Entwicklung chemometrischer Methoden für das *in-silico*-Wirkstoffdesign

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

vorgelegt von

Matthias Busemann
aus Würzburg

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 dieser Dissertation wurden bereits an folgenden Stellen veröffentlicht:

Originalpublikationen

R. Vicik, M. Busemann, K. Baumann, T. Schirmeister. Inhibitors of Cysteine Proteases. *Curr. Top. Med. Chem.* **2006**, 6, 331–353.

R. Vicik, M. Busemann, C. Gelhaus, N. Stiefl, J. Scheiber, W. Schmitz, F. Jenke, M. Mladenovic, B. Engels, M. Leippe, K. Baumann, T. Schirmeister. Aziridide based inhibitors of cathepsin L — Docking of compounds with non-planar amide bonds. (*submitted*)

Konferenzbeiträge

M. Busemann, K. Baumann. Detection of prediction outliers in QSAR analysis using descriptor data only. In: E. Aki-Sener, I. Yalcin (Hrsg.), *EuroQSAR 2004, QSAR & Molecular Modelling in Rational Design of Bioactive Molecules*, Computer Aided Drug Design & Development Society in Turkey, Ankara, Türkei, **2006**, S. 144–146.

Posterbeiträge

M. Busemann, K. Baumann. Improvement of QSAR analyses by detecting prediction outliers. Jahrestagung der Deutschen Pharmazeutischen Gesellschaft, Regensburg **2004**.

M. Busemann, K. Baumann. Detection of prediction outliers in QSAR analysis using descriptor data only. 15th European Symposium on QSAR, Istanbul, Türkei **2004**.

M. Busemann, N. Stiefl, T. Schirmeister, K. Baumann. 3D-QSAR Studies of epoxysuccinyl peptides as potent inhibitors of Cathepsin B. Jahrestagung der Deutschen Pharmazeutischen Gesellschaft, Würzburg **2003**.

Diese Dissertation entstand im Zeitraum von
Oktober 2003 bis Januar 2006
unter Anleitung von Herrn PD Dr. Knut Baumann,
Institut für Pharmazie und Lebensmittelchemie, Universität Würzburg.

Durchgeführt wurden die Arbeiten in der
Abteilung Chem- & Bioinformatics der 4SC AG, Martinsried.

Meinen Eltern

In Erinnerung an Loni Breunig

Inhaltsverzeichnis

Abkürzungen und Symbole

$\mathbf{1_n}$	Spaltenvektor $n \times 1$, alle Elemente gleich Eins
1D	Eindimensional
2D	Zweidimensional
3D	Dreidimensional
_autoskal	autoskaliert
Azet	Azetidin-2-carbonsäure
Azi	Aziridin-2,3-dicarbonsäure
Azy	Aziridin-2-carbonsäure
b	Regressionskoeffizient
BOC	*tert*-Butyloxycarbonyl
Bin/Avg	Binär, Durchschnitt
Bin/Min	Binär, Minimum
c	Cut-off (Grenzwert)
CB	Cathepsin B
CL	Cathepsin L
CV	Cross-validation
_CV	Kreuzvalidiert
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis
D	Dimensionalität
DIBSI	Distance-Based Similarity Search
_dir	Direkt (im Gegensatz zu _ens)
e	Residuum
ED	Euklidische Distanz
ED/Avg	Euklidische Distanz, Durchschnitt
ED/Min	Euklidische Distanz, Minimum
EF	Enrichment Factor
_ens	Ensemble
ΔG	Freie Bindungsenthalpie
GPCR	G-Protein Coupled Receptor
HOMO	Highest Occupied Molecular Orbital
$\mathbf{I_n}$	Einheitsmatrix mit der Dimension $n \times n$
IC_{50}	Inhibitorkonzentration mit 50prozentiger Inhibition

Int/Avg	Integer, Durchschnitt
Int/LDA15	Integer, LDA mit 15 Hauptkomponenten
Int/LDA80	Integer, LDA mit 80 Hauptkomponenten
Int/Min	Integer, Minimum
k	Ensemblegröße
L50%O-CV	LMO-CV mit 50 Prozent ausgelassenen Daten
LDA	Lineare Diskriminanzanalyse
LMO-CV	Leave-multiple-out Cross-validation
LOO-CV	Leave-one-out Cross-validation
logP	logarithmierter Oktanol-Wasser-Verteilungskoeffizient
LR	Lineare Regression
LUMO	Lowest Unoccupied Molecular Orbital
MD	Mahalanobis-Distanz
MDDR	MDL Drug Data Report
MLR	Multiple Lineare Regression
MSEP	Mean Squared Error of Prediction
n	Anzahl der Zeilen (Matrix) bzw. Elemente (Vektor)
N_{el}	Anzahl eliminierter Objekte
NND	Nearest neighbour distance
$\mathrm{NND}_{\mathrm{Te/Tr}}$	NND zwischen Test- und Trainingsdaten
$\mathrm{NND}_{\mathrm{Tr/Tr}}$	NND der Trainingsdaten untereinander
Nip	Nipecotinsäure
ODD	Outlier Detection by Distance towards training data
$_{-\mathrm{orig}}$	original
p	Anzahl der Variablen bzw. der Spalten einer Matrix
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PDB	Protein Data Bank
PLS	Partial Least Squares
PRESS	Predictive Residual Sum of Squares
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
RMSD	Root Mean Squared Deviation
RMSEP	Root Mean Squared Error of Prediction

$RMSEP_{el}$	RMSEP nach Outlier-Eliminierung
RSS	Residual Sum of Squares
s	Anzahl Simulationen
σ^2	Standardabweichung
SIFt	Structural Interactions Fingerprint
SMARTS	Smiles Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry Specification
SVD	Singular Value Decomposition
$_{\text{Test}}$	Testdaten
$_{\text{Train}}$	Trainingsdaten
TS	Tabu-Suche
T_c	Tanimoto-Koeffizient
vHTS	Virtual High-Throughput Screening
VS	Virtual Screening
WDI	World Drug Index
y	Experimentell bestimmte Aktivität
$\hat{y}$	Vorhergesagte Aktivität
$_{\text{zentr}}$	zentriert

Mathematische Notation

In dieser Arbeit werden Skalare in kursiven Kleinbuchstaben (x) dargestellt, Vektoren in fetten Kleinbuchstaben ($\mathbf{x}$) und Matrizen in fetten Großbuchstaben ($\mathbf{X}$).

Vektoren werden stets als Spaltenvektoren angegeben. Die transponierte Form eines Vektors $\mathbf{x}$ wird mit $\mathbf{x}^T$ bezeichnet, gleiches gilt für transponierte Matrizen ($\mathbf{X}^T$). Die Dimension einer Matrix wird durch die Anzahl n ihrer Zeilen und die Anzahl p ihrer Spalten charakterisiert. Die i-te Spalte der Matrix $\mathbf{X}$ ist der Vektor $\mathbf{x_i}$.

Geschätzte Vektoren werden als $\hat{\mathbf{x}}$ symbolisiert. Die Norm eines Vektors $\mathbf{x}$ ist $\|\mathbf{x}\|$. Der Mittelwert über alle Elemente von $\mathbf{x}$ heißt $\bar{x}$; entsprechend ergibt die Gesamtheit aller Spaltenmittelwerte einer Matrix den Vektor $\bar{\mathbf{x}}$.

Hinweise zum Sprachgebrauch

Die deutsche Übersetzung üblicher englischer Begriffe wurde in dieser Arbeit der besseren Lesbarkeit und Konsistenz des Textes untergeordnet. Während etwa „Outlier“ ohne weiteres mit „Ausreißer“ übersetzt werden könnte, würde die sprachliche Einheitlichkeit bereits mit dem Begriff „Inlier“ gesprengt, für den kein deutscher Ausdruck gebräuchlich ist. Da Englisch als Sprache der internationalen Wissenschaft etabliert ist, wird die maßvolle Verwendung von Anglizismen auch in der vorliegenden Dissertation für tolerierbar gehalten.

Diese Arbeit folgt den bis 1998 verbindlichen Regeln der deutschen Rechtschreibung.

Kapitel 1

Einleitung

Ratlosigkeit und Unzufriedenheit sind die ersten Vorbedingungen des Fortschritts.
—Thomas Alva Edison, US-amerikanischer Ingenieur und Erfinder, † 1931

Die Entwicklung eines neuen Medikaments ist ein immens kostspieliger und zeitaufwendiger Prozeß; laut FDA* liegt die durchschnittliche Gesamtdauer bis zur Markteinführung bei 12 Jahren. Die Kosten summieren sich in dieser Zeit im Mittel auf etwa 800 Millionen US-Dollar. Pro Jahr wenden die Pharmaunternehmen Europas und Nordamerikas insgesamt 20 Milliarden US-Dollar für die Erforschung und Entwicklung neuer Wirkstoffe auf.[1]

Für die pharmazeutische Forschung ist es von höchstem Interesse, die begrenzten Ressourcen optimal einzusetzen. Schon aufgrund der geschätzten Anzahl von 10^{60} theoretisch synthetisierbaren *drug-like*-Verbindungen[2,3], also Molekülen mit prinzipiell wirkstoffähnlichen Strukturen und Eigenschaften, kann die Wirkstoffsuche allein nach dem Prinzip von Versuch und Irrtum nicht effektiv sein. Vielmehr ist ein rationales Wirkstoffdesign gefragt, das heißt die Bemühungen gehen in Richtung des zielgerichteten und maßgeschneiderten Entwurfs aktiver Moleküle.[4] Innerhalb dieser Randbedingungen hat der Einsatz computergestützter Verfahren — sogenannter *in-silico*-Methoden† — in den letzten Jahren stetig an Bedeutung gewonnen.

*Food and Drug Administration (US-amerikanische Zulassungsbehörde); Zahlen aus dem Jahr 2004.

†*in silico*: „im Computer"; Analogie zu den etablierten Begriffen *in vitro* (im Reagenzglas) und *in vivo* (im lebenden Organismus).

Gerade die Anwendung strukturbasierter Techniken hat wiederholt zu Erfolgen in der Wirkstoffentwicklung geführt.[5] Dabei wird die beispielsweise röntgenkristallographisch erhaltene dreidimensionale Struktur eines Zielproteins genutzt, um einen paßgenauen Inhibitor zu entwerfen — analog dem Zurechtfeilen eines Schlüsselrohlings für ein Schloß mit bekanntem Schließmechanismus. Die in jüngster Zeit wieder aktuell gewordenen Neuraminidase-Hemmer Relenza[6] (GSK) und Tamiflu[7] (Roche) sind Beispiele für das erfolgreiche strukturbasierte Design eines Wirkstoffs.

Doch auch die sogenannten ligandbasierten Verfahren besitzen weiterhin ihre Daseinsberechtigung.[8] Sie kommen zum Einsatz, wenn bereits die Aktivität einer Serie von Verbindungen bekannt ist und spiegeln gewissermaßen eines der Grundprinzipien der Wissenschaft wider: Durch genaue Beobachtung des Bekannten können Gesetzmäßigkeiten abgeleitet und auf das Unbekannte übertragen werden. So ist es möglich, ligandbasierte Information für die Formulierung einer quantitativen Beziehung von Struktur und Wirkung zu nutzen, die zur Vorhersage der Aktivität neuer Moleküle befähigt.

