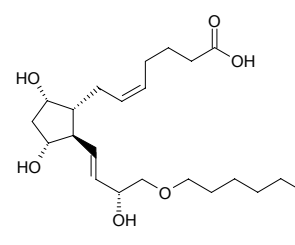
(9) 0.398



(10) 0.398

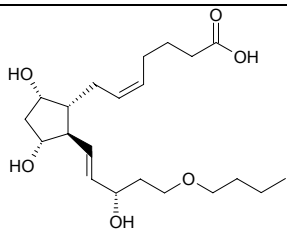


(11) 1.000

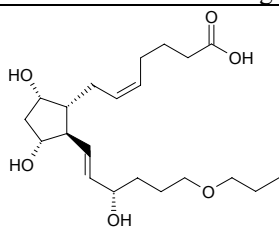


(12) 0.699

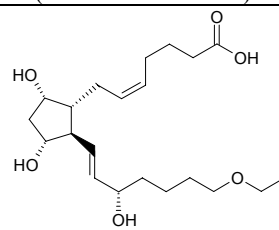
Tabelle 26: Die Strukturen des PGF_{2α}-Datensatzes und deren biologische Aktivität (Nummer 13 bis 24).



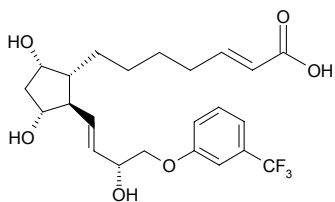
(13) -0.301



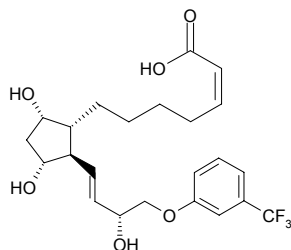
(14) -0.602



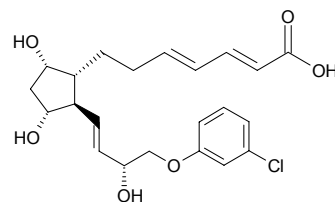
(15) 0.699



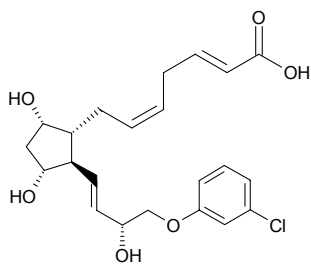
(16) 0.602



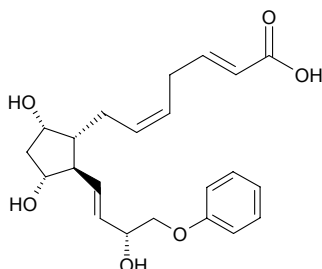
(17) 0.482



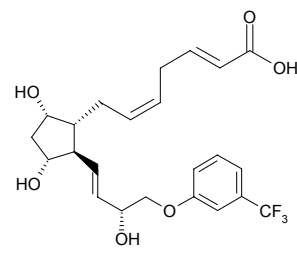
(18) 0.777



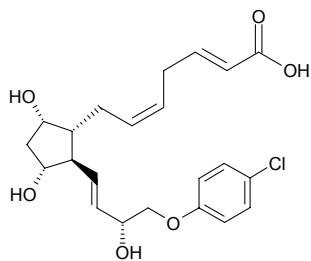
(19) 2.301



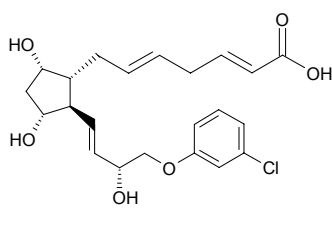
(20) 2.481



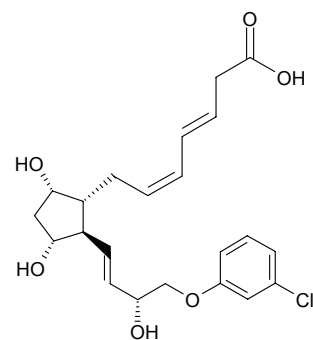
(21) 2.000



(22) 2.000

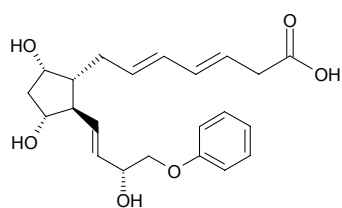


(23) 2.000

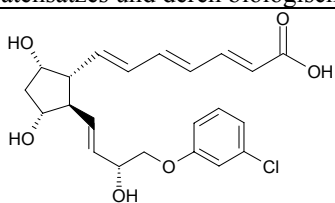


(24) 2.886

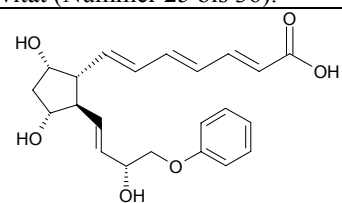
Tabelle 27: Die Strukturen des PGF_{2α}-Datensatzes und deren biologische Aktivität (Nummer 25 bis 36).



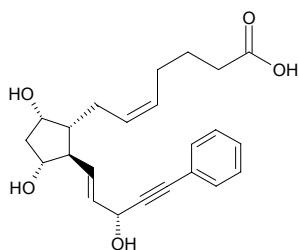
(25) 2.301



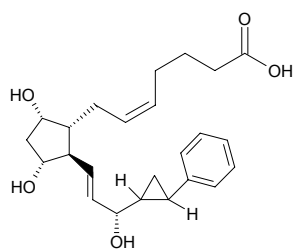
(26) 0.000



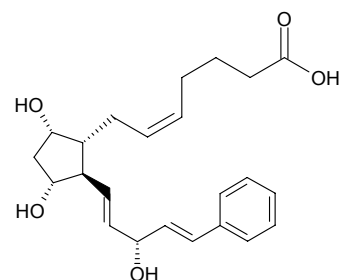
(27) 0.301



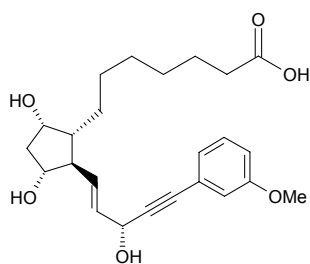
(28) 1.699



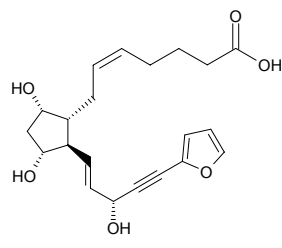
(29) 1.669



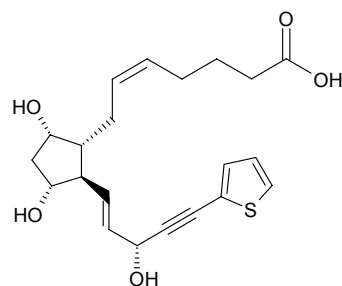
(30) 0.699



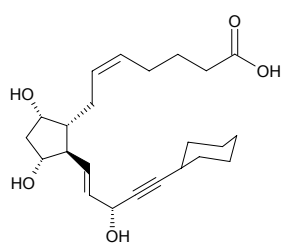
(31) 0.669



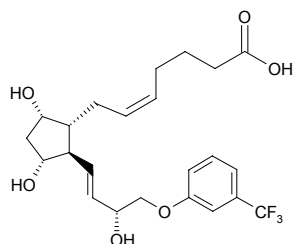
(32) 1.699



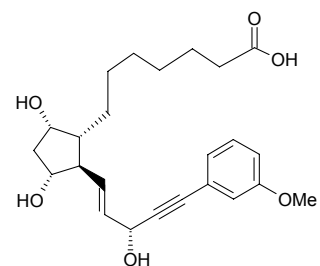
(33) 1.398



(34) -0.602

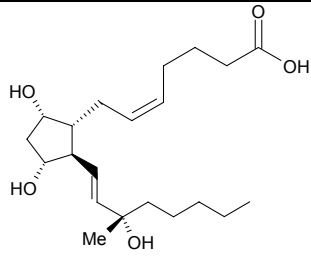


(35) 2.000

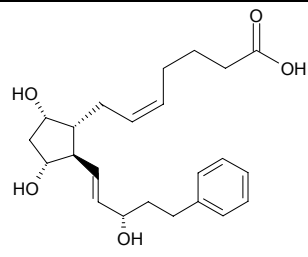


(36) 0.000

Tabelle 28: Die Strukturen des PGF_{2α}-Datensatzes und deren biologische Aktivität (Nummer 37 und 38).



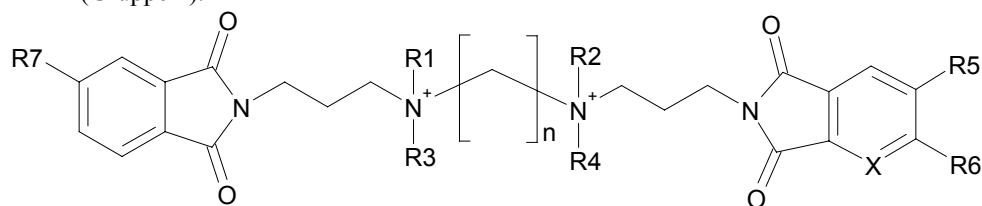
(37) 0.602



(38) 1.959

II.3. Modulatoren des muskarinischen M₂-Rezeptors (M₂)

Tabelle 29: Chemische Strukturen und biologische Aktivitäten der muskarinischen M₂-Rezeptor-Modulatoren (Gruppe 1).



No.	R1	R2	R3	R4	R5	R6	R7	X	n	pEC ₅₀ (-log ₁₀ [nM])
1	Me	Me	Me	Me	Me	H	H	C	6	6.490
7	Me	Me	Me	Me	H	H	H	C	6	5.842
11	Me	Me	Me	Me	H	H	H	C	3	5.420
12	Me	Me	Me	Me	H	H	H	C	4	5.570
13	Me	Me	Me	Me	H	H	H	C	5	5.857
14	Me	Me	Me	Me	H	H	H	C	7	6.409
15	Me	Me	Me	Me	H	H	H	C	8	6.244
16	Me	Me	Me	Me	H	H	H	C	10	6.276
34	H	H	C ₃ H ₆ Ph	C ₃ H ₆ Ph	H	H	H	C	6	6.004
35	H	H	C ₂ H ₄ CN	C ₂ H ₄ CN	H	H	H	C	6	6.131
36	H	H	Cyc-C ₆ H ₁₁	Cyc-C ₆ H ₁₁	H	H	H	C	6	7.347
37	H	H	Cyc-C ₆ H ₁₁	Cyc-C ₆ H ₁₁	Me	H	Me	C	6	7.222
38	Me	Me	Me	Me	H	H	H	N	6	5.630
39	Me	Me	Me	Me	Cl	Cl	H	C	6	6.940
40	Me	Me	Me	Me	F	F	H	C	6	6.340

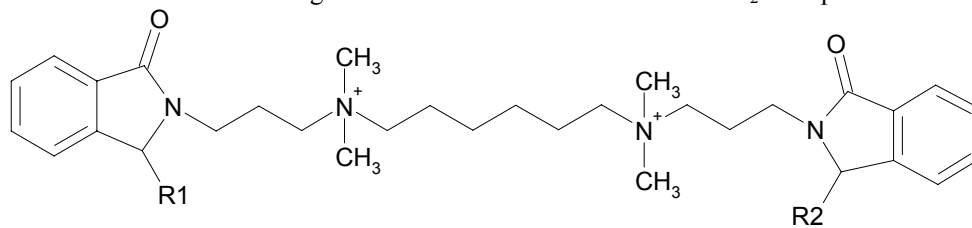
Tabelle 30: Chemische Strukturen und biologische Aktivitäten der muskarinischen M₂-Rezeptor-Modulatoren (Gruppe 2).

No.	R1	R2	R3	R4	n	R5	R6	pEC ₅₀
5	Me	Me	Me	Me	7			4.917
6	Me	Me	Me	Me	7			4.495
9	Me	Me	Me	Me	6			6.125
19	Me	Me	Me	Me	7			6.097
20	Me	Me	Me	Me	7			5.347
21	Me	Me	Me	Me	7			5.658

Tabelle 31: Chemische Strukturen und biologische Aktivitäten der muskarinischen M₂-Rezeptor-Modulatoren (Gruppe 2, Fortsetzung).

No.	R1	R2	R3	R4	n	R5	R6	pEC ₅₀
22	Me	Me	Me	Me	7			3.821
41	Me	Me	Me	Me	6			6.830
42	Me	Me	Me	Me	6			7.200
43	Me	Me	Me	Me	6			6.400
44	Me	Me	Me	Me	6			5.010

Tabelle 32: Strukturen und biologische Aktivitäten der muskarinischen M₂-Rezeptor Modulatoren (Gruppe 3).



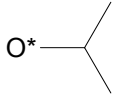
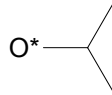
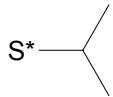
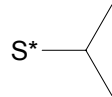
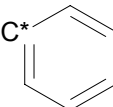
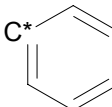
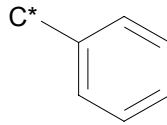
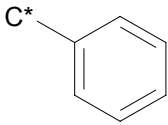
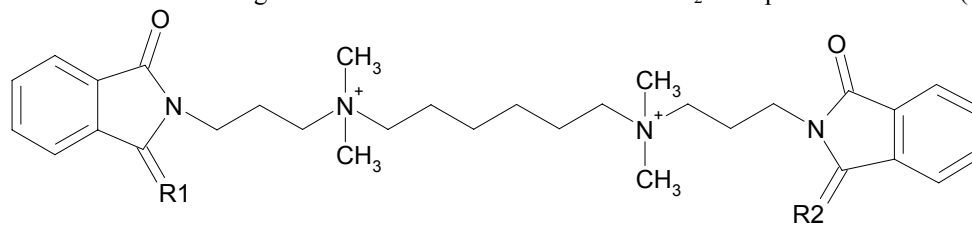
No.	R1	R2	pEC ₅₀
23	H	H	5.700
24	OH	OH	5.220
25*	OMe	OMe	5.170
26	OEt	OEt	5.640
27*			5.800
28			5.900
29			6.100
33			6.000

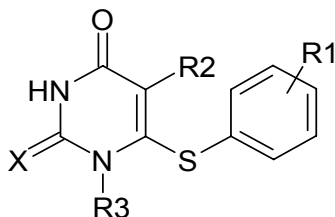
Tabelle 33: Strukturen und biologische Aktivitäten der muskarinischen M₂-Rezeptor Modulatoren (Gruppe 4).