Nie aus dem Blickfeld geraten darf jedoch das kritische Hinterfragen computergestützter Vorhersagen. Ein blindes Vertrauen in die *in-silico*-Techniken ist nicht zielführend, denn „diese Methoden sind mächtige Hilfsmittel beim rationalen Wirkstoffdesign, doch sie werden häufig inadäquat eingesetzt. Wir betreiben Forschung *in vitro* und *in silico*, vernachlässigen dabei aber die so wichtige *in-cerebro*-Komponente der Wirkstoffentwicklung.“[9]

Die in dieser Dissertation präsentierte Entwicklung einer Methode zur Outlier-Identifizierung liefert (obgleich auch sie computergestützt arbeitet) einen Beitrag zur kritischen Überprüfung eines häufig angewandten ligandbasierten Verfahrens. Sie geht der Frage nach, ob ein Modell (eine quantitative Struktur-Wirkungs-Beziehung) überhaupt auf die gewünschten Moleküle anwendbar ist und wo seine Grenzen liegen: Wer nämlich in blindem Vertrauen ein etwa in Kairo zuverlässig funktionierendes Wettermodell auf Meßwerte in Hamburg anwendet, wird möglicherweise schnell im Regen stehen.

Die ebenfalls vorgestellte Untersuchung von Ensemble-Techniken zielt in dieselbe Richtung. Sie liefert Hinweise darauf, wie Erstellung und Kombination vieler leicht differierender Modelle die Zuverlässigkeit der Vorhersage im Vergleich zu einem Einzelmodell erhöhen können.

Darüber hinaus wurden Methoden entwickelt, die für das sogenannte virtuelle Screening nützlich sind. Mit diesem Ansatz werden große virtuelle Substanzbibliotheken durchsucht, die oft mehrere Millionen Moleküle enthalten. Ziel ist es, eine Vorauswahl aussichtsreicher Verbindungen zu identifizieren, die anschließend experimentell getestet werden. Bei der sprichwörtlichen Suche nach der Nadel im Heuhaufen kommt dem virtuellen Screening also die Aufgabe zu, vorab die Größe des Heuhaufens zu reduzieren und so die Erfolgsaussichten für nachfolgende Suchverfahren zu erhöhen. Dazu können ligand- und strukturbasierte Verfahren gleichermaßen herangezogen werden; in dieser Arbeit wurde vor allem der Frage nachgegangen, wie die Information aus diesen häufig separat betrachteten Teilbereichen effektiv zusammengeführt und der durch die Kombination entstehende Erkenntnisgewinn für das Wirkstoffdesign nutzbar gemacht werden kann.

Schließlich wurden Dockingexperimente durchgeführt, die zum genaueren Verständnis der Selektivität einer Serie von Cysteinprotease-Inhibitoren führten. Die Herausforderung lag dabei in der kovalenten Bindung dieser Inhibitoren, die die Entwicklung eines angepaßten Dockingprotokolls erforderte. Der hier gewählte Weg besitzt eine breite Anwendungsmöglichkeit und kann daher auch dem Docking an anderen kovalenten Protein-Ligand-Systemen zugutekommen.

Die im Rahmen der vorliegenden Doktorarbeit entwickelten und untersuchten Methoden erstrecken sich also auf ein breites Spektrum der aktuellen *in-silico*-Wirkstoffentwicklung und adressieren einige in der Praxis häufig auftretende Probleme. Auch wenn computergestützte Verfahren allein keine Wunder vollbringen können, so haben sie sich doch in den vergangenen Jahren und Jahrzehnten zu einem aus der modernen Wirkstoffentwicklung nicht mehr wegzudenkenden Hilfsmittel etabliert; angesichts der stetig anwachsenden Informationsfülle wird ihre Bedeutung weiter zunehmen: „*In-silico*-Verfahren werden zwangsläufig ihren Platz in der pharmazeutischen und biotechnologischen Industrie finden. Wie es kürzlich ausgedrückt wurde: ‚The *Insilicoids* are coming and will save the world'[10].“[11]

Kapitel 2

Theoretische Grundlagen

Klarheit ist ein intellektueller Wert an sich; Genauigkeit und Präzision aber sind es nicht. Absolute Präzision ist unerreichbar; und es ist zwecklos, genauer sein zu wollen, als es unsere Problemsituation verlangt.
—Karl R. Popper, Philosoph und Wissenschaftstheoretiker, † 1994

Die vorliegende Dissertation beschäftigt sich mit verschiedenen Aspekten der *in-silico*-Wirkstoffentwicklung. Diese Disziplin der modernen chemisch-pharmazeutischen Forschung bedient sich leistungsfähiger Computerverfahren, um die Entwicklung von Wirkstoffen zu beschleunigen und durch effizientere Abläufe die Kosten zu reduzieren. Eine zentrale Methode ist dabei die Analyse von quantitativen Struktur-Wirkungs-Beziehungen (engl. *Quantitative structure-activity relationship*, QSAR).

Ziel der QSAR-Analyse ist die Vorhersage der biologischen Aktivität von Substanzen. Bereits im Jahr 1868 gelangten CRUM-BROWN und FRASER durch Untersuchungen an Alkaloiden zu der Erkenntnis, daß die physiologische Aktivität Φ einer Substanz eine Funktion ihrer chemischen Konstitution C darstellt.[12] Sie formulierten diese Beziehung mit der einfachen Gleichung

$$\Phi = f(C) \tag{2.1}$$

Ein Jahrhundert später schrieben HANSCH und FUJITA[13] bzw. FREE und WILSON[14] mit ihren Modellen zur Beschreibung quantitativer Struktur-Wirkungs-Beziehungen diese Idee fort, indem sie die biologische Aktivität ei-

nes Moleküls mit dessen strukturellen und physikochemischen Eigenschaften korrelierten. Die Publikation ihrer Methoden im Jahr 1964 gilt heute als die Geburtsstunde der modernen QSAR-Analyse,[15] die sich unverändert auf die fundamentale Theorie stützt, daß die makroskopischen Eigenschaften einer Substanz durch ihre molekulare Struktur bestimmt sind.[16]

In der aktuellen Wirkstoffentwicklung stellt die QSAR-Analyse eine fest etablierte Methode dar, ermöglicht sie doch die Vorhersage von Eigenschaften materiell noch gar nicht vorliegender Verbindungen. Statt der zeit- und kostenaufwendigen chemischen Synthese von Substanzen und deren anschließender Untersuchung etwa im Hinblick auf eine bestimmte biologische Wirkung genügt das virtuelle Molekül im Computer. Die formale Beschreibung seiner Struktur liefert zusammen mit einigen daraus schnell zu berechnenden physikalisch-chemischen Parametern eine breite Datengrundlage für die Vorhersage der biologischen Aktivität.

Aus dem Alltag ist eine ganz ähnliche Vorgehensweise bekannt: Die Wettervorhersage. Durch die einfache Bestimmung einer Reihe von Meßwerten wie etwa Temperatur, Luftdruck, Windgeschwindigkeit etc. ist mit Hilfe eines zuvor erstellten Wettermodells die Vorhersage des komplexen Wettergeschehens der folgenden Tage möglich.

So wie das Wettermodell „lernt“, aus den heute gemessenen meteorologischen Parametern eine Schlußfolgerung für das morgige Wetter zu ziehen, so muß auch in der QSAR-Analyse zunächst ein Modell „trainiert“ werden, um anschließend Vorhersagen treffen zu können. So wie das Wettermodell durch wiederholte Beobachtung den Zusammenhang zwischen fallendem Luftdruck und nahender Kaltfront „erkennt“, so stellt auch das QSAR-Modell anhand einer Serie experimentell untersuchter Verbindungen eine Korrelation zwischen den physikalisch-chemischen Eigenschaften von Molekülen und der daraus resultierenden biologischen Wirkung her.

2.1 Beschreibung von Molekülen

Allen Ansätzen der QSAR-Analyse gemein ist die Notwendigkeit, die betrachteten Moleküle zunächst in numerischer Form zu repräsentieren und ihre Eigenschaften mathematisch faßbar zu beschreiben. Diese Aufgabe erfül-

len die sogenannten Deskriptoren. Ebenso wie sich zum Beispiel ein Automobil durch die Angabe von Motorleistung, Kraftstoffverbrauch, Drehmoment, Anzahl der Türen oder Kofferraumvolumen beschreiben läßt, kann auch eine chemische Verbindung durch einfach zu bestimmende Parameter charakterisiert werden.

2.1.1 Deskriptoren

Heute steht in der QSAR eine Vielzahl von Deskriptoren zur Verfügung, die alle erdenklichen Eigenschaften eines Moleküls beschreiben und so eine Fülle von Informationen für die mathematische Modellierung zur Verfügung stellen (eine umfassende Übersicht findet sich bei TODESCHINI[17]). Das Spektrum reicht von einfachen Zähl-Deskriptoren, die die Anzahl bestimmter Elemente, funktioneller Gruppen oder Bindungstypen erfassen, über physikochemische Parameter wie Molekulargewicht, logP oder van-der-Waals-Oberfläche bis hin zu komplexen quantenmechanischen Deskriptoren.

Ein wichtiges Klassifizierungsmerkmal solcher Deskriptoren ist die Dimensionalität der zugrundeliegenden Molekülrepräsentation. Während etwa das Molekulargewicht einen klassischen eindimensionalen Deskriptor darstellt und Informationen über die Konnektivität vollständig aus einer zweidimensionalen Darstellung der Moleküle zugänglich sind, wird für die Berechnung von 3D-Deskriptoren wie dem Molekülvolumen eine dreidimensionale Repräsentation benötigt.

Eine Untergruppe der 3D-Deskriptoren bilden wiederum die translations- und rotationsvarianten Deskriptoren. Dabei handelt es sich um Deskriptoren, die von der Orientierung des Moleküls im Raum abhängig sind, beispielsweise das Trägheitsmoment entlang einer ausgewählten Koordinatenachse. Auch alle gitterbasierten Ansätze[18] wie etwa CoMFA[19] (*Comparative Molecular Field Analysis*), CoMSIA[20] (*Comparative Molecular Similarity Indices Analysis*) oder GRID[21,22] liefern solche translations- und rotationsvarianten Deskriptoren (siehe Abb. 2.1). Um für mehrere Moleküle untereinander vergleichbare Deskriptorinformationen zu erhalten, ist daher in diesem Fall zunächst eine gleichartige räumliche Ausrichtung (engl. *Alignment*) zwingend notwendig.

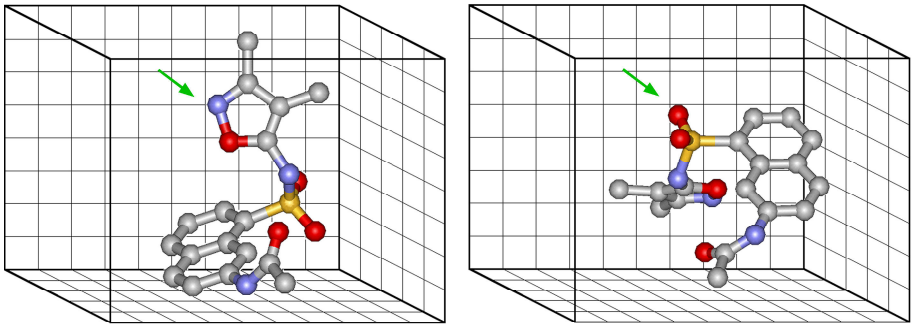


Abbildung 2.1 Translations- und Rotationsvarianz feldbasierter 3D-Deskriptoren. An einem definierten Punkt des gedachten dreidimensionalen Gitters (grüner Pfeil) ergeben sich für ein und dasselbe Molekül bei unterschiedlicher Orientierung unterschiedliche Eigenschaftswerte. Der Deskriptor ist also translations- und rotationsvariant und macht eine gleichsinnige Ausrichtung (engl. *Alignment*) aller Moleküle notwendig.

2.1.2 Fingerprints

Eine spezielle Form von Deskriptoren stellen die sogenannten Fingerprints dar. Dabei handelt es sich um Fragmentvektoren, die die Moleküle entweder anhand der Häufigkeit oder aber anhand der bloßen Existenz bestimmter Substrukturen charakterisieren. Im letzteren Fall besteht der Fingerprint also nur aus Ja/Nein-Informationen (z.B. Carboxylgruppe vorhanden oder nicht), die numerisch als 0 oder 1 dargestellt werden; aus diesem Grund ist auch die Bezeichnung *Bitstring*[*] gebräuchlich.

Eine Schwierigkeit besteht in der Auswahl der berücksichtigten Substrukturen. Ein klassischer 1D-Deskriptor wie das Molekulargewicht oder die van-der-Waals-Oberfläche ist auf jedes denkbare Molekül anwendbar. Zählt ein Fingerprint dagegen nur die Häufigkeit von Substrukturen, die in der gesamten Serie von Molekülen gar nicht vorkommen, so liefert er keine nützliche Information. Etablierte kommerzielle Fingerprints wie beispielsweise die Daylight-Fingerprints[23] verwenden deshalb Substrukturen, die aus großen Bibliotheken von bekannten Wirkstoffmolekülen extrahiert wurden und somit eine repräsentative Auswahl relevanter Strukturmerkmale darstellen.

[*]Bit: b*inary dig*it; eine binäre Variable, die die Zustände 0 oder 1 annehmen kann.

2.2 Multivariate Datenanalyse

Neben einer adäquaten Beschreibung der Eigenschaften von Molekülen liegt die zweite große Herausforderung der QSAR-Analyse in der mathematischen Modellierung des Zusammenhangs zwischen diesen Eigenschaften und den beobachteten biologischen Aktivitäten. In der Geschichte der feldbasierten 3D-QSAR-Techniken stellte sich der (retrospektiv nur vermeintliche) Mangel geeigneter mathematischer Methoden für die Analyse der Daten sogar als limitierender Schritt heraus, der die praktische Umsetzung der Idee über mehrere Jahre hinweg verhinderte.[24]

Das Ziel der mathematischen Modellierung der mit Hilfe von Deskriptoren generierten Daten ist das Sicht- und Nutzbarmachen des Zusammenhangs zwischen der biologischen Aktivität und den molekularen Eigenschaften einer chemischen Substanz, also die Korrelation von X-Daten (unabhängige Variablen, Deskriptoren) und Y-Daten (abhängige Variablen, Aktivitäten). Es soll eine mathematische Gleichung gefunden werden, die eine Vorhersage der Aktivität neuer Moleküle allein aus den zugehörigen Deskriptorwerten ermöglicht.

2.2.1 Datenvorbehandlung

Die absoluten Werte einzelner Variablen sowie die Intervalle, innerhalb derer sie streuen, sind oft nicht untereinander vergleichbar, da sie auf verschiedenen Skalen gemessen werden. Während etwa das Molekulargewicht typischer Wirkstoffmoleküle zwischen 200 und 500 $g \cdot mol^{-1}$ beträgt, nimmt der Deskriptor logP üblicherweise nur Werte < 5 an. Da die Regressionsanalyse von diesem Unterschied beeinflußt werden kann, wird zumeist eine Zentrierung und Autoskalierung der Daten vorgenommen[25].

2.2.1.1 Zentrierung

Eine Datenmatrix $\mathbf{X}$ mit n Zeilen und p Spalten enthalte die jeweils p Deskriptorwerte für die n Moleküle. Für jeden Deskriptor-Vektor $\mathbf{x_i}$ 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 so von der

Rohdatenmatrix subtrahiert, daß jedes Matrixelement um den zugehörigen Spaltenmittelwert vermindert wird:

$$\mathbf{X}_{\text{zentr}} = \mathbf{X}_{\text{orig}} - \mathbf{1}_n \cdot \bar{\mathbf{x}}^T \tag{2.2}$$

Dabei bezeichnet $\mathbf{1}_n$ einen Spaltenvektor der Dimension $n \times 1$, dessen Elemente alle gleich Eins sind.

Die Y-Daten (z. B. biologische Aktivitäten) werden gleichermaßen gemäß

$$\mathbf{y}_{\text{zentr}} = \mathbf{y}_{\text{orig}} - \mathbf{1}_n \cdot \bar{y} \tag{2.3}$$

zentriert, wobei $\bar{y}$ der Mittelwert über alle originalen Y-Daten ist.

Sofern nicht ausdrücklich anders angegeben, sind alle in dieser Arbeit vorgestellten Daten zentriert.

2.2.1.2 Autoskalierung

Bei der Autoskalierung wird jede Variable des Datensatzes mit dem Kehrwert ihrer Standardabweichung gewichtet.

$$\mathbf{X}_{\text{autoskal}} = \mathbf{X}_{\text{orig}} \cdot (\mathbf{1}_n \cdot \text{inv}(\text{std}(\mathbf{X}_{\text{orig}}))) \tag{2.4}$$

Hier ist $\text{inv}(\text{std}(\mathbf{X}_{\text{orig}}))$ eine Funktion, die als Ergebnis einen Zeilenvektor der Dimension $1 \times p$ liefert, der die Kehrwerte der Standardabweichungen der p Spalten von $\mathbf{X}_{\text{orig}}$ enthält.

Sofern nicht ausdrücklich anders angegeben, sind alle in dieser Arbeit vorgestellten Daten autoskaliert.

2.2.2 Regressionstechniken

In der QSAR wird meist ein linearer Zusammenhang zwischen der Datenmatrix $\mathbf{X}$ und den Aktivitäten $\mathbf{y}$ angenommen. Dementsprechend finden hauptsächlich lineare Regressionstechniken Anwendung. Eine Ausnahme bilden die Neuronalen Netzwerke, die in der Vergangenheit zunehmend Verbreitung gefunden haben und auch die Modellierung nichtlinearer Zusammen-

hänge erlauben. Für die vorliegende Arbeit wurden jedoch ausschließlich lineare Verfahren eingesetzt.

Das Ziel einer jeden Regressionstechnik ist die Bestimmung der Regressionskoeffizienten. Diese sollen so gewählt sein, daß mit Hilfe der Regressionsgleichung die abhängige Variable (y) für jedes Objekt des Datensatzes möglichst gut wiedergegeben werden kann.[26]

2.2.2.1 Lineare Regression

In der linearen Regression wird die abhängige Variable **y** des Datensatzes mit der unabhängigen Variable **x** verknüpft:

$$\mathbf{y} = b_0 + b_1 \cdot \mathbf{x} + \mathbf{e} \tag{2.5}$$

Der Regressionskoeffizient b_1 stellt also die Korrelation zwischen dem Deskriptor **x** und der Zielgröße **y** her. Der Achsenabschnitt b_0 beschreibt die Verschiebung dieses Zusammenhangs relativ zum Nullpunkt; bei zentrierten Daten ist $b_0 = 0$. Der Meßfehler der abhängigen Variable wird durch den Fehlervektor **e** beschrieben, während die unabhängige Variable in der QSAR als fehlerfrei angesehen wird.

2.2.2.2 Multiple lineare Regression (MLR)

Normalerweise bestehen die X-Daten in der QSAR aus mehr als einem Deskriptor. Dann wird angenommen, daß sich die Aktivität (**y**) additiv aus der Summe der Beiträge der einzelnen Deskriptoren zusammensetzt, wobei der Einfluß eines jeden Deskriptors durch seinen Regressionskoeffizienten quantifiziert wird. Die entsprechende mathematische Darstellung lautet

$$\mathbf{y} = b_0 + b_1 \cdot \mathbf{x_1} + b_2 \cdot \mathbf{x_2} + \cdots + b_p \cdot \mathbf{x_p} + \mathbf{e} \tag{2.6}$$

oder in Matrixschreibweise

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{e} \tag{2.7}$$

wobei $\mathbf{X}$ die Dimension $n \times p$ besitzt und die Spaltenvektoren der p Deskriptoren für die n Moleküle enthält sowie $\mathbf{b}$ von der Dimension $p \times 1$ ist und die unbekannten Regressionskoeffizienten enthält. Den unabhängigen Fehler der Objekte gibt wiederum $\mathbf{e}$ (Dimension $n \times 1$) an.

Der Regressionskoeffizientenvektor wird in der MLR dann als

$$\hat{\mathbf{b}} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} \tag{2.8}$$

geschätzt. Die Schätzung von $\hat{\mathbf{b}}$ wird dabei so vorgenommen, daß die Summe der quadrierten Abweichungen (RSS, engl. *Residual Sum of Squares*; siehe Gl. 2.16, Seite 21) zwischen den gemäß

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \hat{\mathbf{b}} \tag{2.9}$$

wiedergegebenen und den experimentell ermittelten Werten $\mathbf{y}$ möglichst gering ist.

Mit Gl. 2.9 ist nun auch die Vorhersage der Aktivität neuer Moleküle $\hat{\mathbf{y}}_{\text{neu}}$ aus deren Deskriptorwerten $\mathbf{X}_{\text{neu}}$ möglich.

Durch die Matrixinversion $\mathbf{X}^T\mathbf{X}$ kann Gl. 2.8 prinzipiell direkt gelöst werden. In einigen Fällen sollte jedoch diese Art der Schätzung des Regressionskoeffizientenvektors nicht angewendet werden: Problematisch sind etwa unterbestimmte Gleichungssysteme sowie Systeme, die Multikollinearitäten aufweisen.

Unterbestimmte Gleichungssysteme Für unterbestimmte Gleichungssysteme, sogenannte „fette“ Matrizen, für die gilt $n < p$, ist die Schätzung des Regressionskoeffizientenvektors $\hat{\mathbf{b}}$ nicht mehr eindeutig; es existieren unendlich viele gleichwertige Lösungen. Dieser Fall ist in der QSAR aufgrund der großen Vielfalt zur Verfügung stehender Deskriptoren und den oft nur kleinen Serien von Molekülen bekannter Aktivität recht häufig anzutreffen. „Schlanke“ Matrizen, also überbestimmte Gleichungssysteme mit $n > p$ liegen dagegen nur selten vor.