No.	R1	E/Z-Isomerie an R1	R2	E/Z-Isomerie an R2	pEC ₅₀
30		E		E	7.150
31		Z		Z	7.07
32		E		Z	7.18

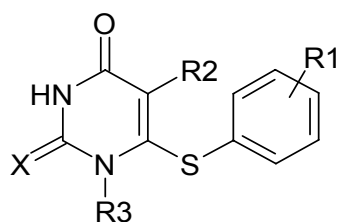
II.4. Inhibitoren der HIV-1 Reversen Transkriptase (HEPT)

Tabelle 34: Chemische Strukturen der HIV1-RT-Inhibitoren



No.	R1	R2	R3	X	pIC ₅₀ (-log ₁₀ [nM])
1	2-Me	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.15
2	2-NO ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.85
3	2-OMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.72
4	3-Me	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.59
5	3-Et	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.57
6	3- <i>t</i> Bu	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.92
7	3-CF ₃	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.35
8	3-F	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.48
9	3-Cl	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.89
10	3-Br	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.24
11	3-I	Me	CH ₂ OCH ₂ CH ₂ OH	O	5
12	3-NO ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.47
13	3-OH	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.09
14	3-OMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.66
15	3,5-Me ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	6.59
16	3,5-Cl ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.89
17	3,5-Me ₂	Me	CH ₂ OCH ₂ CH ₂ OH	S	6.66
18	3-COOMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.1
19	3-COMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.14
20	3-CN	Me	CH ₂ OCH ₂ CH ₂ OH	O	5
21	H	CH ₂ CH=CH ₂	CH ₂ OCH ₂ CH ₂ OH	O	5.6
22	H	Et	CH ₂ OCH ₂ CH ₂ OH	S	6.96
23	H	Pr	CH ₂ OCH ₂ CH ₂ OH	S	5
24	H	<i>i</i> Pr	CH ₂ OCH ₂ CH ₂ OH	S	7.23
25	3,5-Me ₂	Et	CH ₂ OCH ₂ CH ₂ OH	S	8.11
26	3,5-Me ₂	<i>i</i> Pr	CH ₂ OCH ₂ CH ₂ OH	S	8.3
27	3,5-Cl ₂	Et	CH ₂ OCH ₂ CH ₂ OH	S	7.37
28	H	Et	CH ₂ OCH ₂ CH ₂ OH	O	6.92
29	H	Pr	CH ₂ OCH ₂ CH ₂ OH	O	5.47
30	H	<i>i</i> Pr	CH ₂ OCH ₂ CH ₂ OH	O	7.2
31	3,5-Me ₂	Et	CH ₂ OCH ₂ CH ₂ OH	O	7.89
32	3,5-Me ₂	<i>i</i> Pr	CH ₂ OCH ₂ CH ₂ OH	O	8.57
33	3,5-Cl ₂	Et	CH ₂ OCH ₂ CH ₂ OH	O	7.85
34	4-Me	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.66
35	H	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.15
36	H	Me	CH ₂ OCH ₂ CH ₂ OH	S	6.01
37	H	I	CH ₂ OCH ₂ CH ₂ OH	O	5.44
38	H	CH=CH ₂	CH ₂ OCH ₂ CH ₂ OH	O	5.69
39	H	CH=CHPh	CH ₂ OCH ₂ CH ₂ OH	O	5.22
40	H	CH ₂ Ph	CH ₂ OCH ₂ CH ₂ OH	O	4.37
41	H	CH=CPh ₂	CH ₂ OCH ₂ CH ₂ OH	O	6.07
42	H	Me	CH ₂ OCH ₂ CH ₂ OMe	O	5.06
43	H	Me	CH ₂ OCH ₂ CH ₂ OAc	O	5.17
44	H	Me	CH ₂ OCH ₂ CH ₂ OCOPh	O	5.12

Tabelle 35: Chemische Strukturen der HIV1-RT-Inhibitoren (fortgesetzt)



No.	R1	R2	R3	X	pIC ₅₀ (-log ₁₀ [nM])
45	H	Me	CH ₂ OCH ₂ Me	O	6.48
46	H	Me	CH ₂ OCH ₂ CH ₂ Cl	O	5.82
47	H	Me	CH ₂ OCH ₂ CH ₂ N ₃	O	5.24
48	H	Me	CH ₂ OCH ₂ CH ₂ F	O	5.96
49	H	Me	CH ₂ OCH ₂ CH ₂ Me	O	5.48
50	H	Me	CH ₂ OCH ₂ Ph	O	7.06
51	H	Et	CH ₂ OCH ₂ Me	O	7.72
52	H	Et	CH ₂ OCH ₂ Me	S	7.58
53	3,5-Me ₂	Et	CH ₂ OCH ₂ Me	O	8.24
54	3,5-Me ₂	Et	CH ₂ OCH ₂ Me	S	8.3
55	H	Et	CH ₂ OCH ₂ Ph	O	8.23
56	3,5-Me ₂	Et	CH ₂ OCH ₂ Ph	O	8.55
57	H	Et	CH ₂ OCH ₂ Ph	S	8.09
58	3,5-Me ₂	Et	CH ₂ OCH ₂ Ph	S	8.14
59	H	<i>i</i> Pr	CH ₂ OCH ₂ Me	O	7.99
60	H	<i>i</i> Pr	CH ₂ OCH ₂ Ph	O	8.51
61	H	<i>i</i> Pr	CH ₂ OCH ₂ Me	S	7.89
62	H	<i>i</i> Pr	CH ₂ OCH ₂ Ph	S	8.14
63	H	Me	CH ₂ OMe	O	5.68
64	H	Me	CH ₂ OBu	O	5.33
65	H	Me	Et	O	5.66
66	H	Me	Bu	O	5.92
67	3,5-Cl ₂	Et	CH ₂ OCH ₂ Me	S	7.89
68	H	Et	CH ₂ O <i>i</i> Pr	S	6.66
69	H	Et	CH ₂ O <i>c</i> Hex	S	5.79
70	H	Et	CH ₂ OCH ₂ <i>c</i> Hex	S	6.45
71	H	Et	CH ₂ OCH ₂ C ₆ H ₄ (4-Me)	S	7.11
72	H	Et	CH ₂ OCH ₂ C ₆ H ₄ (4-Cl)	S	7.92
73	H	Et	CH ₂ OCH ₂ CH ₂ Ph	S	7.04
74	3,5-Cl ₂	Et	CH ₂ OCH ₂ Me	O	8.13
75	H	Et	CH ₂ O <i>i</i> Pr	O	6.47
76	H	Et	CH ₂ O <i>c</i> Hex	O	5.4
77	H	Et	CH ₂ OCH ₂ <i>c</i> Hex	O	6.35
78	H	Et	CH ₂ OCH ₂ CH ₂ Ph	O	7.02
79	H	<i>c</i> Pr	CH ₂ OCH ₂ Me	S	7.02
80	H	<i>c</i> Pr	CH ₂ OCH ₂ Me	O	7

II.5. Dopamin-Antagonisten (D₁)

Tabelle 36: Übersicht über die Strukturen des D₁-Datensatzes zusammen mit den zugehörigen K_i-Werten für die Wirksamkeit als Antagonisten am D₁-Rezeptor

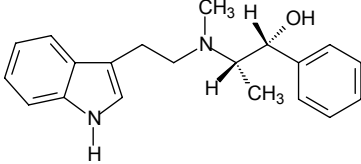
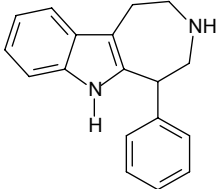
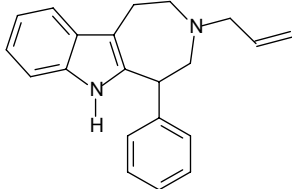
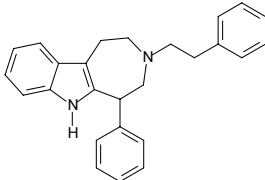
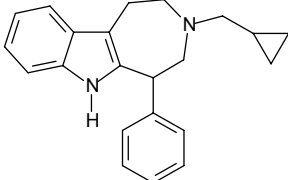
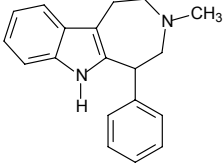
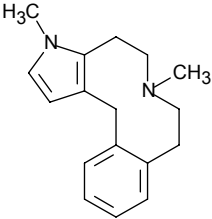
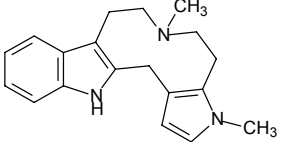
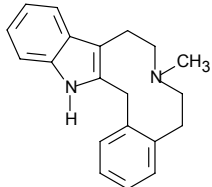
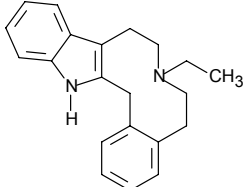
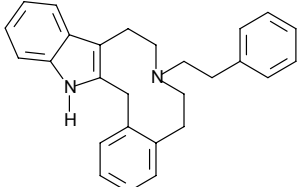
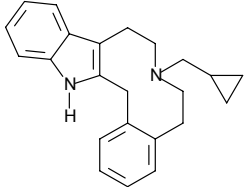
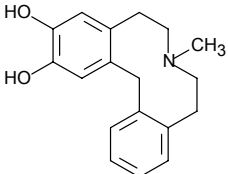
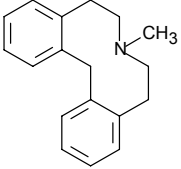
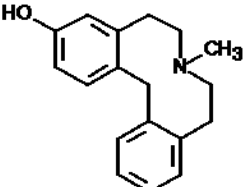
		
DE-17 1752	DE-03 2112	DE-04 2781
		
DE-05 3901	DE-06 2114	DE-07 250
		
SH-3 61	SH-4 101	LE-RU-300 1.9
		
LE-RU-301 16.15	LE-RU-304 655	LE-RU-302 767
		
LE-403 341	LE-410 4.5	LE-404 0.39

Tabelle 37: Übersicht über die Strukturen des D₁-Datensatzes zusammen mit den zugehörigen K_i-Werten für die Wirksamkeit als Antagonisten am D₁-Rezeptor (fortgeführt)

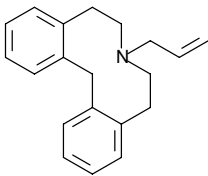
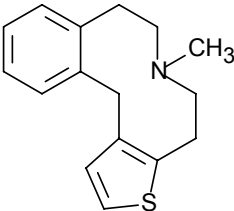
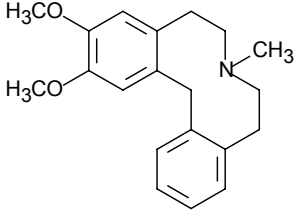
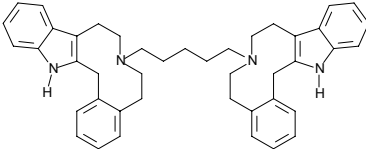
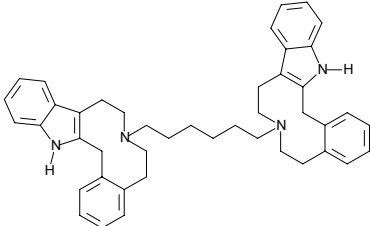
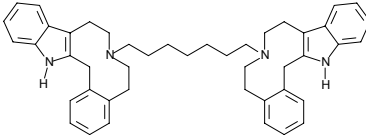
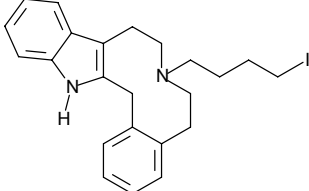
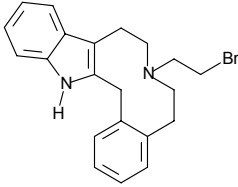
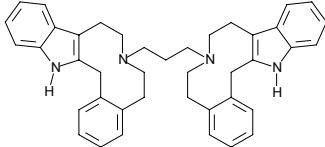
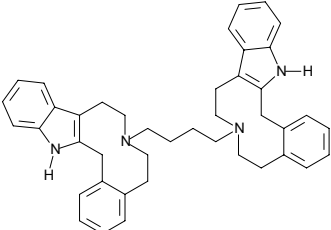
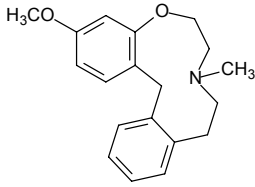
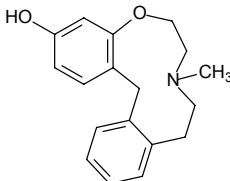
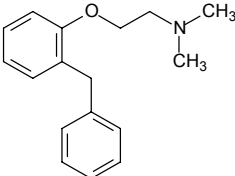
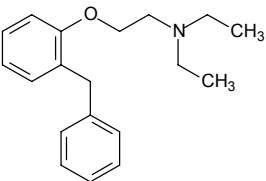
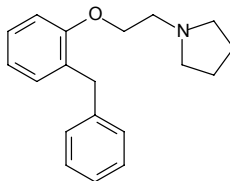
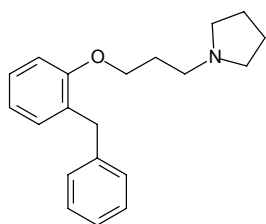
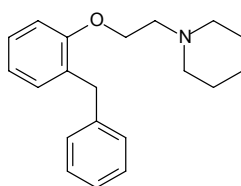
		
LE-411 603	LE-420 10.7	LE-400 509
		
Lan-D5 183.5	Lan-D6 44	Lan-D7 72.4
		
Ab-1, AHA-D9 97.5	Ab2, AHA-D11 260.4	Lan-D3, AHA-D13 1037
		
Lan-D4, AHA-D8 247	LE-WW-450 35.5	LE-WW-451 3.2
		
LE-WW-452 343	LE-WW-453 1959	LE-WW-454 379

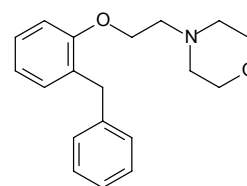
Tabelle 38: Übersicht über die Strukturen des D₁-Datensatzes zusammen mit den zugehörigen K_i-Werten für die Wirksamkeit als Antagonisten am D₁-Rezeptor (fortgeführt)



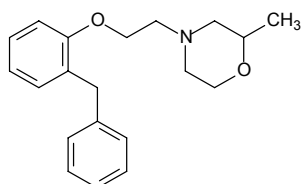
LE-WW-459
113



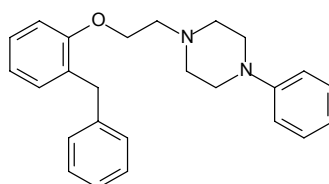
LE-WW-455
33.7



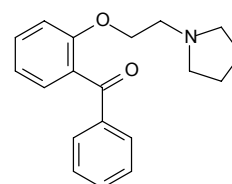
LE-WW-456
201



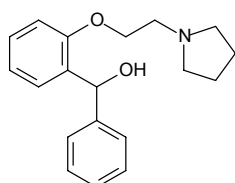
LE-WW-457
1645



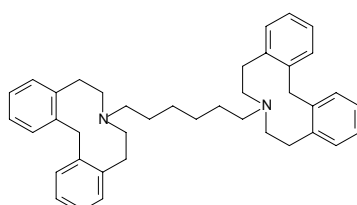
LE-WW-458
1480



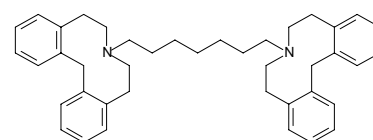
LE-WW-460
1651



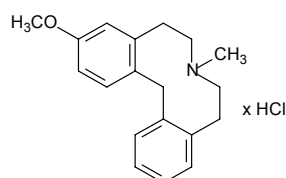
LE-WW-461
2621



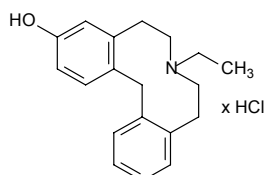
LE-WW-463
5383



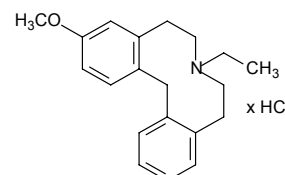
LE-WW-464
5956



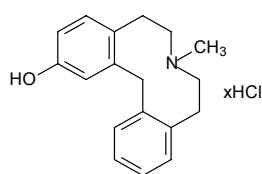
LE-PM-425
23.35



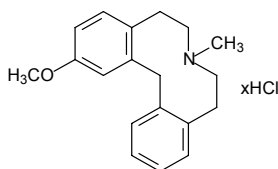
LE-PM-426
3.8



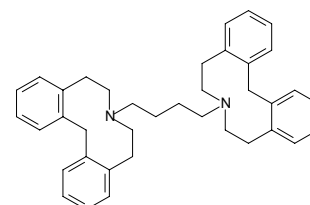
LE-PM-427
33



LE-PM-428
8.92

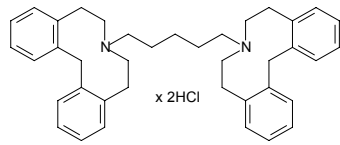


LE-PM-429
82

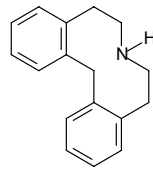


LE-PM-430
52140

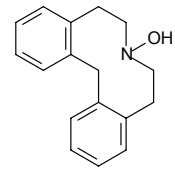
Tabelle 39: Übersicht über die Strukturen des D₁-Datensatzes zusammen mit den zugehörigen K_i-Werten für die Wirksamkeit als Antagonisten am D₁-Rezeptor (fortgeführt)



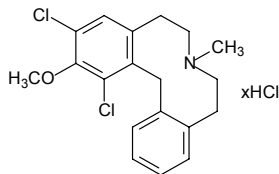
LE-PM-431
4227



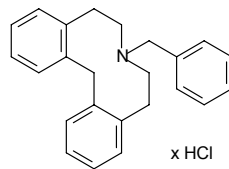
LE-PM-432
2847



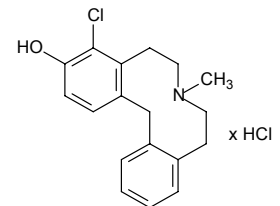
LE-PM-433
4377



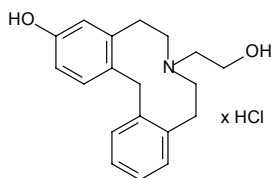
LE-PM-434
25.27



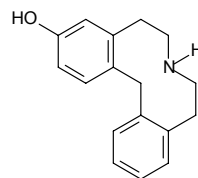
LE-PM-435
543



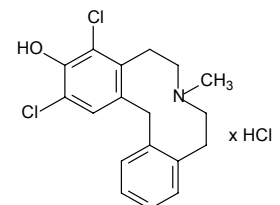
LE-PM-436
0.83



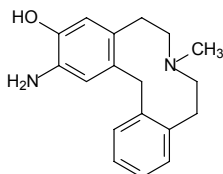
LE-PM-437
41



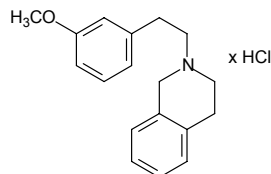
LE-PM-440
68.18



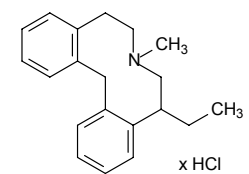
LE-PM-442
3.2



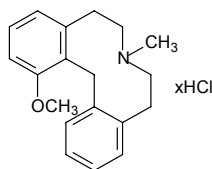
LE-PM-443
9.3



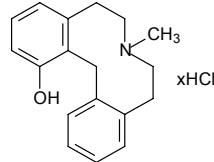
LE-PM-444
364



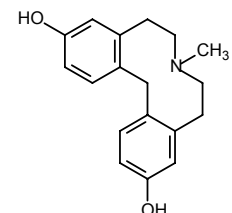
LE-PM-446
18.9



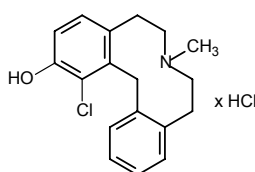
LE-PM-448
7.55



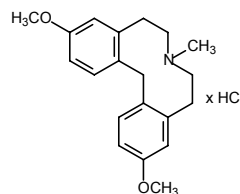
LE-PM-449
8.7



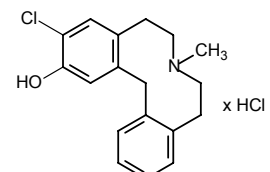
LE-PM-451
2.03



LE-PM-452
0.46



LE-PM-453
9.4



LE-PM-454
3.07

Tabelle 40: Übersicht über die Strukturen des D₁-Datensatzes zusammen mit den zugehörigen K_i-Werten für die Wirksamkeit als Antagonisten am D₁-Rezeptor (fortgeführt)

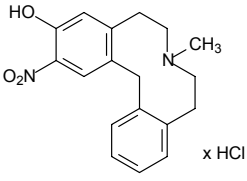
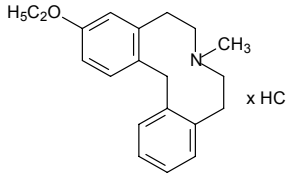
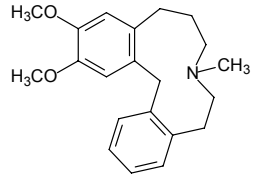
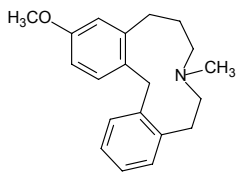
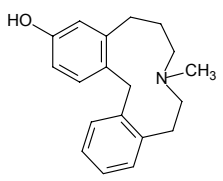
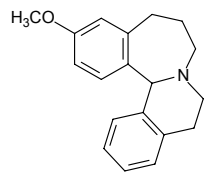
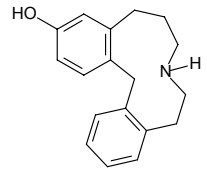
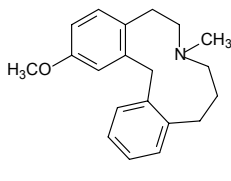
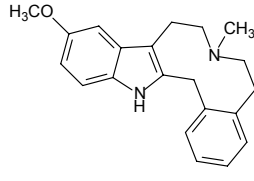
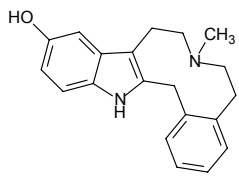
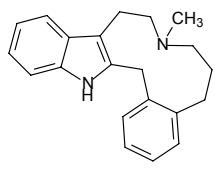
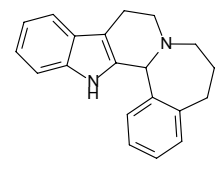
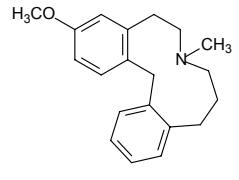
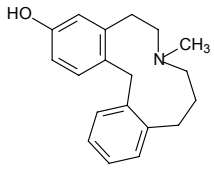
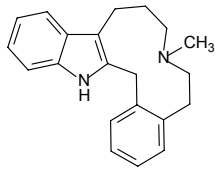
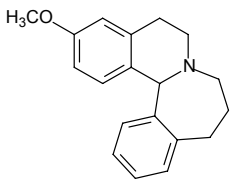
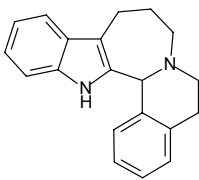
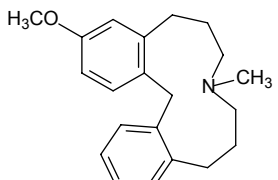
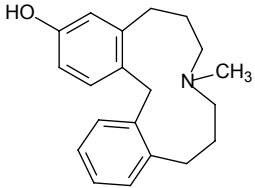
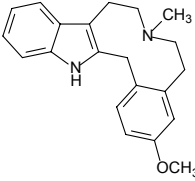
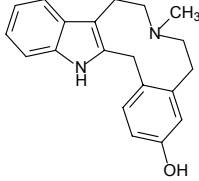
 <p>LE-PM-456 49</p>	 <p>LE-PM-457 32.1</p>	 <p>LE-CE-510 597</p>
 <p>LE-CE-520 29.5</p>	 <p>LE-CE-521 10.9</p>	 <p>LE-CE-522 24750</p>
 <p>LE-CE-523 292</p>	 <p>LE-CE-530 195</p>	 <p>LE-CE-550 0.83</p>
 <p>LE-CE-551 0.56</p>	 <p>LE-CE-560 2.15</p>	 <p>LE-CE-562 61000</p>
 <p>LE-CE-570 18.5</p>	 <p>LE-CE-571 3.23</p>	 <p>LE-CE-580 97</p>
 <p>LE-CE-572 4160</p>	 <p>LE-CE-582 41000</p>	 <p>LE-CE-590 23.5</p>

Tabelle 41: Übersicht über die Strukturen des D₁-Datensatzes zusammen mit den zugehörigen K_i-Werten für die Wirksamkeit als Antagonisten am D₁-Rezeptor (fortgeführt)

		
LE-CE-591 93	LE-CE-500 (End 1) 19	LE-CE-501 (End 2) 3.7

II.6. Glukokortikoide (GK)

Tabelle 42: Übersicht über die Strukturen des GK-Datensatzes zusammen mit den zugehörigen relativen Rezeptoraffinitäten (RRA) relativ zum Dexamethason (Nr. 11)

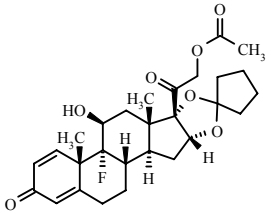
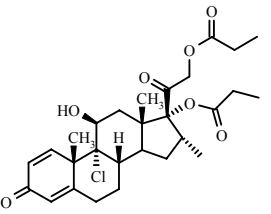
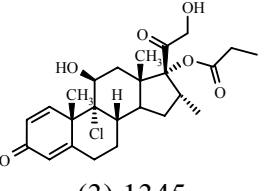
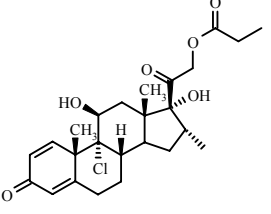
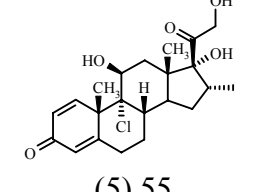
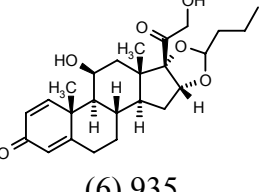
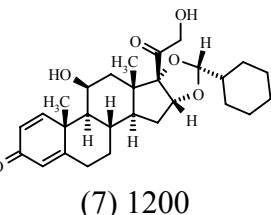
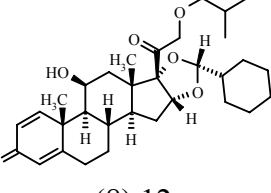
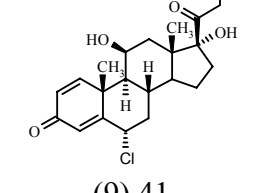
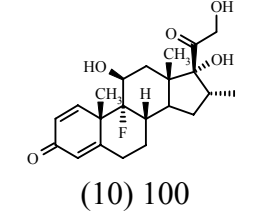
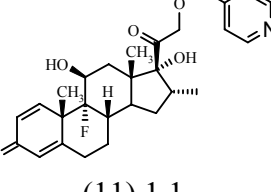
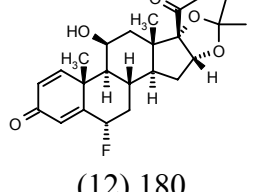
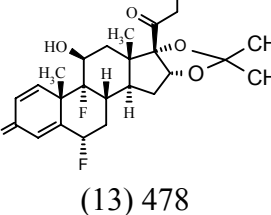
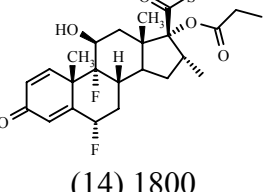
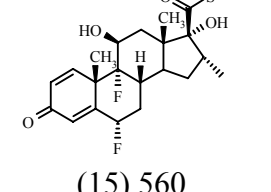
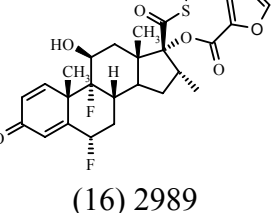
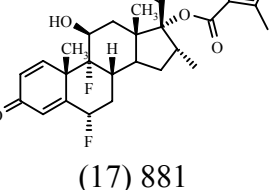
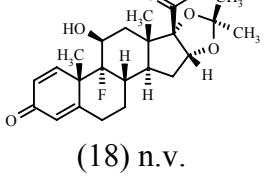
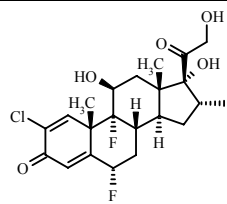
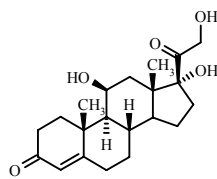
		
(1) 24	(2) 53	(3) 1345
		
(4) 0.9	(5) 55	(6) 935
		
(7) 1200	(8) 12	(9) 41
		
(10) 100	(11) 1.1	(12) 180
		
(13) 478	(14) 1800	(15) 560
		
(16) 2989	(17) 881	(18) n.v.