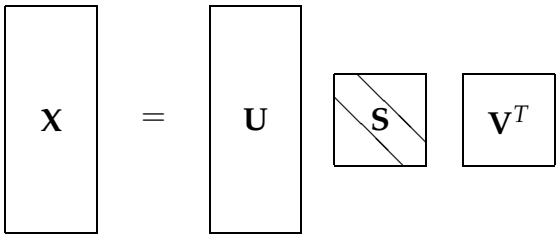
Multikollinearitäten Sind mehrere Spalten der Datenmatrix $\mathbf{X}$ linear voneinander abhängig, so liegt eine exakte Multikollinearität vor. Diese kann durch

Entfernen einer der beiden Spalten leicht behoben werden. Sind zwei Spalten dagegen nur näherungsweise linear voneinander abhängig, liegt eine sogenannte Nahezu-Kollinearität vor. Hier ist eine Behebung des Problems durch Eliminierung einer der Spalten offensichtlich nicht möglich. Die Regressionskoeffizienten können in diesem Fall zwar eindeutig bestimmt werden, doch die Varianz der einzelnen Koeffizienten ist möglicherweise inakzeptabel hoch. Eine mathematische Begründung dafür wird im folgenden aus Gl. 2.14 ersichtlich.

2.2.2.3 Singulärwertzerlegung (SVD)

Ein mathematisches Verfahren, das die Grundlage zur Lösung der genannten Problemfälle in der MLR schafft, ist die Singulärwertzerlegung (engl. *Singular value decomposition*, SVD).[27] Es zerlegt die Datenmatrix $\mathbf{X}$ in zwei orthonormale Matrizen $\mathbf{U}$ und $\mathbf{V}$ und eine Diagonalmatrix $\mathbf{S}$, die in der Hauptdiagonalen die Singulärwerte enthält. Die Anwendung der SVD auf die Datenmatrix $\mathbf{X}$ liefert

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \tag{2.10}$$



Hier entsprechen die Spalten der Matrix $\mathbf{U}$ den Eigenvektoren von $\mathbf{X}\mathbf{X}^T$, die Diagonalmatrix $\mathbf{S}$ enthält die Singulärwerte (positive Wurzeln der Eigenwerte) von $\mathbf{X}^T\mathbf{X}$, und die Spalten von $\mathbf{V}$ geben die Eigenvektoren der Matrix $\mathbf{X}^T\mathbf{X}$ an. Es gilt $\mathbf{U}^T\mathbf{U} = \mathbf{I_n}$ und $\mathbf{V}^T\mathbf{V} = \mathbf{I_p}$.

Hauptkomponentenanalyse Aus den mittels SVD erhaltenen Matrizen sind die wichtigsten Größen der Hauptkomponentenanalyse (engl. *Principal component analysis*, PCA) direkt zugänglich[28]:

Die Scores **T** berechnen sich als

$$\mathbf{T} = \mathbf{U} \cdot \mathbf{S} \qquad \text{bzw.} \qquad \mathbf{T} = \mathbf{X} \cdot \mathbf{V} \tag{2.11}$$

Die Loadings entsprechen den Eigenvektoren **V**, die quadrierten Diagonalelemente von **S** sind die Eigenwerte.

Berechnung der Regressionskoeffizienten Entsprechend Gl. 2.9 (Seite 12) und Gl. 2.10 ergibt sich nach Anwendung der SVD auf die Datenmatrix der Zusammenhang

$$\mathbf{y} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \cdot \hat{\mathbf{b}} \tag{2.12}$$

Unter der Berücksichtigung, daß $\mathbf{U} \cdot \mathbf{U}^T = \mathbf{I}$, erhält man als geschätzten Regressionskoeffizientenvektor

$$\hat{\mathbf{b}} = \mathbf{V} \cdot \mathbf{S}^{-1} \cdot \mathbf{U}^T \cdot \mathbf{y} \tag{2.13}$$

Aufgrund der notwendigen Matrixinversion $\mathbf{S}^{-1}$ darf **S** offensichtlich keine Singulärwerte enthalten, die gleich Null sind. Dieser Fall tritt genau dann ein, wenn zwei Spalten der Datenmatrix **X** linear abhängig voneinander sind.

Sehr kleine Singulärwerte sind aus einem weiteren Grund problematisch: Die Varianz eines jeden Regressionskoeffizienten $\hat{b}$ ist umgekehrt proportional zum zugehörigen Singulärwert[27]:

$$\mathrm{var}(\hat{b}_i) = \sum_{k=1}^{r} \frac{v_{i,k}^2}{s_{k,k}^2} \cdot \sigma^2 \tag{2.14}$$

Hier sind $v_{i,k}^2$ und $s_{k,k}^2$ die entsprechenden Elemente der **V**- bzw. **S**-Matrix; r ist der Rang der Matrix. Für kleine Singulärwerte im Nenner von Gl. 2.14 ergibt sich also eine hohe Varianz der geschätzten Regressionskoeffizienten. Da gleichzeitig die Größe eines Singulärwerts den Informationsgehalt der entsprechenden Variable in der Originaldatenmatrix widerspiegelt, ist es naheliegend, kleine Singulärwerte zu vernachlässigen. Genau diesen Ansatz verfolgt die im nächsten Abschnitt beschriebene Hauptkomponentenregression (engl. *Principal component regression*, PCR).

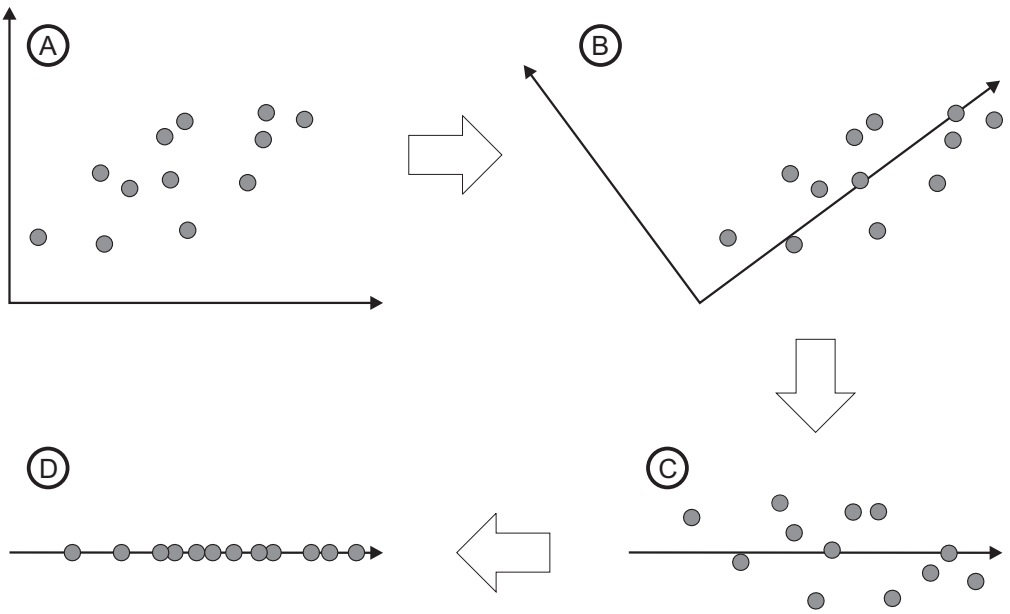


Abbildung 2.2 Geometrische Deutung der *Singular value decomposition* (SVD). Das Koordinatensystem der Originaldaten (A) wird so gedreht, daß die erste Hauptachse den größten Teil der Varianz der Daten erfaßt (B). Wird dieses neue Koordinatensystem auf nur eine Achse reduziert (C), geht nur ein sehr geringer Teil der Information verloren. Auch im neuen eindimensionalen Koordinatensystem (D) sind alle Objekte eindeutig voneinander unterscheidbar.

Geometrische Deutung der SVD Da die Spalten der Eigenvektormatrix $\mathbf{V}$ wie bereits erwähnt aus orthogonalen Einheitsvektoren besteht, können diese als Koordinatenachsen eines euklidischen Raums interpretiert werden. Die Koordinaten der Originaldaten in diesem neuen Koordinatensystem entsprechen den Scores $\mathbf{T}$, d. h. die Originaldaten werden gemäß $\mathbf{T} = \mathbf{X} \cdot \mathbf{V}$ in das neue System projiziert.

Im neuen Koordinatensystem erstreckt sich der Hauptanteil der Varianz der Daten entlang der ersten Koordinatenachse; die zweite Achse erfaßt den größten Teil der verbleibenden Varianz usw. Die SVD schafft also ein Koordinatensystem, das gegenüber dem Raum der Originaldaten so gedreht ist, daß die Streuung der Daten weitestgehend von einigen wenigen Koordinatenachsen erfaßt wird (siehe Abb. 2.2). Es bildet also die in den Daten enthaltene Information sehr effizient ab, da bereits mit wenigen Raumachsen ein Großteil der Information darstellbar ist.

2.2.2.4 Hauptkomponentenregression (PCR)

Für eine Datenmatrix $\mathbf{X}$ vom Rang r erhält man aus der SVD r Hauptkomponenten. Um die in Gl. 2.14 (Seite 14) beschriebene Varianz der Regressionskoeffizienten zu reduzieren, verwendet die Hauptkomponentenregression nur die ersten q Hauptkomponenten. Analog zu Gl. 2.13 (Seite 14) berechnet die PCR den Regressionskoeffizientenvektor also gemäß

$$\hat{\mathbf{b}}_{\mathrm{PCR}} = \mathbf{V}_q \cdot \mathbf{S}_q^{-1} \cdot \mathbf{U}_q^T \cdot \mathbf{y} \tag{2.15}$$

Die entscheidende Frage bei der PCR ist die nach der Anzahl q der verwendeten Hauptkomponenten. Einerseits soll versucht werden, eine möglichst hohe Varianzreduktion von $\hat{\mathbf{b}}$ zu erzielen, andererseits führt die Vernachlässigung der Information der übrigen $r - q$ Hauptkomponenten zu einer Verzerrung (engl. *Bias*) der Ergebnisse. Dieser *Bias-variance tradeoff* wird in der Praxis meist mittels Kreuzvalidierung optimiert (siehe 2.2.3.1, Seite 18).

2.2.2.5 Partial Least Squares Regression (PLS)

Ein der PCR verwandtes Verfahren ist die *Partial least squares regression* (PLS).[29] Wie die Hauptkomponentenanalyse projiziert auch PLS die Daten in ein neues Koordinatensystem. Für dessen Bestimmung werden jedoch nicht nur die X-Daten, sondern auch die Y-Daten genutzt; durch eine leichte Rotation der Eigenvektoren wird die Korrelation der Komponenten (latenten Variablen) mit $\mathbf{y}$ maximiert. Als gemeinsamer PLS-Vektor wird gewissermaßen der optimale Kompromiß zwischen dem X- und dem Y-Eigenvektor gesucht, indem die Regressionskoeffizienten unter Berücksichtigung der Kovarianz von $\mathbf{X}$ und $\mathbf{y}$ optimiert werden.

Verglichen mit PCA liefert PLS häufig weniger komplexe Modelle mit einer geringeren Zahl von Komponenten. Die Anzahl der tatsächlich verwendeten Freiheitsgrade ist jedoch im allgemeinen für beide Methoden nahezu identisch. Demnach ist in der QSAR-Analyse PLS nicht grundsätzlich besser geeignet als PCR.[30]

2.2.3 Modellvalidierung

Das Ziel der QSAR-Analyse ist üblicherweise die Vorhersage der biologischen Aktivität neuer Substanzen. Da die Deskriptoren im Computer anhand „virtueller“ Moleküle berechnet werden können, ist die Vorhersage der Aktivität von Verbindungen möglich, die noch gar nicht im Labor synthetisiert wurden — ein immens wichtiges Mittel zur Effizienzsteigerung und Kostenreduktion in der Wirkstoffentwicklung.