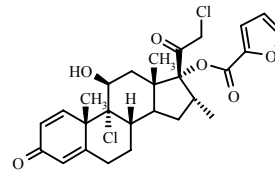
Tabelle 43: Übersicht über die Strukturen des GK-Datensatzes zusammen mit den zugehörigen relativen Rezeptoraffinitäten (RRA) (Fortsetzung)



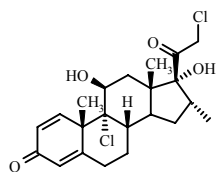
(19) 598



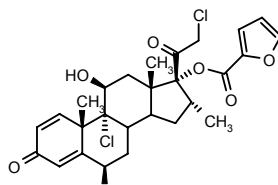
(20) 7.8



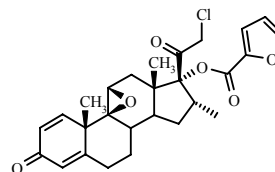
(21) 2244



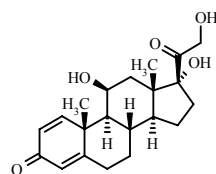
(22) 780



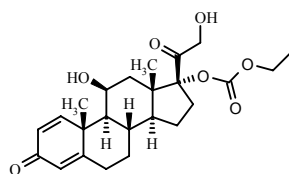
(23) 220



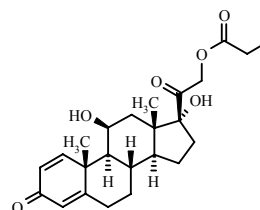
(24) 206



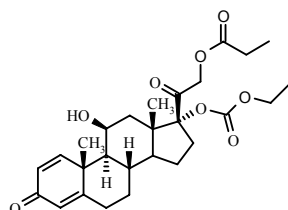
(25) 10.4



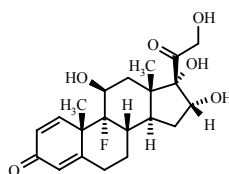
(26) 103



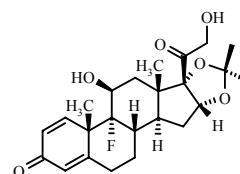
(27) 0.5



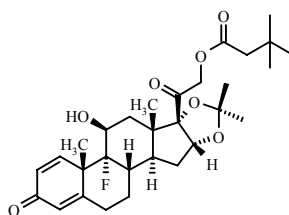
(28) 7.3



(29) 1.1



(30) 361



(31) 3

Tabelle 44: Übersicht über die mit verschiedenen Methoden [268] vorhergesagten logP-Werte für die Verbindungen des GK-Datensatzes

	logP(Broto)	virtual logP	milogP	alogPs	IAlogP	clogP	XLOGP
1	2.179	2.5377	3.175	3.16	1.97	2.88	2.34
2	1.69	2.2837	3.783	3.67	2.39	4.26	2.83
3	0.608	1.259	2.844	2.5	2.85	3.19	1.84
4	0.499	1.2339	2.844	2.65	3.1	3.19	1.84
5	-0.583	0.1302	1.906	2.1	2.45	2.13	1.28
6	1.382	1.9967	2.444	3.1	2.57	3.42	2.33
7	1.34	1.7504	2.948	3.77	0.96	3.88	3.01
8	2.612	3.1037	4.042	4.08	3.03	5.25	4.3
9	-1.162	-0.221	1.406	1.87	1.5	1.41	0.88
10	-0.645	0.1121	1.898	1.95	1.82	1.78	1.14
11	0.15	0.7465	3.358	2.62	2.65	2.62	1.91
12	0.078	0.8713	1.33	2.2	1.74	2.41	1.12
13	0.884	1.4256	2.279	2.49	2.83	1.96	1.56
14	1.988	2.2977	3.713	3.69	4.15	3.8	3.39
15	0.797	1.299	2.775	2.69	2.59	2.63	2.4
16	2.435	2.7879	4.592	3.73	3.67	4.26	3.55
17	2.944	3.1648	4.31	3.71	4.93	4.05	3.57
18	2.289	2.4585	3.311	3.58	3.34	2.85	2.72
19	-0.898	0.1372	2.236	2.4	2.53	2.6	1.61
20	-0.666	-0.0014	1.445	1.71	1.71	1.7	0.52
21	2.363	2.7484	4.92	4.27	3.34	4.12	3.28
22	0.725	1.1633	3.103	2.82	3.44	2.76	2.12
23	1.01	1.7224	3.939	3.59	0.1	2.41	2.3
24	1.754	2.205	3.565	4.07	2.79	3.31	3.01
25	-1.342	-0.3874	1.205	1.64	1.38	1.42	1.02
26	0.051	0.8711	1.447	2.48	2.17	2.43	1.93
27	-0.26	0.6998	2.143	2.59	2.36	2.49	1.59
28	1.133	2.1141	2.386	3.08	2.02	3.5	2.92
29	-2.188	-1.0137	0.762	0.85	0.74	0.71	0.24
30	0.981	1.409	2.114	2.311	2.06	2.21	1.22
31	3.509	3.8574	4.047	3.89	2.97	4.6	3.9

II.7. Naphtylisochinolin-Alkaloide (NIQ)

Tabelle 45: Verbindungen des NIQ-Datensatzes und deren biologische Aktivität.

Nr.	Name	pIC ₅₀	SZ ^a	R1	R2	R3	R4	R5	R6	R7
1a	Dioncopeltin A ^c	1.880	1R/3R	H	H	H	CH ₂ OH	Me	H	H
1b	Ancistrocladisin ^d	-0.236	3S	-	OMe	Me	Me	Me	Me	H
1c	<i>Cis</i> -1,2-Dihydroancistrocladisin	-0.237	1R/3S	H	OMe	Me	Me	Me	Me	H
1d	<i>Trans</i> -1,2-Dihydroancistrocladisin	-0.265	1S/3S	H	OMe	Me	Me	Me	Me	H
1e	Ancistrocongolin D	-0.687	1R/3R	H	OH	Me	Me	Me	Me	H
1f	Ancistrogriffin A	0.729	1S/3S	H	OH	Me	H	Me	Me	Me
1g	Ancistrobertsonin D	-0.639	1R/3S	H	OH	Me	Me	Me	Me	H
1h	<i>N</i> -Benzyl-dioncopeltin A	0.299	1R/3R	Bn	H	H	CH ₂ OH	Me	H	H
1i	<i>N</i> -5'- <i>O</i> -Dibenzyl-dioncopeltin A	-0.001	1R/3R	Bn	H	Bn	CH ₂ OH	Me	H	H
1j	Habropetalin A	1.896	1R/3R	H	H	H	CH ₂ OH	Me	Me	H
1k	Dioncophyllin A	0.419	1R/3R	H	H	H	Me	Me	Me	H
1l	5'- <i>O</i> -Demethyl-dioncophyllin A	0.381	1R/3R	H	H	H	Me	Me	H	H
1m	4- <i>O</i> -5- <i>O</i> -Didemethyl-dioncophyllin A	-0.343	1R/3R	H	H	H	Me	H	H	H
2a	7- <i>epi</i> -Dioncopeltin A	0.538	1R/3R	H	H	H	CH ₂ OH	Me	H	H
2b	Ancistrogriffin C	-0.138	1S/3S	H	OMe	H	H	H	Me	Me
2c	<i>N</i> -Benzyl-7- <i>epi</i> -dioncopeltin A	0.257	1R/3R	Bn	H	H	CH ₂ OH	Me	H	H
2d	7- <i>epi</i> -Dioncophyllin A ^c	0.516	1R/3R	H	H	H	Me	Me	Me	H
2e	<i>N</i> -Methyl-7- <i>epi</i> -dioncophyllin A	-0.172	1R/3R	Me	H	H	Me	Me	Me	H
2f	Dioncophyllin D ^e	0.797	1R/3R	H	H	H	H	H	Me	Me
2g	-- ^c	-0.369	1S/3S	Me	OMe	Me	H	Me	Me	Me

Tabelle 46: Verbindungen des NIQ-Datensatzes (fortgeführt).

Nr.	Name	pIC ₅₀	SZ ^a	R1	R2	R3	R4	R5	R6	R7
3a	Dioncophyllin C ^c	1.782	1R/3R	H	H	H	Me	Me	H	H
3b	Hamatein ^b	-0.261	-	-	H	Me	Me	Me	Me	H
3c	Ancistrocongolol C ^{c,d}	-0.853	1R/3R	Me	OH	Me	H	Me	Me	Me
3d	Ancistrolikokin B	0.226	1S/3R	H	OH	Me	H	Me	H	Me
3e	Ancistrolikokin C	-0.356	1R/3R	Me	OH	Me	H	Me	H	Me
3f	Ancistrobertsonin B	-0.868	1R/3S	Me	OMe	Me	Me	Me	Me	H
3g	<i>N</i> -Methyldioncophyllin C	0.453	1R/3R	Me	H	H	Me	Me	H	H
3h	Korupensamin B	0.936	1R/3R	H	OH	H	H	H	Me	Me
4a	Ancistrocladin	-0.542	1S/3S	H	H	Me	Me	Me	Me	H
4b	<i>N</i> -Methylancistrocladin	-0.329	1S/3S	Me	H	Me	Me	Me	Me	H
4c	Ancistrocongolol A	0.300	1R/3R	Me	H	H	H	H	Me	Me
4d	Ancistrocongolol B	0.423	1R/3R	H	Me	H	H	Me	Me	Me
4e	Ancistroealain A ^{c,d}	-0.401	3S	-	Me	Me	H	Me	Me	Me
4f	Ancistroealain B	-0.102	1S/3S	H	Me	Me	H	H	Me	Me
4g	Ancistrolikokin A	0.464	1R/3R	Me	Me	H	H	H	Me	Me
4h	Ancistrolikokin D ^d	-0.305	3S	-	H	Me	H	Me	H	Me
4i	Ancistrobertsonin A	-1.051	1S/3S	Me	H	Me	H	Me	Me	Me
4j	Ancistrobertsonin C	-0.347	1R/3S	Me	Me	Me	H	Me	Me	Me
4k	Ancistrocladein ^b	-0.277	-	-	H	Me	Me	Me	Me	H
4l	Korupensamin A ^c	0.738	1R/3R	H	H	H	H	H	Me	Me
5a	Dioncophyllinol B ^c	1.074	1R	OH	H	-	-	-	-	-
5b	1- <i>epi</i> -Dioncophyllin B	0.370	1S	H	H	-	-	-	-	-
5c	8- <i>O</i> -Methyl-1- <i>epi</i> -dioncophyllin B	1.431	1S	H	Me	-	-	-	-	-

Tabelle 47: Verbindungen des NIQ-Datensatzes (fortgeführt).

Nr.	Name	pIC ₅₀	SZ ^a	R1	R2	R3	R4	R5	R6	R7
5d	Dioncophyllin B	0.636	1R	H	H	-	-	-	-	-
5e	8- <i>O</i> -Methyldioncophyllinol B	-0.221	1R	OH	Me	-	-	-	-	-
6a	Dioncolacton A ^c	-0.085	H	-	-	-	-	-	-	-
6b	<i>N</i> -Benzylidionco-lacton A	-0.392	Bn	-	-	-	-	-	-	-
7	--	1.454	-	-	-	-	-	-	-	-
8a	--	0.363	Me	Bn	H	H	H	H	OCH ₂ OCH ₃	-
8b	--	0.047	Me	Bn	H	H	H	Br	OCH ₂ OCH ₃	-
8c	--	-0.652	H	Bn	Me	H	OMe	H	OMe	-
8d	--	-0.321	H	H	Me	Br	OMe	H	OMe	-
8e	--	-0.036	H	Bn	Me	Br	OMe	H	OMe	-
9a	--	-0.101	H	H	H	-	-	-	-	-
9b	--	-0.647	Br	Bn	Bn	-	-	-	-	-
9c	--	-0.857	Ph	H	H	-	-	-	-	-
9d	--	-0.617	H	Bn	Bn	-	-	-	-	-

^a SCC: stereogene Zentren. ^b Isochinolinringsystem vollständig ungesättigt. ^c Referenzstruktur der Gruppe. ^d Doppelbindung zwischen C₁ und N₂. ^e Biarylachse ist bei Raumtemperatur frei drehbar. Bn: Benzyl. Ph: Phenyl.

Anhang III. Der Ein- und Dreibuchstabencode für Aminosäuren

Name der Aminosäure	1-Buchstaben-Code	3-Buchstaben-Code
Alanin	A	Ala
Cystein	C	Cys
Aspartat	D	Asp
Glutamat	E	Glu
Phenylalanin	F	Phe
Glycin	G	Gly
Histidin	H	His
Isoleucin	I	Ile
Lysin	K	Lys
Leucin	L	Leu
Methionin	M	Met
Asparagin	N	Asn
Prolin	P	Pro
Glutamin	Q	Gln
Arginin	R	Arg
Serin	S	Ser
Threonin	T	Thr
Valin	V	Val
Tryptophan	W	Trp
Tyrosin	Y	Tyr

Literaturverzeichnis

- [1] E. Fischer; Einfluß der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985-2993.
- [2] H. Berman, K. Henrick, H. Nakamura; Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **2003**, *10*, 980-980.
- [3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne; The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
- [4] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Z. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, L. S. L. Yeh; The universal protein resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, D154-D159.
- [5] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Z. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, L. S. L. Yeh; UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115-D119.
- [6] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Z. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek; The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* **2006**, *34*, D187-D191.
- [7] H. Kopp; Ueber den Zusammenhang zwischen der chemischen Constitution und einigen physikalischen Eigenschaften bei flüssigen Verbindungen. *Annalen der Chemie und Pharmacie* **1844**, *50*, 71-144.
- [8] A. Crum Brown, Fraser TR.; On the Connection between Chemical Constitution and Physiologic Action. Part 1. On the Physiological Action of Salts of the Ammonium Bases, Derived from Strychnia, Brucia, Thebia, Codeia, Morphia and Nicotia. *Trans Roy. Soc. Edinburgh* **1868**, *25*, 151-203.
- [9] R. Todeschini, V. Consonni; *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, Germany **2000**
- [10] K. Baumann; Distance Profiles (DiP): A translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules. *Quant. Struct. Act. Rel.* **2002**, *21*, 507-519.
- [11] C. Cruciani, P. Crivori, P. A. Carrupt, B. Testa; Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *Journal of Molecular Structure-Theochem* **2000**, *503*, 17-30.
- [12] G. Cruciani, M. Pastor, W. Guba; VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29-S39.
- [13] M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi; GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233-3243.
- [14] N. Stiefl, 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.
- [15] A. J. Hopfinger, S. Wang, J. S. Tokarski, B. Q. Jin, M. Albuquerque, P. J. Madhav, C. Duraiswami; Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society* **1997**, *119*, 10509-10524.

- [16] V. E. Kuz'min, A. G. Artemenko, V. P. Lozitsky, E. N. Muratov, A. S. Fedtchouk, N. S. Dyachenko, L. N. Nosach, T. L. Gridina, L. I. Shitikova, L. M. Mudrik, A. K. Mescheriakov, V. A. Chelombitko, A. I. Zheltvay, J. J. Vanden Eynde; The analysis of structure-anticancer and antiviral activity relationships for macrocyclic pyridinophanes and their analogues on the basis of 4D QSAR models (simplex representation of molecular structure). *Acta Biochim. Pol.* **2002**, *49*, 157-168.
- [17] A. Vedani, K. Briem, M. Dobler, H. Dollinger, D. R. McMasters; Multiple-conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. *Journal of Medicinal Chemistry* **2000**, *43*, 4416-4427.
- [18] A. Vedani, D. R. McMasters, M. Dobler; Multi-conformational ligand representation in 4D-QSAR: Reducing the bias associated with ligand alignment. *Quantitative Structure-Activity Relationships* **2000**, *19*, 149-161.
- [19] N. Stiefl; Entwicklung, Validierung und Anwendung einer neuen translations- und rotationsinvarianten 3D-QSAR Methodik. *Dissertation* **2004**, *212 Seiten*, Institut für Pharmazie und Lebensmittelchemie Bayerische Julius-Maximilians-Universität Würzburg.
- [20] C. Hansch, P. P. Maloney, T. Fujita; Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178-&.
- [21] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney; Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3-25.
- [22] A. T. Balaban, I. Motoc, D. Bonchev, O. Mekenyan; Topological Indexes for Structure-Activity Correlations. *Top. Curr. Chem.* **1983**, *114*, 21-55.
- [23] P. J. Hansen, P. C. Jurs; Chemical Applications of Graph-Theory.1. Fundamentals and Topological Indexes. *J. Chem. Educ.* **1988**, *65*, 574-580.
- [24] A. R. Katritzky, E. V. Gordeeva; Traditional Topological Induces Vs Electronic, Geometrical, and Combined Molecular Descriptors in Qsar Qspr Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835-857.
- [25] M. I. Stankevich, I. V. Stankevich, N. S. Zefirov; Topological Indices in Organic-Chemistry. *Usp. Khim.* **1988**, *57*, 337-366.
- [26] D. Bawden; Computerized Chemical Structure-Handling Techniques in Structure Activity Studies and Molecular Property Prediction. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14-22.
- [27] K. C. Chu, R. J. Feldmann, M. B. Shapiro, G. F. Hazard, R. I. Geran; Pattern-Recognition and Structure-Activity Relationship Studies - Computer-Assisted Prediction of Antitumor Activity in Structurally Diverse Drugs in an Experimental Mouse-Brain Tumor System. *J. Med. Chem.* **1975**, *18*, 539-545.
- [28] S. M. Free, J. W. Wilson; Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395-&.
- [29] T. Fujita, T. Ban; Studies on Structure-Activity Relationship.3. Structure-Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters. *J. Med. Chem.* **1971**, *14*, 148-&.
- [30] L. Hodes, G. F. Hazard, R. I. Geran, S. Richman; Statistical-Heuristic Method for Automated Selection of Drugs for Screening. *J. Med. Chem.* **1977**, *20*, 469-475.
- [31] R. E. Carhart, D. H. Smith, R. Venkataraghavan; Atom Pairs as Molecular-Features in Structure Activity Studies - Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
- [32] V. Turau; *Algorithmische Graphentheorie*, Addison-Wesley: Bonn, Germany **1996**

- [33] L. J. Soltzberg, C. L. Wilkins; Molecular Transforms - Potential Tool for Structure-Activity Studies. *J. Am. Chem. Soc.* **1977**, *99*, 439-443.
- [34] R. D. Cramer, D. E. Patterson, 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.
- [35] G. Klebe, U. Abraham, T. Mietzner; Molecular Similarity Indexes in a Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. *J. Med. Chem.* **1994**, *37*, 4130-4146.
- [36] P. J. Goodford; A Computational-Procedure for Determining Energetically Favorable Binding-Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.
- [37] A. M. Doweyko; The Hypothetical Active-Site Lattice - an Approach to Modeling Active-Sites from Data on Inhibitor Molecules. *J. Med. Chem.* **1988**, *31*, 1396-1406.
- [38] K. Baumann; Neue chemometrische Methoden zur Berechnung von quantitativen Struktur-Wirkungs-Beziehungen. *Habilitationsschrift* **2003**, 302 Seiten, Institut für Pharmazie und Lebensmittelchemie Bayerische Julius-Maximilians-Universität Würzburg.
- [39] J. A. Calder, J. A. Wyatt, D. A. Frenkel, J. E. Casida; Comfa Validation of the Superposition of 6 Classes of Compounds Which Block Gaba Receptors Noncompetitively. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 45-60.
- [40] Y. Kato, A. Itai, Y. Iitaka; A Novel Method for Superimposing Molecules and Receptor Mapping. *Tetrahedron* **1987**, *43*, 5229-5236.
- [41] G. Klebe; *Structural Alignment of Molecules*, 173-199 in Book; H. Kubinyi: Structural Alignment of Molecules, ESCOM: Leiden, Netherlands **1993**
- [42] C. Lemmen, T. Lengauer; Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215-232.
- [43] Y. C. Martin, M. G. Bures, E. A. Danaher, J. Delazzer, I. Lico, P. A. Pavlik; A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83-102.
- [44] S. Namasivayam, P. M. Dean; Statistical-Method for Surface Pattern-Matching between Dissimilar Molecules - Electrostatic Potentials and Accessible Surfaces. *Journal of Molecular Graphics* **1986**, *4*, 46-&.
- [45] G. Moreau, P. Broto; The Auto-Correlation of a Topological-Structure - a New Molecular Descriptor. *Nouveau Journal De Chimie-New Journal of Chemistry* **1980**, *4*, 359-360.
- [46] P. Broto, G. Moreau, C. Vandycke; Molecular-Structures - Perception, Auto-Correlation Descriptor and Sar Studies - System of Atomic Contributions for the Calculation of the Normal-Octanol Water Partition-Coefficients. *Eur. J. Med. Chem.* **1984**, *19*, 71-78.
- [47] R. Todeschini, M. Lasagni; New Molecular Descriptors for 2D and 3D Structures - Theory. *J. Chemom.* **1994**, *8*, 263-272.
- [48] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani; MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79-92.
- [49] E. Gancia, G. Bravi, P. Mascagni, A. Zaliani; Global 3D-QSAR methods: MS-WHIM and autocorrelation. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 293-306.
- [50] R. Todeschini, G. Moro, R. Boggia, L. Bonati, U. Cosentino, M. Lasagni, D. Pitea; Modeling and prediction of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 65-73.

- [51] A. M. Ferguson, T. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan, P. J. Snaith; EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143-152.
- [52] R. Bursi, T. Dao, T. van Wijk, M. de Gooyer, E. Kellenbach, P. Verwer; Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861-867.
- [53] K. Baumann; An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 26-35.
- [54] K. Baumann, J. T. Clerc; Computer-assisted IR spectra prediction - Linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348*, 327-343.
- [55] J. T. Clerc, A. L. Terkovic; Versatile Topological-Structure Descriptor for Quantitative Structure Property Studies. *Anal. Chim. Acta* **1990**, *235*, 93-102.
- [56] K. Baumann; Uniform-length molecular descriptors for quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR): classification studies and similarity searching. *Trac-Trends in Analytical Chemistry* **1999**, *18*, 36-46.
- [57] A. Bender, H. Y. Mussa, G. S. Gill, R. C. Glen; Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569-6583.
- [58] A. Bender, H. Y. Mussa, R. C. Glen, S. Reiling; Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708-1718.
- [59] A. Bender, S. Reiling, H. Y. Mussa, R. C. Glen; Similarity searching using atom environments, information gain based feature selection and the naive Bayesian classifier. *Abstracts of Papers of the American Chemical Society* **2004**, *227*, U907-U907.
- [60] M. F. Sanner, A. J. Olson, J. C. Spohner; Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305-320.
- [61] P. J. Goodford; *The Basic Principles of GRID*, 3-26 in Book; C. Cruciani: *The Basic Principles of GRID*, Wiley-VCH: Weinheim, Germany **2006**
- [62] A. Bender, R. C. Glen; Discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369-1375.
- [63] A. Bender, A. Klamt, K. Wichmann, M. Thormann, R. C. Glen; Molecular similarity searching using COSMO screening charges (COSMO/3PP). *Computational Life Sciences, Proceedings* **2005**, *3695*, 175-185.
- [64] A. Bender, H. Y. Mussa, R. C. Glen; Screening for dihydrofolate reductase inhibitors using MOLPRINT 2D, a fast fragment-based method employing the naive Bayesian classifier: Limitations of the descriptor and the importance of balanced chemistry in training and test sets. *Journal of Biomolecular Screening* **2005**, *10*, 658-666.
- [65] M. Karthikeyan, R. C. Glen, A. Bender; General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* **2005**, *45*, 581-590.
- [66] F. Fontaine; Development and Applications of new 3D molecular descriptors. *Dissertation* **2004**, *92 Seiten*, Computer-assisted drug design laboratory, Research Group on Biomedical Informatics Pompeu Fabra University Barcelona.
- [67] F. Fontaine, M. Pastor, F. Sanz; Incorporating molecular shape into the alignment-free GRid-INdependent Descriptors. *J. Med. Chem.* **2004**, *47*, 2805-2815.

- [68] R. Mannhold, G. Berellini, E. Carosati, P. Benedetti; *Use of MIF-based VolSurf Descriptors in Physicochemical and Pharmacokinetic Studies*, 173-196 in Book; C. Cruciani: *Use of MIF-based VolSurf Descriptors in Physicochemical and Pharmacokinetic Studies*, Wiley-VCH: Weinheim, Germany **2006**
- [69] N. Jain, S. H. Yalkowsky; Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234-252.
- [70] F. Lombardo, R. S. Obach, M. Y. Shalaeva, F. Gao; Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J. Med. Chem.* **2004**, *47*, 1242-1250.
- [71] S. H. Yalkowsky, S. C. Valvani; Solubility and Partitioning. 1. Solubility of Non-Electrolytes in Water. *J. Pharm. Sci.* **1980**, *69*, 912-922.
- [72] M. Pastor; *Alignment-independent Descriptors from Molecular Interaction Fields*, 117-141 in Book; G. Cruciani: *Alignment-independent Descriptors from Molecular Interaction Fields*, Wiley-VCH: Weinheim, Germany **2006**
- [73] L. Afzelius, I. Zamora, C. M. Masimirembwa, A. Karlen, T. B. Andersson, S. Mecucci, M. Baroni, G. Cruciani; Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors. *J. Med. Chem.* **2004**, *47*, 907-914.
- [74] F. Fontaine, M. Pastor, I. Zamora, F. Sanz; Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687-2694.
- [75] G. Berellini, G. Cruciani, R. Mannhold; Pharmacophore, drug metabolism, and pharmacokinetics models on non-peptide AT(1), AT(2), and AT(1)/AT(2) angiotensin II receptor antagonists. *J. Med. Chem.* **2005**, *48*, 4389-4399.
- [76] P. Braiuca, L. Boscarol, C. Ebert, P. Linda, L. Gardossi; 3D-QSAR applied to the quantitative prediction of penicillin G amidase selectivity. *Adv. Synth. Catal.* **2006**, *348*, 773-780.
- [77] F. Broccolo, G. Cainelli, G. Caltabiano, C. E. A. Cocuzza, C. G. Fortuna, P. Galletti, D. Giacomini, G. Musumarra, R. Musumeci and A. Quintavalla; Design, synthesis, and biological evaluation of 4-alkyliden-beta lactams: New products with promising antibiotic activity against resistant bacteria. *J. Med. Chem.* **2006**, *49*, 2804-2811.
- [78] R. Budriesi, E. Carosati, A. Chiarini, B. Cosimelli, G. Cruciani, P. Ioan, D. Spinelli, R. Spisani; A new class of selective myocardial calcium channel modulators. 2. Role of the acetal chain in oxadiazol-3-one derivatives. *J. Med. Chem.* **2005**, *48*, 2445-2456.
- [79] E. Carosati, H. Lemoine, R. Spogli, D. Grittner, R. Mannhold, O. Tabarrini, S. Sabatini, V. Cecchetti; Binding studies and GRIND/ALMOND-based 3D QSAR analysis of benzothiazine type K-ATP-channel openers. *Bioorg. Med. Chem.* **2005**, *13*, 5581-5591.
- [80] N. Stiefl, K. Baumann; Structure-based validation of the 3D-QSAR technique MaP. *J. Chem. Inf. Model.* **2005**, *45*, 739-749.
- [81] N. Stiefl, G. Bringmann, C. Rummey, 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.
- [82] D. P. Zlotos, S. Buller, N. Stiefl, K. Baumann, Y. Mohr; Probing the pharmacophore for allosteric ligands of muscarinic M-2 receptors: SAR and QSAR studies in a series of bisquaternary salts of caracurine V and related ring systems. *J. Med. Chem.* **2004**, *47*, 3561-3571.
- [83] J. L. Pascualahir, E. Silla; Gepol - an Improved Description of Molecular-Surfaces. 1. Building the Spherical Surface Set. *J. Comput. Chem.* **1990**, *11*, 1047-1060.