Die Qualität und Verläßlichkeit eines QSAR-Modells kann jedoch prinzipiell nur durch den Vergleich von Vorhersagewerten und tatsächlich experimentell bestimmten Aktivitätsdaten ermittelt werden. Um ein aussagekräftiges Maß der Vorhersagekraft des Modells zu erhalten, müssen zur Validierung zudem unabhängige Daten (sogenannte Testdaten) verwendet werden, die nicht bereits in der Modellbildung zum Einsatz kamen. Da jedoch meist insgesamt nur eine geringe Anzahl von Verbindungen mit experimentell bestimmten Aktivitäten zur Verfügung steht, ist bei der Modellerstellung der Verzicht auf einen für die Validierung nutzbaren Subdatensatz oft nicht akzeptabel. Die Datenbasis für die Erstellung des Modells (der sogenannte Trainingsdatensatz) wäre sonst so klein, daß das Modell selbst instabil würde, also nur mit einer hohen Variabilität geschätzt werden könnte. Um diesem Problem entgegenzuwirken, wird häufig die im folgenden beschriebene Methode der wiederholten Stichprobenziehung eingesetzt, die als Kreuzvalidierung (engl. *Cross-validation*, CV) bekannt ist.[31]

Die Bestimmung der Vorhersagekraft eines QSAR-Modells dient nicht in erster Linie der Generierung eines Qualitätsmaßes an sich. Vielmehr wird sie zur Auswahl wichtiger Modellparameter eingesetzt. So kann etwa die geeignete Anzahl q der für die Modellbildung verwendeten Hauptkomponenten nur dadurch ermittelt werden, daß Modelle für verschiedene Werte von q erstellt und miteinander verglichen werden. Der optimale Wert für q ist dann derjenige, der zum Modell mit der besten Vorhersagekraft führt. Solche datengestützten Entscheidungen werden also anhand einer internen Validierung getroffen, bei der alle Daten in die Modellselektion eingehen. Gleiches gilt für die Methode der Variablenselektion (siehe 2.2.6, Seite 29), bei der nur ein Teil der zur Verfügung stehenden Deskriptoren für die Modellgenerierung genutzt wird; auch hier muß die optimale Zusammenstellung der

schließlich verwendeten Variablen durch Validierung vieler einzelner Modelle ermittelt werden, die aus unterschiedlichen Untermengen von Variablen hervorgegangen sind.

2.2.3.1 Kreuzvalidierung

Bei der Kreuzvalidierung wird dem Datensatz zunächst eine gewisse Anzahl von Objekten entnommen. Die Modellbildung erfolgt dann ausschließlich anhand der verbleibenden Objekte. Mit dem so erstellten Modell wird die Zielgröße der zuvor ausgelassenen Objekte vorhergesagt und mit dem wahren Wert verglichen. Als Gütefunktion dient hierbei meist die Fehlerquadratsumme (siehe Gl. 2.16, Seite 21). Der gesamte Vorgang wird mehrmals wiederholt, wobei im ersten Schritt immer wieder andere Objekte ausgelassen werden. Entsprechend der genauen Vorgehensweise werden verschiedene Arten der Kreuzvalidierung unterschieden:

Leave-one-out-Kreuzvalidierung (LOO-CV) Bei der Leave-one-out-Kreuzvalidierung wird dem Datensatz von n Objekten genau ein Objekt entnommen. Die Modellbildung wird mit den übrigen $n - 1$ Datenpunkten („Konstruktionsdaten“) vorgenommen und schließlich die Zielgröße des einen ausgelassenen Objekts („Validierdaten“) vorhergesagt. Die Prozedur wird n mal so wiederholt, daß schließlich jedes Objekt genau einmal ausgelassen und seine Zielgröße vorhergesagt wurde. Es werden also insgesamt n Modelle erstellt. Eine schematische Darstellung der Kreuzvalidierung zeigt Abb. 2.3. In dieser Arbeit wird ein Bezug zur LOO-CV durch das Subscript „CV-1“ gekennzeichnet.

ν-fache Kreuzvalidierung (ν-CV) Bei der ν-fachen Kreuzvalidierung (engl. *ν-fold Cross-validation*) wird der Datensatz zufällig in ν möglichst gleich große Gruppen unterteilt. Die Modellerstellung erfolgt anhand von $\nu - 1$ Gruppen des Datensatzes, während zur Validierung die verbleibende ν-te Gruppe verwendet wird. Somit werden bei der ν-CV genau ν Wiederholungen von Modellbildung und -validierung durchgeführt, jede der ν Gruppen wird genau einmal ausgelassen und die Zielgrößen der enthaltenen Objekte vorhergesagt. Bei kleinen Werten von ν reduziert sich also die Anzahl der generierten

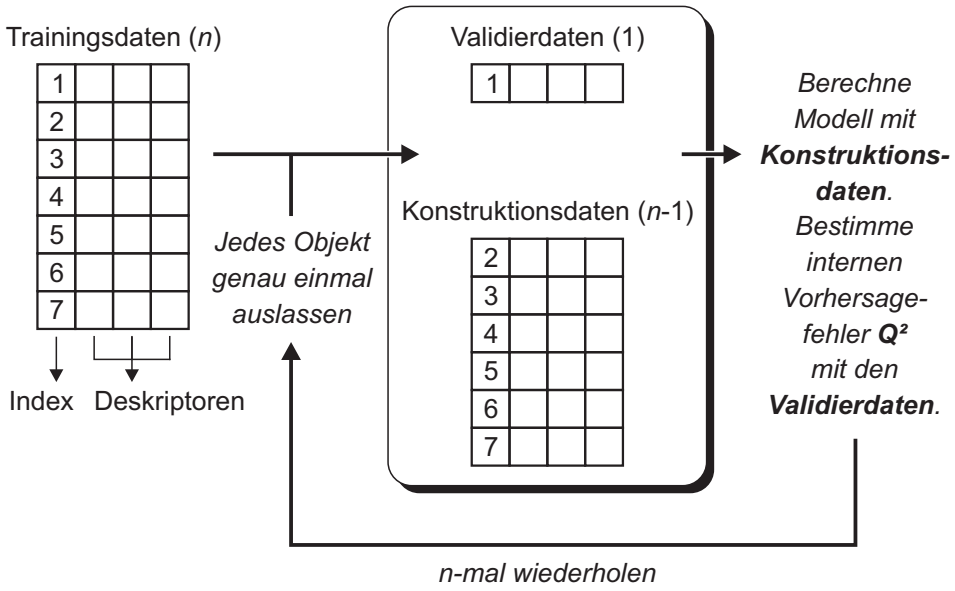


Abbildung 2.3 Schematische Darstellung der Leave-one-out-Kreuzvalidierung (LOO-CV). Jedes der n Objekte des Trainingsdatensatzes wird genau einmal ausgelassen.

Modelle und damit auch der Rechenaufwand, während der Grenzfall $v = n$ der Leave-one-out-Kreuzvalidierung entspricht.

Leave-multiple-out-Kreuzvalidierung (LMO-CV) Wird nicht nur ein einzelnes Objekt, sondern eine Anzahl $d > 1$ von Objekten ausgelassen, so spricht man von einer Leave-multiple-out-Kreuzvalidierung. Die Modellbildung wird also mit den $n - d$ Objekten des Konstruktionsdatensatzes vorgenommen, die Vorhersage erfolgt für die d ausgelassenen Objekte.[32] Auch diese Prozedur wird mehrmals wiederholt, wobei die d auszulassenden Objekte jeweils erneut durch eine Zufallsauswahl bestimmt werden. Aufgrund dieser zufälligen Auswahl ist eine hohe Zahl von Wiederholungen B notwendig; als robust haben sich Werte zwischen $B = 2 \cdot n$ und $B = 3 \cdot n$ erwiesen.[33] Die Vorgehensweise der Leave-multiple-out-Kreuzvalidierung ist schematisch in Abb. 2.4 dargestellt. In dieser Arbeit wird ein Bezug zur LMO-CV durch das Subscript „CV-d“ gekennzeichnet; in den meisten Fällen wurde die Hälfte der Objekte ausgelassen, also $d = n/2$ (50 %) gewählt.

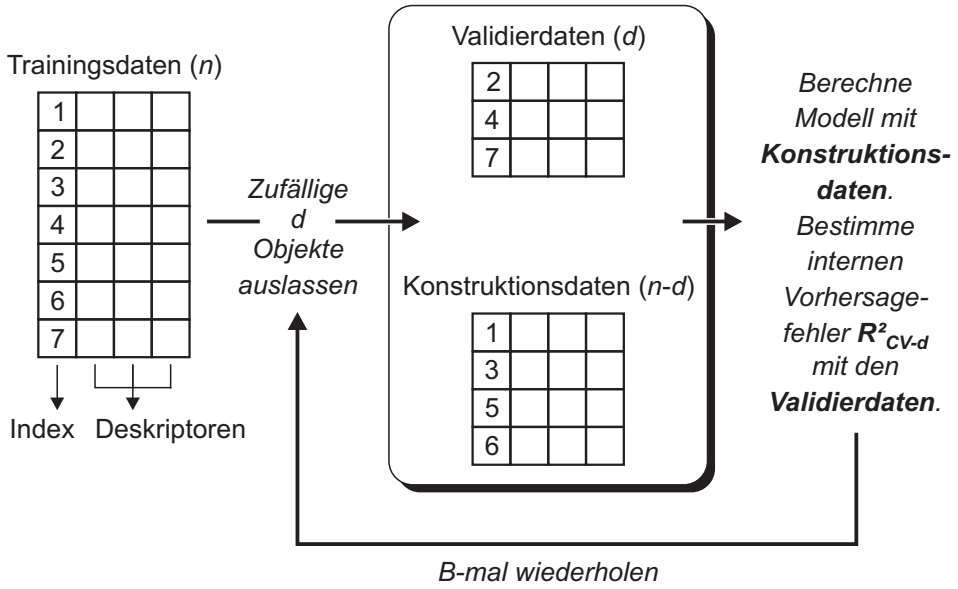


Abbildung 2.4 Schematische Darstellung der Leave-multiple-out-Kreuzvalidierung (LMO-CV). Von den n Trainingsdaten werden zufällige d Objekte ausgelassen; mit den übrigen n-d wird das Modell erstellt. Die gesamte Prozedur wird B-mal wiederholt, wobei üblicherweise $B = 3 \cdot n$ gewählt wird.

2.2.3.2 Gütekriterien

Bei der Validierung eines Modells muß zwischen interner Validierung (z. B. Kreuzvalidierung) und externer Validierung (Vorhersage separater Testdaten) unterschieden werden. Auch wenn die interne Validierung Vorteile hinsichtlich der Effizienz der Datennutzung bietet, so trifft sie dennoch tendenziell zu optimistische Aussagen bezüglich der Modellqualität. Eine wirklich verläßliche Modellvalidierung ist nur durch die Vorhersage der Zielgrößen eines externen Testdatensatzes möglich, der nicht bereits für die Erstellung und Optimierung des Modells genutzt wurde.

Aus diesem Grund existieren für die unterschiedlichen Formen der Validierung unterschiedliche Gütekriterien. Diese sind zwar eng miteinander verwandt oder sogar identisch, tragen aber häufig gesonderte Bezeichnungen, um ihren Bezug deutlich zu machen.