- [84] G. Schneider, W. Neidhart, T. Giller, G. Schmid; "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie-International Edition* **1999**, *38*, 2894-2896.
- [85] S. Renner, U. Fechner, G. Schneider; *Alignment-free Pharmacophore Patterns - A Correlation-vector Approach*, 49-79 in Book; T. Langer, R. Hoffmann: *Alignment-free Pharmacophore Patterns - A Correlation-vector Approach*, Wiley-VCH: Weinheim, Germany **2006**
- [86] S. Renner, G. Schneider; Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006**, *2006*, 181-185.
- [87] B. L. Bush, R. P. Sheridan; Patty - a Programmable Atom Typer and Language for Automatic Classification of Atoms in Molecular Databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756-762.
- [88] J. S. Duca, A. J. Hopfinger; Estimation of molecular similarity based on 4D-QSAR analysis: Formalism and validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367-1387.
- [89] X. Hong, A. J. Hopfinger; 3D-pharmacophores of flavonoid binding at the benzodiazepine GABA(A) receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324-336.
- [90] X. Hong, M. D. Krasowski, A. J. Hopfinger, N. L. Harrison; 4D-QSAR analysis of a set of propofol analogs: Mapping binding sites for an anesthetic phenol on the GABA(A) receptor. *Abstracts of Papers of the American Chemical Society* **2002**, *223*, U493-U493.
- [91] A. J. Hopfinger, A. Reaka, P. Venkatarangan, J. S. Duca, S. Wang; Construction of a virtual nigh throughput screen by 4D-QSAR analysis: Application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151-1160.
- [92] A. J. Hopfinger, P. Vankatarangan, Y. F. Tseng, S. Wang, J. S. Duca; Evaluation of alignment dependence in 3D-QSAR model construction using 4D-QSAR analysis. *Internet Journal of Chemistry* **2000**, *3*, art. no.-10.
- [93] C. D. P. Klein, A. J. Hopfinger; Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. *Pharm. Res.* **1998**, *15*, 303-311.
- [94] M. D. Krasowski, X. A. Hong, A. J. Hopfinger, N. L. Harrison; 4D-QSAR analysis of a set of propofol analogues: Mapping binding sites for an anesthetic phenol on the GABA(A) receptor. *J. Med. Chem.* **2002**, *45*, 3210-3221.
- [95] J. Z. Liu, D. H. Pan, Y. F. Tseng, A. J. Hopfinger; 4D-QSAR analysis of a series of antifungal P450 inhibitors and 3D-pharmacophore comparisons as a function of alignment. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2170-2179.
- [96] D. H. Pan, Y. F. Tseng, A. J. Hopfinger; Quantitative structure-based design: Formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1591-1607.
- [97] M. Ravi, A. J. Hopfinger, R. E. Hormann, L. Dinan; 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1587-1604.
- [98] N. C. Romeiro, M. G. Albuquerque, R. B. de Alencastro, M. Ravi, A. J. Hopfinger; Construction of 4D-QSAR models for use in the design of novel p38-MAPK inhibitors. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 385-400.
- [99] O. A. Santos, A. J. Hopfinger; A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1-12.
- [100] O. A. Santos, A. J. Hopfinger; The 4D-QSAR paradigm: Application to a novel set of nonpeptidic HIV protease inhibitors. *Quant. Struct. Act. Rel.* **2002**, *21*, 369-381.

- [101] O. A. Santos, A. J. Hopfinger; Structure-based QSAR analysis of a set of 4-hydroxy-5,6-dihydropyrones as inhibitors of HIV-1 protease: An application of the receptor-dependent (RD) 4D-QSAR formalism. *J. Chem. Inf. Model.* **2006**, *46*, 345-354.
- [102] C. L. Senese, A. J. Hopfinger; A simple clustering technique to improve QSAR model selection and predictivity: Application to a receptor independent 4D-QSAR analysis of cyclic urea derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2180-2193.
- [103] C. L. Senese, A. J. Hopfinger; Receptor-independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1297-1307.
- [104] P. Venkatarangan, A. J. Hopfinger; Prediction of ligand-receptor binding free energy by 4D-QSAR analysis: Application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141-1150.
- [105] E. F. F. da Cunha, M. G. Albuquerque, O. A. C. Antunes, R. B. de Alencastro; 4D-QSAR models of HOE/BAY-793 analogues as HIV-1 protease inhibitors. *QSAR Comb. Sci.* **2005**, *24*, 240-253.
- [106] S. J. Weiner, P. A. Kollman, D. T. Nguyen, D. A. Case; An All Atom Force-Field for Simulations of Proteins and Nucleic-Acids. *J. Comput. Chem.* **1986**, *7*, 230-252.
- [107] Seascape-Learning.com; 4D-QSAR Users Manual. **2005**,
- [108] M. Hahn; Receptor Surface Models.1. Definition and Construction. *J. Med. Chem.* **1995**, *38*, 2080-2090.
- [109] M. Hahn, D. Rogers; Receptor Surface Models.2. Application to Quantitative Structure-Activity-Relationships Studies. *J. Med. Chem.* **1995**, *38*, 2091-2102.
- [110] A. Vedani, M. Dobler, P. Zbinden; Quasi-atomistic receptor surface models: A bridge between 3-D QSAR and receptor modeling. *J. Am. Chem. Soc.* **1998**, *120*, 4471-4477.
- [111] M. Greener; QSAR: Prediction beyond the fourth dimension. *Drug Discovery & Development* **2005**, *5*,
- [112] M. A. Lill, A. Vedani; Exploring QSAR beyond 3D: From optimization of binding affinities to prediction of adverse reactions. *Erlangen* **2005**,
- [113] M. A. Lill, F. Winiger, A. Vedani, B. Ernst; Impact of induced fit on ligand binding to the androgen receptor: A multidimensional QSAR study to predict endocrine-disrupting effects of environmental chemicals. *J. Med. Chem.* **2005**, *48*, 5666-5674.
- [114] A. Vedani; 4D-QSAR and beyond. *Quant. Struct. Act. Rel.* **2002**, *21*, 347-347.
- [115] A. Vedani, M. Dobler; Multidimensional QSAR: Moving from three- to five-dimensional concepts. *Quant. Struct. Act. Rel.* **2002**, *21*, 382-390.
- [116] A. Vedani, M. Dobler; 5D-QSAR: The key for simulating induced fit? *Journal of Medicinal Chemistry* **2002**, *45*, 2139-2149.
- [117] A. Vedani, M. Dobler, H. Dollinger, K. M. Hasselbach, F. Birke, M. A. Lill; Novel ligands for the chemokine receptor-3 (CCR3): A receptor-modeling study based on 5D-QSAR. *J. Med. Chem.* **2005**, *48*, 1515-1527.
- [118] A. Vedani, M. Dobler, M. A. Lill; Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **2005**, *48*, 3700-3703.
- [119] V. E. Kuz'min, A. G. Artemenko, R. N. Lozytska, A. S. Fedtchouk, V. P. Lozitsky, E. N. Muratov, A. K. Mescheriakov; Investigation of anticancer activity of macrocyclic Schiff bases by means of 4D-QSAR based on simplex representation of molecular structure. *SAR QSAR Environ. Res.* **2005**, *16*, 219-230.

- [120] V. A. Potemkin, R. M. Arslambekov, E. V. Bartashevich, M. A. Grishina, A. V. Belik, S. Perspicace, S. Guccione; Multiconformational method for analyzing the biological activity of molecular structures. *Journal of Structural Chemistry* **2002**, *43*, 1045-1049.
- [121] E. V. Bartashevich, V. A. Potemkin, M. A. Grishina, A. V. Belik; A method for multiconformational modeling of the three-dimensional shape of a molecule. *Journal of Structural Chemistry* **2002**, *43*, 1033-1039.
- [122] J. B. Bhonsle, Z. X. Wang, H. Tamamura, N. Fujii, S. C. Peiper, J. O. Trent; A simple, automated quasi-4D-QSAR, quasi-multi way PLS approach to develop highly predictive QSAR models for highly flexible CXCR4 inhibitor cyclic pentapeptide ligands using scripted common molecular modeling tools. *QSAR Comb. Sci.* **2005**, *24*, 620-630.
- [123] A. Davison, D. Hinkley; *Bootstrap Methods and their Application*, Cambridge University Press: Cambridge, UK **1997**
- [124] A. McQuarrie, C. Tsai; *Regression and Time Series Selection*, World Scientific Publishing Co. Pte. Ltd.: Singapore **1998**
- [125] A. Miller; *Subset selection in Regression*, Chapman&Hall: London, UK **1990**
- [126] J. Shao, D. Tu; *The Jackknife and the Bootstrap*, Springer: New York, USA **1995**
- [127] R. Kramer; *Chemometric techniques for quantitative analysis*, Marcel Dekker AG: Basel, Switzerland **1998**
- [128] J. Mandel; Use of the Singular Value Decomposition in Regression-Analysis. *American Statistician* **1982**, *36*, 15-24.
- [129] I. Jolliffe; *Principal Components Analysis*, Springer: New York, USA **1986**
- [130] S. Wold, H. Martens, H. Wold; 286-293 **in** Book; A. Ruhe, B. Kagström: Springer: Heidelberg, Germany **1983**
- [131] K. Baumann; Cross-validation as the objective function for variable-selection techniques. *Trac-Trends in Analytical Chemistry* **2003**, *22*, 395-406.
- [132] S. Wold; Cross-Validatory Estimation of Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397-405.
- [133] S. Geisser; Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association* **1975**, *70*, 320-328.
- [134] J. Shao; Linear-Model Selection by Cross-Validation. *Journal of the American Statistical Association* **1993**, *88*, 486-494.
- [135] A. Golbraikh, A. Tropsha; Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20*, 269-276.
- [136] T. G. Dietterich; Ensemble Methods in Machine Learning. *Technical Report, Oregon State University* **2000**,
- [137] T. G. Dietterich; *Ensemble Methods in Machine Learning*, 1-15 **in** Book; F. R. J. Kittler: Ensemble Methods in Machine Learning, Springer: **2001**
- [138] M. Busemann; Entwicklung chemometrischer Methoden für das in-silico-Wirkstoffdesign. **2006**, *173 Seiten*, Institut für Pharmazie und Lebensmittelchemie *Bayerische Julius-Maximilians-Universität Würzburg*.
- [139] L. Breiman; Bagging predictors. *Machine Learning* **1996**, *24*, 123-140.
- [140] P. Buhlmann, B. Yu; Analyzing bagging. *Annals of Statistics* **2002**, *30*, 927-961.

- [141] J. H. Friedman, B. E. Pospescu; Importance Sampled Learning Ensembles. *Technical Report, Stanford University* **2003**, 32.
- [142] C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, G. L. Cote; Theoretical justification of wavelength selection in PLS calibration development of a new algorithm. *Anal. Chem.* **1998**, *70*, 35-44.
- [143] J. G. Topliss, R. J. Costello; Chance Correlations in Structure-Activity Studies Using Multiple Regression-Analysis. *J. Med. Chem.* **1972**, *15*, 1066-&.
- [144] H. Sies; A New Parameter for Sex-Education. *Nature* **1988**, *332*, 495-495.
- [145] B. Naturschutz; VI. Internationaler Weißstorchzensus. **2005**,
- [146] K. Baumann, H. Albert, 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.
- [147] K. Baumann, M. von Korff, H. Albert; A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. Practical applications. *J. Chemom.* **2002**, *16*, 351-360.
- [148] K. Baumann; Chance correlation in variable subset regression: Influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR Comb. Sci.* **2005**, *24*, 1033-1046.
- [149] J. G. Topliss, R. P. Edwards; Chance Factors in Studies of Quantitative Structure-Activity-Relationships. *J. Med. Chem.* **1979**, *22*, 1238-1244.
- [150] Accelrys; Catalyst. **2005**,
- [151] Eyesopen; Omega. **2005**,
- [152] J. Kirchmair, G. Wolber, C. Laggner, T. Langer; Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**,
- [153] A. Smellie, S. D. Kahn, S. L. Teig; Analysis of Conformational Coverage.2. Application of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295-304.
- [154] A. Smellie, S. D. Kahn, S. L. Teig; Analysis of Conformational Coverage.1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285-294.
- [155] A. Smellie, S. L. Teig, P. Towbin; Poling - Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16*, 171-187.
- [156] J. Kirchmair, C. Laggner, G. Wolber, T. Langer; Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422-430.
- [157] A. K. Ghose, G. M. Crippen; Atomic Physicochemical Parameters for 3-Dimensional Structure-Directed Quantitative Structure-Activity-Relationships.1. Partition-Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565-577.
- [158] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski; Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762-3772.
- [159] K. Baumann, N. Stiefl; Validation tools for variable subset regression. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 549-562.
- [160] R. Koradi, M. Billeter, K. Wuthrich; MOLMOL: A program for display and analysis of macromolecular structures. *Journal of Molecular Graphics* **1996**, *14*, 51-&.