Da ein QSAR-Modell zur Vorhersage von biologischen Aktivitäten dient, ist es naheliegend, die Genauigkeit dieser Vorhersage als Gütekriterium zu

verwenden. Der in diesem Zusammenhang intuitivste Wert ist die Differenz zwischen dem Vorhersagewert und dem wahren Wert. Die Summe über alle quadrierten Differenzen wird als Fehlerquadratsumme oder Summe der Abweichungsquadrate (engl. *Residual sum of squares, RSS*) bezeichnet.

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.16}$$

Beziehen sich die vorhergesagten Werte auf einen externen Testdatensatz, so spricht man von der *Predictive residual sum of squares* (PRESS).

$$\mathrm{PRESS} = \mathrm{RSS}_{\mathrm{Test}} = \sum_{i=1}^{n} (y_{i,\,\mathrm{Test}} - \hat{y}_{i,\,\mathrm{Test}})^2 \tag{2.17}$$

Der quadrierte Korrelationskoeffizient R^2 (auch als Bestimmtheitsmaß bezeichnet) beschreibt, zu welchem Anteil das Modell die Varianz der abhängigen Variable erklären kann. Es gilt $0 < R^2 < 1$ und

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{SYY}} \tag{2.18}$$

mit

$$\mathrm{SYY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.19}$$

Der Korrelationskoeffizient der Leave-one-out-Kreuzvalidierung trägt die gesonderte Bezeichnung Q^2 oder q^2.

$$Q^2 = R^2_{\text{CV-1}} = 1 - \frac{\mathrm{RSS}_{\text{CV-1}}}{\mathrm{SYY}} = 1 - \frac{\mathrm{PRESS}}{\mathrm{SYY}} \tag{2.20}$$

Der mittlere quadrierte Fehler der Vorhersage (engl. *Mean squared error of prediction, MSEP*) wird sowohl in der Kreuzvalidierung als auch bei der

externen Testdatenvorhersage verwendet.

$$\text{MSEP} = \frac{1}{n} \cdot \text{PRESS} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.21}$$

Üblicherweise wird aber die Quadratwurzel dieses Fehlers angegeben, die RMSEP (engl. *Root mean squared error of prediction*) genannt wird.

$$\text{RMSEP} = \sqrt{\text{MSEP}} \quad = \quad \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2.22}$$

Je nach Bezug wird diese Größe beispielsweise als $\text{RMSEP}_{\text{CV-1}}$ (LOO-CV) oder $\text{RMSEP}_{\text{Test}}$ (externe Testdatenvorhersage) bezeichnet.

2.2.4 Identifizierung von Prediction Outliern

Ebenso wie auf Grundlage der Wettervorhersage wichtige Entscheidungen etwa in der Landwirtschaft, der Luftfahrt oder dem Alpinismus getroffen werden, liefert die QSAR-gestützte Vorhersage von Moleküleigenschaften die Argumente für wichtige Entscheidungen in der Wirkstoffentwicklung. So legt beispielsweise die vorhergesagte schlechte Wasserlöslichkeit einer Substanz oftmals den Entschluß nahe, diese Verbindung erst gar nicht zu synthetisieren und zu testen. Andererseits kann eine fälschlicherweise als gut löslich vorhergesagte Verbindung unnötige Kosten verursachen und damit zu — im jeweiligen Kontext — ähnlich negativen Konsequenzen führen wie ein nicht vorhergesagtes Unwetter.

Ebenso wie ein mit Daten aus Berlin kalibriertes Wettermodell zwar noch in Frankfurt anwendbar ist, für Shanghai aber sicherlich einen immens hohen Vorhersagefehler aufweisen wird, hat auch ein QSAR-Modell keine Gültigkeit für das gesamte chemische Universum. Auch ein QSAR-Modell kann nur in sehr beschränktem Maße extrapolieren: Verläßt man mit den Vorhersagen den chemischen Raum der zur Kalibrierung verwendeten Verbindungen, so wächst die Gefahr eines übergroßen Vorhersagefehlers. Der QSAR-Anwender muß sich also bewußt sein, daß die Aktivitäten mancher Verbindungen

schlicht nicht vorhersagbar sind, daß der vom Modell für eine solche Verbindung gelieferte Vorhersagewert nicht mehr wert ist als eine Zufallszahl.

Nur durch die Identifizierung der jenseits der Modellgrenzen, also außerhalb des Kalibrierungsraums liegenden Verbindungen kann die Gefahr gebannt werden, Entscheidungen auf prinzipiell nicht verläßliche Vorhersagewerte zu stützen. Eine Auswahl von bekannten und neuen Methoden zur Identifizierung solcher *Prediction Outlier* werden im folgenden vorgestellt.

Inlier Einen speziellen Fall stellen die sogenannten *Inlier* dar: Auch für Objekte, die im Datenraum innerhalb der Grenzen liegen, die durch die äußersten Kalibrierobjekte beschrieben werden, sind unter bestimmten Umständen keine zuverlässigen Vorhersagen möglich. Ein solcher Fall tritt auf, wenn innerhalb des Kalibrierdatenraums „leere Blasen“ existieren — gewissermaßen Enklaven in einem ansonsten gut durch Trainingsobjekte beschriebenen Raum, die ihrerseits nicht durch die Anwesenheit von Trainingsdaten charakterisiert sind. Auch innerhalb dieser „leeren“ Regionen im Datenraum ist die Gültigkeit des Modells ungewiß und daher eine vertrauenswürdige Vorhersage nicht möglich.

2.2.4.1 Mahalanobis-Distanz

Eine intuitives Maß zur Identifizierung von Outliern ist der Abstand eines Objekts zum Zentroiden des Datensatzes. Diese Distanz zeigt, wie weit das Objekt vom gedachten Schwerpunkt der Daten entfernt liegt, also wie stark es sich vom „durchschnittlichen“ Objekt unterscheidet.

Zur Bestimmung dieses Abstands wird nicht die gewohnte euklidische Distanz verwendet, sondern die Mahalanobis-Distanz. Dieses Verfahren ist eine der am häufigsten angewendeten Techniken zur Identifizierung von Outliern.[34] Eine detaillierte Beschreibung der Mahalanobis-Distanz folgt in 2.3.2 (Seite 34) im allgemeinen Abschnitt über Distanz- und Ähnlichkeitsmaße.

Die Frage, ab welcher Distanz eine Objekt „zu weit“ entfernt vom Datenschwerpunkt liegt, kann nur durch die Einführung eines Schwellenwerts beantwortet werden, der eben diese Grenze markiert. Da die Mahalanobis-Distanzen von p-dimensionalen Vektoren aus einer normalverteilten Grund-

gesamtheit einer χ^2-Verteilung mit $p-1$ Freiheitsgraden folgen, können die Quantile dieser Verteilung als Schwellenwerte herangezogen werden: Ist die Mahalanobis-Distanz eines Objekts größer als der entsprechende Wert der χ^2-Verteilung, so gilt das Objekt als Outlier.

2.2.4.2 Leverages

Die sogenannten Leverages stellen eine weitere etablierte Methode zur Identifizierung von Prediction Outliern dar. Ihre separate Nennung ist im Grunde nur unter historischen Gesichtspunkten gerechtfertigt, da sie mit der Mahalanobis-Distanz sehr eng verwandt sind — eine Tatsache, die jedoch lange Zeit nicht bekannt war und deshalb in der Vergangenheit zur getrennten Behandlung beider Methoden führte.

Die Leverages eines Datensatzes können aus der Projektions-Matrix $\mathbf{H}$ (auch Hat-Matrix genannt) berechnet werden[35], die als

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \tag{2.23}$$

definiert ist. Für den Leverage h_i eines Objekts $\mathbf{x_i}$ gilt

$$h_i = \mathbf{x_i} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x_i}^T \tag{2.24}$$

Die Bezeichnung „Hebel“ (engl. *Leverage*) folgt der Sichtweise, daß h_i beschreibt, in welchem Ausmaß sich der Vorhersagewert $\hat{y}_i$ bei einer Änderung von y_i ändert, also mit welchen Hebel y_i auf $\hat{y}_i$ wirkt.[36]

Die enge Verwandtschaft zwischen dem Leverage und der Mahalanobis-Distanz MD eines Objekts wird anhand der folgenden Gleichung deutlich:

$$h_i = \frac{1}{n-1} \cdot \mathrm{MD}^2(i) + \frac{1}{n} \tag{2.25}$$

Als Schwellenwert zur Beurteilung des Leverages wird meist der Wert $2p/n$ verwendet:* Besitzt ein Objekt einen Leverage $h_i > 2p/n$, so gilt es als Outlier.[37]

*Auch der Schwellenwert $3p/n$ ist gebräuchlich.

2.2.4.3 Applicability Domain

Zusätzlich zu dem in der Chemometrie schon seit Jahrzehnten gebräuchlichen Begriff der Prediction Outlier[34,36,38–40] wurde in jüngerer Zeit in Arbeiten auf dem Gebiet der QSAR eine weitere Definition eingeführt. Mit der Bezeichnung *Applicability domain* wird hier derjenige Bereich des Datenraums beschrieben, innerhalb dessen das Modell zu Vorhersagen in der Lage ist[41,42]. Mit anderen Worten sind also alle Testobjekte, die außerhalb der Applicability domain liegen, als Outlier anzusehen.

Leider wird in den entsprechenden Publikationen nicht auf die viel ältere Definition der Prediction Outlier verwiesen, obwohl der neue Begriff keinerlei Erweiterung dieses in der Chemometrie bewährten Konzepts bietet. „Prediction outlier“ und „Applicability domain“ beschreiben also ein und dasselbe Phänomen lediglich aus verschiedenen Blickwinkeln. Der Applicability domain entsprechende oder ähnliche Ansätze finden sich in der chemometrischen Literatur unter der Bezeichnung „konvexe Hülle“[40,43] bzw. „*Effective Prediction Domain*“[44].

2.2.4.4 *k*-Nächster-Nachbar-Methode

Die im Rahmen dieser Arbeit entwickelte Methode zur Outlier-Identifizierung bedient sich des Konzepts der *k*-Nächster-Nachbar-Distanz (engl. *k-Nearest neighbour*, kNN)[45]; eine detaillierte Erläuterung der Vorgehensweise folgt im Ergebnisteil (3.1.2). An dieser Stelle sei aber darauf verwiesen, daß davon unabhängig von TROPSHA ein ähnlicher Ansatz publiziert wurde, der jedoch nur ansatzweise und ausschließlich zur Verwendung in sogenannten kNN-QSPR-Modellen beschrieben wird[41,46]. Dabei handelt es sich um einen speziellen Fall der *Quantitative structure property relationships*, bei dem keine Datenmodellierung etwa mit PCR oder PLS (siehe 2.2.2.5, Seite 16) vorgenommen wird, sondern ein Nächster-Nachbar-Ansatz verfolgt wird. Stark vereinfacht verwendet diese Idee als Vorhersagewert der Zielgröße eines Testobjekts den bekannten y-Wert des nächstgelegenen Trainingsobjekts.

Ein ausführlicher Vergleich sowie eine Abgrenzung der eigenen Arbeiten zu der genannten Methode erfolgt ebenfalls im Ergebnisteil (3.1.2, Seite 60 ff).