- [161] W. Kabsch; Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors. *Acta Crystallographica Section A* **1978**, *34*, 827-828.
- [162] U. Pieper, N. Eswar, H. Braberg, M. S. Madhusudhan, F. P. Davis, A. C. Stuart, N. Mirkovic, A. Rossi, M. A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C. C. Huang, T. E. Ferrin, A. Sali; MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **2004**, *32*, D217-D222.
- [163] U. Pieper, N. Eswar, F. P. Davis, H. Braberg, M. S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B. M. Webb, D. Eramian, M. Y. Shen, L. Kelly, F. Melo, A. Sali; MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **2006**, *34*, D291-D295.
- [164] U. Pieper, N. Eswar, A. C. Stuart, V. A. Ilyin, A. Sali; MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **2002**, *30*, 255-259.
- [165] J. C. Cole, C. W. Murray, J. W. M. Nissink, R. D. Taylor, R. Taylor; Comparing protein-ligand docking programs is difficult. *Proteins-Structure Function and Bioinformatics* **2005**, *60*, 325-332.
- [166] M. D. Cummings, R. L. DesJarlais, A. C. Gibbs, V. Mohan, E. P. Jaeger; Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48*, 962-976.
- [167] I. Halperin, B. Y. Ma, H. Wolfson, R. Nussinov; Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins-Structure Function and Genetics* **2002**, *47*, 409-443.
- [168] T. Lengauer, M. Rarey; Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* **1996**, *6*, 402-406.
- [169] V. Mohan, A. C. Gibbs, M. D. Cummings, E. P. Jaeger, R. L. DesJarlais; Docking: Successes and challenges. *Curr. Pharm. Des.* **2005**, *11*, 323-333.
- [170] R. D. Taylor, P. J. Jewsbury, J. W. Essex; A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151-166.
- [171] W. Sippl, J. M. Contreras, I. Parrot, Y. M. Rival, C. G. Wermuth; Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 395-410.
- [172] F. Frolow, M. Harel, J. L. Sussman, M. Mevarech, M. Shoham; Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin (vol 3, pg 452, 1996). *Nature Structural Biology* **1996**, *3*, 1055-1055.
- [173] M. Harel, J. L. Hyatt, B. Brumshtein, C. L. Morton, R. M. Wadkins, I. Silman, J. L. Sussman, P. M. Potter; The 3D structure of the anticancer prodrug CPT-11 with Torpedo californica acetylcholinesterase rationalizes its inhibitory action on AChE and its hydrolysis by butyrylcholinesterase and carboxylesterase. *Chem.-Biol. Interact.* **2005**, *157*, 153-157.
- [174] M. Harel, J. L. Hyatt, B. Brumshtein, C. L. Morton, K. J. P. Yoon, R. M. Wadkins, I. Silman, J. L. Sussman, P. M. Potter; The crystal structure of the complex of the anticancer prodrug 7-ethyl-10-[4-(1-piperidino)-1-piperidino]carbonyloxycamptothecin (CPT-11) with Torpedo californica acetylcholinesterase provides a molecular explanation for its cholinergic action. *Mol. Pharmacol.* **2005**, *67*, 1874-1881.
- [175] J. L. Hyatt, L. Tsurkan, C. L. Morton, K. J. P. Yoon, M. Harel, B. Brumshtein, I. Silman, J. L. Sussman, R. M. Wadkins, P. M. Potter; Inhibition of acetylcholinesterase by the anticancer prodrug CPT-11. *Chem.-Biol. Interact.* **2005**, *157*, 247-252.
- [176] C. B. Millard, G. Kryger, A. Ordentlich, H. M. Greenblatt, M. Harel, M. L. Raves, Y. Segall, D. Barak, A. Shafferman, I. Silman, J. L. Sussman; Crystal structures of aged phosphonylated acetylcholinesterase: Nerve agent reaction products at the atomic level. *Biochemistry* **1999**, *38*, 7032-7039.

- [177] I. Silman, M. Harel, P. Axelsen, M. Raves, J. L. Sussman; 3-Dimensional Structures of Acetylcholinesterase and of Its Complexes with Anticholinesterase Agents. *Biochem. Soc. Trans.* **1994**, *22*, 745-749.
- [178] J. L. Sussman, M. Harel, I. Silman; 3-Dimensional Structure of Acetylcholinesterase and of Its Complexes with Anticholinesterase Drugs. *Chem.-Biol. Interact.* **1993**, *87*, 187-197.
- [179] C. G. Wermuth, A. Exinger; 3-(2-Morpholino-Ethylamino)-4-Methyl-6-Phenyl Pyridazine Dihydrochloride (Agr 1240). *Agressologie* **1972**, *13*, 285-&.
- [180] H. M. Bryson, P. Benfield; Donepezil. *Drugs & Aging* **1997**, *10*, 234-239.
- [181] G. L. Ellman, K. D. Courtney, V. Andres, R. M. Featherstone; A New and Rapid Colorimetric Determination of Acetylcholinesterase Activity. *Biochem. Pharmacol.* **1961**, *7*, 88-&.
- [182] G. Kryger, I. Silman, J. L. Sussman; Three-dimensional structure of a complex of E2020 with acetylcholinesterase from *Torpedo californica*. *Journal of Physiology-Paris* **1998**, *92*, 191-194.
- [183] G. Kryger, I. Silman, J. L. Sussman; Structure of acetylcholinesterase complexed with E2020 (Aricept (R)): implications for the design of new anti-Alzheimer drugs. *Structure* **1999**, *7*, 297-307.
- [184] B. Kramer, G. Metz, M. Rarey, T. Lengauer; Ligand docking and screening with FlexX. *Med. Chem. Res.* **1999**, *9*, 463-478.
- [185] B. Kramer, M. Rarey, T. Lengauer; CASP2 experiences with docking flexible ligands using FLEXX. *Proteins-Structure Function and Genetics* **1997**, 221-225.
- [186] B. Kramer, M. Rarey, T. Lengauer; Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins-Structure Function and Genetics* **1999**, *37*, 228-241.
- [187] D. G. Fletcher, K. H. Gibson, H. R. Moss, D. R. Sheldon, E. R. H. Walker; Synthesis and Biological-Activity of 16,17-Configurationaly-Rigid-17-Aryl 18,19,20-Trinor-Prostaglandins. *Prostaglandins* **1976**, *12*, 493-500.
- [188] M. Hayashi, Y. Arai, H. Wakatsuka, M. Kawamura, Y. Konishi, T. Tsuda, K. Matsumoto; Prostaglandin Analogs Possessing Antinidatory Effects.2. Modification of the Alpha-Chain. *J. Med. Chem.* **1980**, *23*, 525-535.
- [189] M. Hayashi, H. Miyake, S. Kori, T. Tanouchi, H. Wakatsuka, Y. Arai, T. Yamato, I. Kajiwara, Y. Konishi, T. Tsuda, K. Matsumoto; Prostaglandin Analogs Possessing Antinidatory Effects.1. Modification of the Omega-Chain. *J. Med. Chem.* **1980**, *23*, 519-524.
- [190] T. A. Martinek, F. Otvos, M. Dervarics, G. Toth, F. Fulop; Ligand-based prediction of active conformation by 3D-QSAR flexibility descriptors and their application in 3+3D-QSAR models. *J. Med. Chem.* **2005**, *48*, 3239-3250.
- [191] M. I. S. Inc.; MDL ISIS Draw. *San Leandro, California, USA* **2005**,
- [192] J. Gasteiger, C. Hiller, C. Rudolph, J. Sadowski; Automatic-Generation of 3D-Atomic Coordinates for Organic-Molecules. *Abstracts of Papers of the American Chemical Society* **1991**, *202*, 36-Cinf.
- [193] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer; Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030-1037.
- [194] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, V. Steinhauer; Chemical Information in 3D Space. *Abstracts of Papers of the American Chemical Society* **1995**, *210*, 46-Cinf.
- [195] J. Bostrom, J. R. Greenwood, J. Gottfries; Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449-462.

- [196] E. Perola, P. S. Charifson; Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499-2510.
- [197] U. Holzgrabe, K. Mohr; Allosteric modulators of ligand binding to muscarinic acetylcholine receptors. *Drug Discovery Today* **1998**, *3*, 214-222.
- [198] J. M. Luco, F. H. Ferretti; QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392-401.
- [199] A. L. Hopkins, J. S. Ren, H. Tanaka, M. Baba, M. Okamoto, D. I. Stuart, D. K. Stammers; Design of MKC-442 (emivirine) analogues with improved activity against drug-resistant HIV mutants. *J. Med. Chem.* **1999**, *42*, 4500-4505.
- [200] H. Tanaka, R. T. Walker, A. L. Hopkins, J. Ren, E. Y. Jones, K. Fujimoto, M. Hayashi, T. Miyasaka, M. Baba, D. K. Stammers, D. I. Stuart; Allosteric inhibitors against HIV-1 reverse transcriptase: design and synthesis of MKC-442 analogues having an omega-functionalized acyclic structure. *Antiviral Chem. Chemother.* **1998**, *9*, 325-332.
- [201] M. Busemann; Vergleich und Optimierung feldbasierter 3D-QSAR-Techniken. **2003**, *75 Seiten*, *Diplomarbeit*, Institut für Pharmazie und Lebensmittelchemie *Bayerische Julius-Maximilians-Universität Würzburg*.
- [202] A. H. Abadi, S. Lankow, B. Hoefgen, M. Decker, M. U. Kassack, J. Lehmann; Dopamine/serotonin receptor ligands, Part III: Synthesis and biological activities of 7,7'-alkylene-bis-6,7,8,9,14,15-hexahydro-5H-benz[d]indolo[2,3-g]azecines - Application of the bivalent ligand approach to a novel type of dopamine receptor antagonist. *Archiv Der Pharmazie* **2002**, *335*, 367-373.
- [203] M. Decker, D. Appenroth, J. Lehmann, C. Fleck; Testing of novel tri- and tetracyclic inhibitors of cholinesterases in vitro and in vivo on rats. *Naunyn-Schmiedebergs Archives of Pharmacology* **2006**, *372*, 38-38.
- [204] M. Decker, A. König, E. Glusa, J. Lehmann; Synthesis and vasorelaxant properties of hybrid molecules out of NO-donors and the beta-receptor blocking drug propranolol. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4995-4997.
- [205] M. Decker, F. Krauth, J. Lehmann; Novel tricyclic quinazolinimines and related tetracyclic nitrogen bridgehead compounds as cholinesterase inhibitors with selectivity towards butyrylcholinesterase. *Bioorg. Med. Chem.* **2006**, *14*, 1966-1977.
- [206] M. Decker, J. Lehmann; Dopamine receptor ligands. Part VII [1]: Novel 3-substituted 5-phenyl-1,2,3,4,5,6-hexahydro-azepino-[4,5-b]indoles as ligands for the dopamine receptors. *Archiv Der Pharmazie* **2003**, *336*, 466-476.
- [207] M. Decker, J. Lehmann; LE300 - New results on its ability to antagonize the discriminative stimulus effects of cocaine - Dopamine/serotonin receptor antagonists, part XI. *Pharmazie* **2006**, *61*, 248-250.
- [208] M. Decker, T. T. T. Nguyen, J. Lehmann; Investigations into the mechanism of lactamization of lactones yielding in a novel route to biologically active tryptamine derivatives. *Tetrahedron* **2004**, *60*, 4567-4578.
- [209] M. Decker, K. J. Schleifer, M. Nieger, J. Lehmann; Dopamine/serotonin receptor ligands. Part VIII: the dopamine receptor antagonist LE300-modelled and X-ray structure plus further pharmacological characterization, including serotonin receptor binding, biogenic amine transporter testing and in vivo testings. *Eur. J. Med. Chem.* **2004**, *39*, 481-489.
- [210] H. El-Subbagh, T. Wittig, M. Decker, S. Elz, M. Nieger, J. Lehmann; Dopamine/serotonin receptor ligands. Part IV [1]: Synthesis and pharmacology of novel 3-benzazecines and 3-benzazonines as potential 5-HT_{2A} and dopamine receptor ligands. *Archiv Der Pharmazie* **2002**, *335*, 443-448.