2.2.5 Ensemble-Techniken

Durch die Anwendung von Ensemble-Techniken kann die Vorhersagegenauigkeit von QSAR-Modellen signifikant verbessert werden. Dazu wird nicht nur ein einziges Modell erstellt, sondern ein ganzes Ensemble von Einzelmodellen, die parallel für die Vorhersage verwendet werden.

Zur Generierung dieser einzelnen Modelle, aus denen das Ensemble besteht, wird zunächst der Trainingsdatensatz in einer bestimmten Art und Weise manipuliert. Ziel ist es, einen neuen, leicht veränderten Datensatz zu erhalten, der aber weiterhin die Charakteristik der ursprünglichen Daten besitzt. Dazu kann beispielsweise die abhängige Größe mit einem zusätzlichen Signalrauschen versehen werden, das dem Zufallsfehler der experimentellen Messung dieser Größe entspricht. Diese Manipulation (Störung, engl. *Perturbation*) der Ursprungsdaten wird mehrmals durchgeführt, so daß man ein Ensemble von leicht veränderten und untereinander verschiedenen Trainingsdatensätzen erhält. Aus diesen Daten wird dann jeweils ein QSAR-Modell erstellt, was zu einem Ensemble von Modellen führt.

Jedes der so erzeugten k Modelle des Ensembles wird nun zur Vorhersage benutzt, so daß schließlich für jede vorherzusagende Zielgröße $\hat{y}_i$ eine Anzahl k an Ensemble-Vorhersagewerten $\hat{y}_{\mathrm{ens,1}}$, $\hat{y}_{\mathrm{ens,2}}$, $\ldots$, $\hat{y}_{\mathrm{ens,k}}$ existiert. Als endgültiger Vorhersagewert wird schließlich der Mittelwert über alle Ensemble-Vorhersagen berechnet (siehe auch Abb. 2.5):

$$\hat{\mathbf{y}}_{\mathrm{ens}} = \frac{1}{k} \cdot \sum_{i=1}^{k} \hat{y}_{\mathrm{ens,i}} \tag{2.26}$$

Die den Ensemble-Modellen zugrundeliegenden manipulierten Datensätze können mit unterschiedlichen Methoden erzeugt werden, die die Daten auf verschiedene Art und Weise stören. Die in dieser Arbeit verwendeten Störungsmethoden werden im folgenden kurz erläutert; eine weitergehende Übersicht findet sich beispielsweise bei DIETTERICH[47].

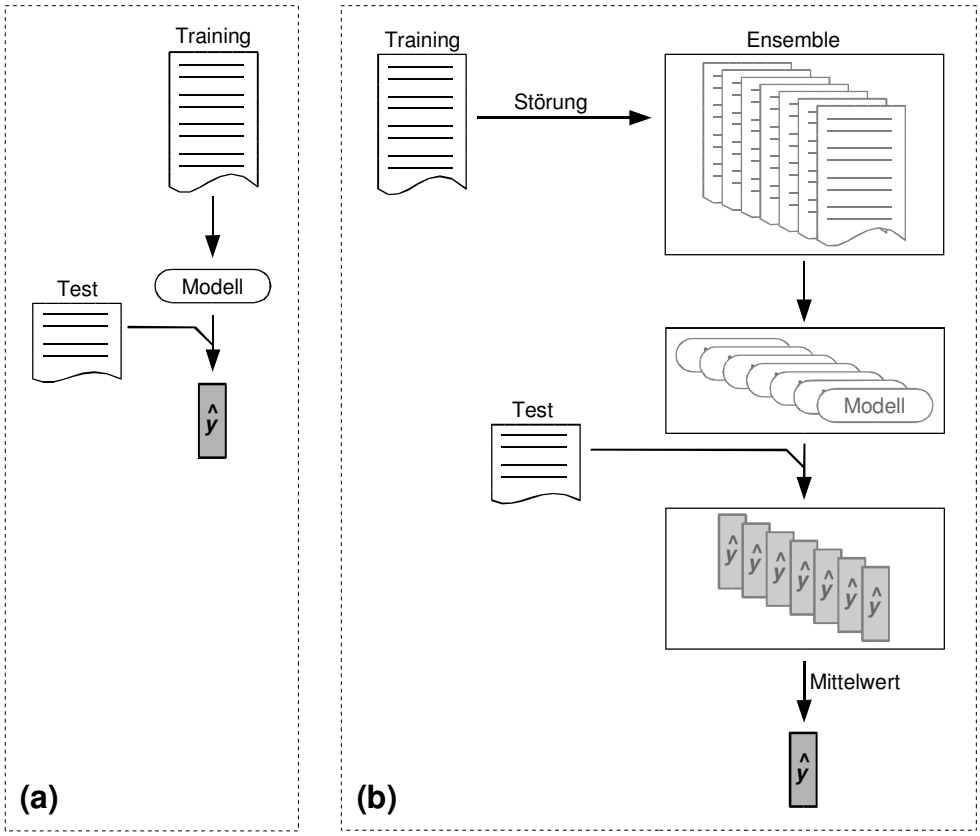


Abbildung 2.5 Vorhersage der $\hat{y}$-Werte des Testdatensatzes (a) mit einem Einzelmodell und (b) mit einem Ensemble.

2.2.5.1 Noise Addition

Bei der bereits einführend beschriebenen *Noise addition*[48] handelt es sich um eine Störungsfunktion im eigentlichen Sinne. Hier wird der **y**-Vektor des Datensatzes durch die Addition von zufälligem, normalverteiltem Rauschen beeinflußt. Die **X**-Matrix bleibt dagegen unverändert. Diese Vorgehensweise entspricht also einer ungenauen, stark verrauschten Messung der abhängigen Variable. Die Stärke des addierten Rauschens wird üblicherweise festgelegt auf 60 Prozent des Vorhersagefehlers des unverrauschten Modells.[48]

$$\mathbf{X}_{\text{noise}} = \mathbf{X} \tag{2.27}$$

$$y_{i,\text{noise}} = y_i + (\text{randn}[0,1] \cdot 0.6 \cdot \text{RMSEP}_{\text{original}}) \tag{2.28}$$

Hier bezeichnet randn$[0,1]$ eine Funktion, die als Ergebnis eine normalverteilte Zufallszahl mit dem Mittelwert 0 und der Standardabweichung 1 liefert.

2.2.5.2 Konvexe Pseudodaten

Eine von BREIMAN 1998 vorgestellte Methode ist die Erzeugung konvexer Pseudodaten.[49] Hierbei werden nicht wie bei der Noise addition die Originalobjekte selbst manipuliert. Vielmehr werden aus dem gegebenen Datensatz neue Objekte gewissermaßen interpoliert, indem zwei bestehende Datenpunkte miteinander kombiniert werden.

Der Algorithmus wählt aus dem ursprünglichen Trainingsdatensatz zunächst zwei zufällige Objekte $\mathbf{x_1}$ und $\mathbf{x_2}$ sowie die zugehörigen abhängigen Variablen y_1 und y_2 aus. Dann wird eine Zufallszahl v bestimmt, die im Intervall $[0,d]$ mit $d = [0,1]$ liegt; d ist der einzige Parameter der Methode und wurde in der vorliegenden Arbeit auf $d = 0.5$ gesetzt.* Außerdem wird $u = 1 - v$ berechnet.

Schließlich wird ein neues Objekt $\mathbf{x}'$ (mit entsprechendem y') gemäß

$$x' = u \cdot \mathbf{x_1} + v \cdot \mathbf{x_2} \tag{2.29}$$

$$y' = u \cdot y_1 + v \cdot y_2 \tag{2.30}$$

erzeugt, also ein zufällig gewichteter Mittelwert der beiden originalen Objekte $\mathbf{x_1}$ und $\mathbf{x_2}$. In dieser Arbeit wurden mit der Methode der konvexen Pseudodaten stets n neue Objekte $(\mathbf{x_i}, y_i)$ erzeugt und die n Objekte des ursprünglichen Datensatzes verworfen.

2.2.5.3 Bagging

Das *Bagging*[50] ist eine Bootstrap-Methode, also eine wiederholte Stichprobennahme mit Zurücklegen. Aus dem Datensatz wird n Mal ein Objekt $\mathbf{x}$ (und seine zugehörige abhängige Variable y) gezogen, dem neuen Datensatz hinzugefügt und wieder zurückgelegt.

*Der Parameters d hat für die hier gezeigte Anwendung nur eine untergeordnete Bedeutung und wurde anhand einer kurzen empirischen Untersuchung (hier nicht gezeigt) auf den genannten Wert festgelegt.

Aufgrund des Zurücklegens besteht die Möglichkeit, daß einige Objekte mehrmals gezogen werden. Im neuen Datensatz kann also ein Objekt auch mehrfach enthalten sein. Es läßt sich theoretisch zeigen, daß der neue Datensatz etwa 63 Prozent Unikate enthält, d. h. etwa 27 Prozent der ursprünglichen Daten gelangen nicht in den neuen Datensatz.

Die Objekte selbst werden also — im Gegensatz zu den konvexen Pseudodaten und der Noise addition — beim Bagging nicht verändert; die Störung des Datensatzes besteht ausschließlich in dessen veränderter Zusammenstellung.

2.2.5.4 Subdatensatzauswahl

Die Auswahl eines Subdatensatzes (engl. *Data subsetting*) unterscheidet sich von den übrigen hier vorgestellten Methoden durch die Größe des erzeugten Datensatzes. Der neue (Sub-)Datensatz besteht nämlich aus $f \cdot n$ Objekten, die den ursprünglichen Daten zufällig entnommen werden. In dieser Arbeit wurde stets $f = 0.66$ gewählt, der neue Datensatz besteht also aus zufälligen zwei Dritteln der Originaldaten. Auch hier bleiben wie beim Bagging die Objekte selbst unverändert.

2.2.6 Variablenselektion

Moderne QSAR-Verfahren beschreiben ein Molekül anhand von Hunderten, wenn nicht gar — etwa im Fall von 3D-QSAR wie CoMFA oder CoMSIA — Tausenden von Deskriptoren. Trotz einer möglichen Datenreduktion durch Anwendung von Methoden wie der Hauptkomponentenanalyse stellt sich unweigerlich die Frage, welche dieser Variablen tatsächlich für die konkrete Fragestellung wichtig sind. Denn bei einer solch großen Menge von Variablen ist davon auszugehen, daß einige ein hohes Maß an Information tragen, die mit der Zielgröße korreliert, während andere Variablen nur einen sehr geringen oder keinen entsprechenden Beitrag leisten.

Die Kunst der Variablenselektion besteht also darin, aus den vorhandenen p Variablen $x_1, x_2, \ldots, x_p$ diejenige Submenge auszuwählen, die nur die tatsächlich relevanten Variablen enthält. Da für diese Auswahl 2^p mögliche Submengen existieren, ist die Evaluierung aller daraus resultierender QSAR-Mo-

delle für übliche Datensätze nicht praktikabel. Andererseits erwächst aus der großen Anzahl möglicher Modelle die Gefahr von Zufallskorrelationen[51]. In diesem Fall kann die Zielgröße mathematisch gut modelliert werden, obwohl die ausgewählten Variablen keine sinnvolle Beschreibung des Systems darstellen; die Korrelation ist also rein zufälliger Natur.