- [211] G. P. Gellermann, K. Ullrich, A. Tannert, C. Unger, G. Habicht, S. R. N. Sauter, P. Hortschansky, U. Horn, U. Mollmann, M. Decker, J. Lehmann, M. Fandrich; Alzheimer-like plaque formation by human macrophages is reduced by fibrillation inhibitors and lovastatin. *J. Mol. Biol.* **2006**, *360*, 251-257.
- [212] B. Hoefgen, M. Decker, P. Mohr, A. M. Schramm, S. A. F. Rostom, H. El-Subbagh, P. M. Schweikert, D. R. Rudolf, M. U. Kassack, J. Lehmann; Dopamine/serotonin receptor ligands. 10: SAR studies on azecine-type dopamine receptor Ligands by functional screening at human cloned D-1, D-2L, and D-5 receptors with a microplate reader based calcium assay lead to a novel potent D-1/D-5 selective antagonist. *J. Med. Chem.* **2006**, *49*, 760-769.
- [213] M. U. Kassack, B. Hofgen, M. Decker, N. Eckstein, J. Lehmann; Pharmacological characterization of the benz[d]indolo[2,3-g]azecine LE300, a novel type of a nanomolar dopamine receptor antagonist. *Naunyn-Schmiedebergs Archives of Pharmacology* **2002**, *366*, 543-550.
- [214] P. Mohr, M. Decker, C. Enzensperger, J. Lehmann; Dopamine/serotonin receptor ligands. 12: SAR studies on hexahydro-dibenz[d,g]azecines lead to 4-chloro-7-methyl-5,6,7,8,9,14-hexahydrodibenz[d,g]azecin-3-ol, the first picomolar D-5-selective dopamine-receptor antagonist. *J. Med. Chem.* **2006**, *49*, 2110-2116.
- [215] P. Mohr, J. Lehmann, M. Decker; Synthesis and reactivity of dibenz[d,g]azecin-14(5H)-ones. *Heterocycles* **2006**, *68*, 879-884.
- [216] T. Witt, F. J. Hock, J. Lehmann; 7-methyl-6,7,8,9,14,15-hexahydro-5H-benz[d]indolo[2,3-g]azecine: A new heterocyclic system and a new lead compound for dopamine receptor antagonists. *J. Med. Chem.* **2000**, *43*, 2079-2081.
- [217] T. W. Wittig, M. Decker, J. Lehmann; Dopamine/serotonin receptor ligands. 9. Oxygen-containing mid-sized heterocyclic ring systems and nonrigidized analogues. A step toward dopamine D-5 receptor selectivity. *J. Med. Chem.* **2004**, *47*, 4155-4158.
- [218] T. W. Wittig, C. Enzensperger, J. Lehmann; Dopamine/serotonin receptor ligands VI. Dibenz[g,j]-1-oxa-4-azacycloundecene and dibenz[d,g]-2-azacycloundecene: Synthesis of two new heterocyclic ring systems as potential ligands for dopamine receptor subtypes. *Heterocycles* **2003**, *60*, 887-898.
- [219] <http://de.wikipedia.org/>; Dopamin-Rezeptoren. **2006**,
- [220] E. Mutschler, G. Geisslinger, H. Kroemer, M. Schäfer-Korting; *Arzneimittelwirkungen*, Wissenschaftliche Verlagsges.: **2001**
- [221] Accelrys; DS Viewer Pro. **2002**,
- [222] M. R. McGann, H. R. Almond, A. Nicholls, J. A. Grant, F. K. Brown; Gaussian docking functions. *Biopolymers* **2003**, *68*, 76-90.
- [223] T. Schulz-Gasch, M. Stahl; Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47-57.
- [224] M. Freiwald, A. Valotis, A. Kirschbaum, M. McClellan, T. Murdter, P. Fritz, G. Friedel, M. Thomas, P. Hogger; Monitoring the initial pulmonary absorption of two different beclomethasone dipropionate aerosols employing a human lung reperfusion model. *Respiratory Research* **2005**, *6*, -.
- [225] A. Valotis, P. Hogger; Significant receptor affinities of metabolites and a degradation product of mometasone furoate. *Respiratory Research* **2004**, *5*, -.
- [226] A. Valotis, K. Neukam, O. Elert, P. Hogger; Human receptor kinetics, tissue binding affinity, and stability of mometasone furoate. *J. Pharm. Sci.* **2004**, *93*, 1337-1350.
- [227] A. Valotis; Pharmakokinetische und molekularpharmakodynamische Aspekte inhalativ angewandter Glucocorticoide. *Dissertation* **2005**, *226 Seiten*, Institut für Pharmazie und Lebensmittelchemie Bayerische Julius-Maximilians-Universität Würzburg Würzburg.

- [228] <http://de.wikipedia.org/>; Glukokortikoid. **2006**,
- [229] H. Hatz; *Glucocorticoide*, Wissenschaftliche Verlagsges.: **2005**
- [230] M. Clark, R. D. Cramer, N. van Opdenbosch; Validation of the General-Purpose Tripos 5.2 Force-Field. *J. Comput. Chem.* **1989**, *10*, 982-1012.
- [231] C. Lemmen, C. Hiller, T. Lengauer; RigFit: A new approach to superimposing ligand molecules. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 491-502.
- [232] C. Lemmen, T. Lengauer, G. Klebe; FLEXS: A method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502-4520.
- [233] G. Bringmann, C. Rummey; 3D QSAR investigations on antimalarial naphthylisoquinoline alkaloids by comparative molecular similarity indices analysis (CoMSIA), based on different alignment approaches. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 304-316.
- [234] M. Wagener, J. Sadowski, J. Gasteiger; Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.
- [235] A. Kulkarni, A. J. Hopfinger, R. Osborne, L. H. Bruner, E. D. Thompson; Prediction of eye irritation from organic chemicals using membrane-interaction QSAR analysis. *Toxicol. Sci.* **2001**, *59*, 335-345.
- [236] S. R. Krystek, J. T. Hunt, P. D. Stein, T. R. Stouch; 3-Dimensional Quantitative Structure-Activity-Relationships of Sulfonamide Endothelin Inhibitors. *J. Med. Chem.* **1995**, *38*, 659-668.
- [237] D. D. Robinson, P. J. Winn, P. D. Lyne, W. G. Richards; Self-organizing molecular field analysis: A tool for structure-activity studies. *J. Med. Chem.* **1999**, *42*, 573-583.
- [238] D. B. Turner, P. Willett; The EVA spectral descriptor. *Eur. J. Med. Chem.* **2000**, *35*, 367-375.
- [239] M. A. Avery, M. Alvim-Gaston, C. R. Rodrigues, E. J. Barreiro, F. E. Cohen, Y. A. Sabnis, J. R. Woolfrey; Structure-activity relationships of the antimalarial agent artemisinin. 6. The development of predictive in vitro potency models using CoMFA and HQSAR methodologies. *J. Med. Chem.* **2002**, *45*, 292-303.
- [240] B. Siedle, A. J. Garcia-Pineros, R. Murillo, J. Schulte-Monting, V. Castro, P. Rungeler, C. A. Klaas, F. B. Da Costa, W. Kisiel, I. Merfort; Quantitative structure - Activity relationship of sesquiterpene lactones as inhibitors of the transcription factor NF-kappa B. *J. Med. Chem.* **2004**, *47*, 6042-6054.
- [241] S. Wagner, A. Hofmann, B. Siedle, L. Terfloth, I. Merfort, J. Gasteiger; Development of a structural model for NF-kappa B inhibition of sesquiterpene lactones using self-organizing neural networks. *J. Med. Chem.* **2006**, *49*, 2241-2252.
- [242] H. Vandervoet; Comparing the Predictive Accuracy of Models Using a Simple Randomization Test. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 313-323.
- [243] P. Atkins; *Physical Chemistry*, R.J. Axford: Cichester, Sussex, UK **1987**
- [244] M. D. Cummings, R. L. DesJarlais, A. C. Gibbs, V. Mohan, E. P. Jaeger; Comparison of virtual screening programs. *Abstracts of Papers of the American Chemical Society* **2003**, *226*, U456-U456.
- [245] M. D. Cummings, R. L. DesJarlais, A. C. Gibbs, V. Mohan, E. P. Jaeger; Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962-976.
- [246] S. Prilla, J. Schrobang, J. Ellis, H. D. Holtje, K. Mohr; Allosteric interactions with muscarinic acetylcholine receptors: Complex role of the conserved tryptophan M-2 (422)Trp in a critical cluster of amino acids for baseline affinity, subtype selectivity, and cooperativity. *Mol. Pharmacol.* **2006**, *70*, 181-193.

- [247] U. Voigtlander, K. Jöhren, M. Mohr, A. Raasch, C. Trankle, S. Buller, J. Ellis, H. D. Holtje, K. Mohr; Allosteric site on muscarinic acetylcholine receptors: Identification of two amino acids in the muscarinic M-2 receptor that account entirely for the M-2/M-5 subtype selectivities of some structurally diverse allosteric ligands in N-methylscopolamine-occupied receptors. *Mol. Pharmacol.* **2003**, *64*, 21-31.
- [248] M. U. Kassack, B. Höfgen, J. Lehmann, N. Eckstein, J. M. Quillan, W. Sadee; Functional screening of G protein-coupled receptors by measuring intracellular calcium with a fluorescence microplate reader. *Journal of Biomolecular Screening* **2002**, *7*, 233-246.
- [249] Y. Cheng, W. H. Prusoff; Relationship between Inhibition Constant (K₁) and Concentration of Inhibitor Which Causes 50 Per Cent Inhibition (I₅₀) of an Enzymatic-Reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099-3108.
- [250] J. Scheiber; Die Entwicklung einer auf Moleküldynamik basierenden Methode zur Homologie-Modellierung der Tertiärstruktur von Proteinen. Diplomarbeit **2003**, *106 Seiten*, Institut für Biophysik und Physikalische Biochemie *Universität Regensburg*.
- [251] T. Okada, M. Sugihara, A. N. Bondar, M. Elstner, P. Entel, V. Buss; The retinal conformation and its environment in rhodopsin in light of a new 2.2 angstrom crystal structure. *J. Mol. Biol.* **2004**, *342*, 571-583.
- [252] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, J. D. Thompson; Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* **2003**, *31*, 3497-3500.
- [253] F. Jeanmougin, J. D. Thompson, M. Gouy, D. G. Higgins, T. J. Gibson; Multiple sequence alignment with Clustal x. *Trends in Biochemical Sciences* **1998**, *23*, 403-405.
- [254] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, D. G. Higgins; The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **1997**, *25*, 4876-4882.
- [255] D. G. Higgins, J. D. Thompson, T. J. Gibson; Using CLUSTAL for multiple sequence alignments. *Computer Methods for Macromolecular Sequence Analysis* **1996**, *266*, 383-402.
- [256] J. D. Thompson, D. G. Higgins, T. J. Gibson; Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* **1994**, *22*, 4673-4680.
- [257] J. Ballesteros, H. Weinstein; Integrated methods for the construction of three-dimensional models of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **1995**, *25*, 366-428.
- [258] L. Shi, J. A. Javitch; The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol.* **2002**, *42*, 437-467.
- [259] H. Xhaard, V. V. Rantanen, T. Nyronen, M. S. Johnson; Molecular evolution of adrenoceptors and dopamine receptors: Implications for the binding of catecholamines. *J. Med. Chem.* **2006**, *49*, 1706-1719.
- [260] <http://de.wikipedia.org/>: G-Protein-gekoppelter Rezeptor. **2006**,
- [261] H. X. Lan, C. J. DuRand, M. M. Teeter, K. A. Neve; Structural determinants of pharmacological specificity between D-1 and D-2 dopamine receptors. *Mol. Pharmacol.* **2006**, *69*, 185-194.
- [262] S. W. Lin, T. P. Sakmar; Specific tryptophan UV-absorbance changes are probes of the transition of rhodopsin to its active state. *Biochemistry* **1996**, *35*, 11149-11159.
- [263] J. J. Ruprecht, T. Mielke, R. Vogel, C. Villa, G. F. X. Schertler; Electron crystallography reveals the structure of metarhodopsin I. *EMBO J.* **2004**, *23*, 3609-3620.
- [264] B. Kobilka; Agonist binding: A multistep process. *Mol. Pharmacol.* **2004**, *65*, 1060-1062.

- [265] J. C. Caron, B. Shroot; Determination of Partition-Coefficients of Glucocorticosteroids by High-Performance Liquid-Chromatography. *J. Pharm. Sci.* **1984**, *73*, 1703-1706.
- [266] A. Leo, C. Hansch, P. Y. C. Jow; Dependence of Hydrophobicity of Apolar Molecules on Their Molecular Volume. *J. Med. Chem.* **1976**, *19*, 611-615.
- [267] G. Würthwein, S. Rehder, P. Rohdewald; Lipophilicity and receptor affinity of glucocorticoids. *Pharm. Ztg. Wissensch.* **1992**, *137*, 161-167.
- [268] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko; Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453-463.
- [269] A. Valotis; Würzburg, Germany **2006**, Relative Retentionszeiten der untersuchten Glukokortikoide.
- [270] E. A. Coats; The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspectives in Drug Discovery and Design* **1998**, *12*, 199-213.
- [271] J. Draize, G. Woodard, H. Calvery; Methods for the Study of Irritation and Toxicity of Substances Applied to the Skin and Mucous Membranes. *J. Pharmacol. Exp. Ther.* **1944**, *82*, 377-390.

Lebenslauf

Persönliches:

Name: Josef Heinrich Scheiber
Geburtstag: 20. April 1978
Geburtsort: Tirschenreuth

Schulbildung:

09/1984 – 07/1988 Marienschule Tirschenreuth (Grundschule)
09/1988 – 06/1997 Stiftland-Gymnasium Tirschenreuth
Abschluss: Abitur

Grundwehrdienst:

07/1997 – 04/1998 Grundwehrdienst bei 5./Pionierbataillon 4 in Bogen/Donau

Studium:

10/1998 – 08/2000 Grundstudium Diplom-Biologie an der Universität Regensburg
09/2000 Vordiplom
10/2000 – 05/2002 Hauptstudium mit den Schwerpunkten Biophysik / Bioinformatik / Organische Chemie / Zellbiologie an der Universität Regensburg
06/2002 Diplom-Prüfung
07/2002 – 07/2003 Diplomarbeit am Institut für Biophysik und physikalische Biochemie der Universität Regensburg mit dem Thema „Entwicklung einer auf Moleküldynamik basierenden Methode zur Homologiemodellierung der Tertiärstruktur von Proteinen“ bei Prof. Dr. Dr. Hans Robert Kalbitzer
07/2003 Abschluss: Diplom-Biologe

Promotion:

11/2003 – 11/2006 Promotion als wissenschaftlicher Mitarbeiter am Institut für Pharmazie und Lebensmittelchemie der Universität Würzburg bei Prof. Dr. Knut Baumann (jetzt an der TU Braunschweig)

Postdoc:

ab 12/2006 Postdoctoral Fellow
Discovery Technologies Informatics
Novartis Institutes of Biomedical Research
Cambridge, Massachusetts, USA

Auszeichnungen und Stipendien:

09/2004 Lesmüller-Posterpreis der DPhG
05/2005 Travel Grant der MGMS-DS
08/2005 Travel Grant des EU-ASIA-Netzwerks
09/2005 Best Poster Prize der MGMS
04/2006 Accelrys-Posterpreis der UK QSAR Society
05/2006 Travel Grant der MGMS-DS
05/2006 Vortragspreis der MGMS-DS
09/2006 Vortragspreis der DPhG