Ein oft zitiertes Beispiel für eine Zufallskorrelation ist die gute Korrelation von Geburtenrate und Storchenpopulation in Schleswig-Holstein: Obwohl der verwendete Deskriptor (Storchenpopulation) offensichtlich in keinem kausalen Zusammenhang mit der Zielgröße (Geburtenrate) steht, liegt mathematisch eine hohe Korrelation vor. Das Problem geht so weit, daß die Kalibrierung eines QSAR-Modells mit einer ausreichend großen Anzahl von Zufallszahlen die annähernd exakte Wiedergabe der Zielgröße ermöglicht.

Für die Variablenselektion müssen also zwei Probleme gelöst werden: Einerseits die effiziente Suche nach der „richtigen" Submenge von Variablen, andererseits die besonders strenge Modellvalidierung zur Vermeidung von Zufallskorrelationen. Dabei kommt dem zweiten Punkt die größte Bedeutung zu, da erst die Validierung eine Information darüber liefert, ob die vom Suchalgorithmus vorgeschlagene Submenge von Variablen tatsächlich mit der Zielgröße korreliert. Mit anderen Worten bringt auch der schnellste Suchalgorithmus keinen Vorteil, wenn die Beurteilung der selektierten Variablen mit Hilfe der sogenannten Gütefunktion (engl. *Objective function*) nicht zuverlässig ist.

Unter mathematischen Gesichtspunkten gilt es demnach, eine Variablen-Submenge α zu finden, die die Gütefunktion optimiert. Dabei sei $\alpha \in \mathcal{A}$ und $\mathcal{A}$ die Menge aller 2^p möglichen Lösungen. α bestehe aus einer definierten Anzahl von 0 bis p Variablen, wobei die zur Verfügung stehenden Variablen als $\{1, \ldots, p\}$ definiert seien. Wird als Gütefunktion der Vorhersagefehler $PE(\alpha)$ des Modells verwendet, so lautet das Optimierungsproblem demnach

$$\text{minimiere } PE(\alpha) : \alpha \in \mathcal{A} \tag{2.31}$$

2.2.6.1 Suchalgorithmus

Als Suchalgorithmus zur Auswahl der Variablen-Submenge wurde in der vorliegenden Arbeit die sogenannte Tabu-Suche (TS) verwendet, bei der es

sich um eine schrittweise Methode zur Variablenselektion handelt. In jeder Iteration werden ausgehend von der aktuellen Lösung zunächst alle Nachbarschaftslösungen erzeugt, also diejenigen Lösungen, die sich nur durch den Status (im Modell enthalten oder nicht enthalten) einer einzigen Variable unterscheiden. Auf Grundlage dieser Menge wird derjenige Schritt ausgeführt, der ΔPE gemäß

$$\Delta PE = PE(\alpha_{\text{Nachbar}}) - PE(\alpha) \tag{2.32}$$

minimiert, d. h. die stärkste Verbesserung (engl. *Steepest descent*) der Modellgüte hervorruft. Ist dagegen nur ein Selektionsschritt möglich, der die Modellqualität verschlechtert, so wird derjenige gewählt, der zur geringsten Verschlechterung führt (engl. *Mildest ascent*). Der Startpunkt des Algorithmus ist der Zustand, in dem alle Variablen auf „nicht enthalten“ gesetzt sind.

Durch die Zulässigkeit von Verschlechterungen des Modells kann der Suchalgorithmus nicht in lokalen Minima steckenbleiben, er besitzt also prinzipiell die Möglichkeit zum Auffinden des globalen Minimums. Der Name Tabu-Suche verweist auf die Bedingung, daß eine bereits zuvor ausgewählte Lösung in einem späteren Schritt nicht noch einmal ausgewählt werden darf. Der einzige vom Anwender zu definierende Parameter ist das Abbruchkriterium; in den in dieser Arbeit durchgeführten Variablenselektionen wurden (vorangehenden Untersuchungen entsprechend) stets $3 \cdot p$ Iterationen durchlaufen.

2.2.6.2 Gütefunktion

Wie bereits erwähnt kann die Steuerung des Suchalgorithmus, d. h. die Bewertung der möglichen Lösungen, nur durch die Gütefunktion erfolgen. Die Gütefunktion ist also die wichtigste Komponente der Variablenselektion, weil sie dem Suchalgorithmus gewissermaßen als Landkarte dient und die Richtung hin zur Verbesserung des resultierenden Modells weist.

Gleichzeitig ist die Art der Gütefunktion eng mit der Wahrscheinlichkeit von Zufallskorrelationen verknüpft, was durch Simulationsstudien gezeigt wurde[52,30]. Einige allgemeine Faustregeln zur Reduzierung dieser Wahrscheinlichkeit sollten bei der Variablenselektion beachtet werden:

- Das Objekt/Variablen-Verhältnis sollte nicht kleiner als 6 sein.
- Die Anzahl der Objekte im Datensatz sollte nicht kleiner als 20 sein.
- Wird eine Kreuzvalidierung als Gütefunktion verwendet, so sollte die Anzahl ausgelassener Objekte zwischen 40 und 60 Prozent liegen.

Anhand des letztgenannten Punkts wird deutlich, daß die LOO-CV (siehe 2.2.3.1, Seite 18) als Gütefunktion in der Variablenselektion nicht geeignet ist. Eine ausreichend strenge Modellvalidierung, die weniger zu überoptimistischen Gütekennzahlen neigt, ist erst mit der LMO-CV möglich. In dieser Arbeit wurde bei der Variablenselektion stets die LMO-CV mit 50 Prozent ausgelassenen Objekten (L50%O-CV) als Gütefunktion verwendet; auch die beiden anderen o. g. Kriterien wurden stets eingehalten.

2.2.7 Klassifizierung

Als Klassifizierung bezeichnet man die Einteilung von Objekten in Klassen oder Kategorien anhand ihrer spezifischen Eigenschaften. Aus dem Alltag ist beispielsweise die Taxonomie in der Biologie bekannt, die Einteilung der Lebewesen etwa in Pflanzen/Tiere, Säuger/Reptilien, Nager/Primaten usw. Auch die Zuordnung von Waren zu bestimmten Preisklassen oder die Einordnung von Bankkunden als kreditwürdig oder nicht kreditwürdig stellt eine Klassifizierung dar.

In der vorliegenden Arbeit wurde das Klassifizierungsverfahren der Linearen Diskriminanzanalyse verwendet, um Moleküle anhand ihres Bindungsmodus an einem Rezeptor in verschiedene Klassen einzuteilen. Denkbar ist hier beispielsweise die Unterscheidung der Liganden in zwei Gruppen, je nachdem, ob eine bestimmte Wasserstoffbrückenbindung ausgebildet wird oder nicht. Existiert nämlich eine Serie von Inhibitoren mit bekannt hoher Aktivität, die allesamt diesen Bindungsmodus aufweisen, so kann die entsprechende Klassifizierung bislang unbekannter Moleküle ein Kriterium zur Abschätzung ihrer Aktivität liefern: Diejenigen Verbindungen, die ebenfalls diese Bindung ausbilden, können als potentiell aktiv angesehen werden, die anderen hingegen als inaktiv.

2.2.7.1 Lineare Diskriminanzanalyse (LDA)

Die Lineare Diskriminanzanalyse (engl. *Linear diskriminant analysis*, LDA) ist ein Verfahren zum Auffinden der Trennebene zwischen zwei Klassen; es wurde 1936 von FISHER eingeführt[53]. Im folgenden wird nur kurz der Zwei-Klassen-Fall beschrieben, der für die Problemstellungen im Rahmen dieser Arbeit zutreffend war. Für eine eingehendere Beschreibung der LDA und anderer Klassifizierungsverfahren in der Chemometrie wird auf die entsprechende Literatur[54] verwiesen.

Das einfachste Verfahren zur Klassifizierung per LDA im Zwei-Klassen-Fall bedient sich der Mahalanobis-Distanz (siehe 2.3.2). Zunächst werden die gemeinsame Varianz-Kovarianz-Matrix C beider Klassen gemäß Gl. 2.34 (Seite 35) sowie die beiden Klassenzentroiden μ_1 und μ_2 berechnet. Für jedes zu klassifzierende Objekte muß nun lediglich gemäß Gl. 2.36 (Seite 35) mit Hilfe von C die Mahalanobis-Distanz zu μ_1 bzw. μ_2 bestimmt werden. Das Objekt wird dann derjenigen Klasse zugeordnet, zu deren Zentroiden die Distanz minimal ist.

2.3 Distanz- und Ähnlichkeitsmaße

Viele chemometrische Techniken basieren auf der Bestimmung von Abständen oder Ähnlichkeiten. Die wichtigsten Methoden zur Berechnung solcher Maße werden in den folgenden Abschnitten vorgestellt. Eine Übersicht weiterer gebräuchlicher Distanzmaße findet sich beispielsweise bei WILLETT[55].

2.3.1 Euklidische Distanz

Die euklidische Distanz im zwei- oder dreidimensionalen Raum ist unsere gewohnte „Alltagsentfernung“. Sie entspricht dem Luftlinienabstand zwischen zwei Punkten A und B in der Ebene oder im Raum und erscheint zunächst intuitiv als das einzig sinnvolle Abstandsmaß. Spätestens beim Fußmarsch durch eine Stadt mit ausschließlich rechtwinklig angelegten Straßenzügen wie etwa Manhattan wird jedoch die unzulängliche Aussagekraft einer euklidischen Distanzangabe deutlich. Hier erweist sich die tatsächlich so benannte Manhattan-Distanz* als nützlicher, die ein zum Straßenverlauf paralleles Koordinatensystem anlegt und die Summe der Betragsdifferenzen der beiden Koordinaten von A und B erfaßt.[25]

Dennoch bleibt der euklidische Abstand das am häufigsten angewendete Maß. Die euklidische Distanz ED zwischen zwei Ortsvektoren $\mathbf{x}$ und $\mathbf{y}$ im n-dimensionalen Raum ergibt sich zu

$$\mathrm{ED}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \|\mathbf{x}-\mathbf{y}\| \tag{2.33}$$

2.3.2 Mahalanobis-Distanz

Die Mahalanobis-Distanz[56] besitzt gegenüber der euklidischen Distanz die interessante Eigenschaft, daß sie die Korrelation der Daten berücksichtigt. Sie kann leicht mit Hilfe der Varianz-Kovarianz-Matrix $\mathbf{C}$ berechnet werden; für

*auch als Canberra-, City-Block- oder Taxifahrer-Distanz bekannt.

eine spaltenzentrierte Matrix $\mathbf{X}$ mit n Zeilen ist diese definiert als

$$\mathbf{C_X} = \frac{1}{n-1} \cdot \mathbf{X}^T \cdot \mathbf{X} \tag{2.34}$$

Die Mahalanobis-Distanz MD zwischen zwei Objekten $\mathbf{x_1}$ und $\mathbf{x_2}$ aus dem Datensatz $\mathbf{X}$ ist dann

$$\mathrm{MD}(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{(\mathbf{x_1} - \mathbf{x_2})^T \cdot \mathbf{C_X}^{-1} \cdot (\mathbf{x_1} - \mathbf{x_2})} \tag{2.35}$$

Häufig wird auch die Mahalanobis-Distanz eines Objekts $\mathbf{x_i}$ zum Zentroiden des Datensatzes benötigt, für die entsprechend gilt

$$\mathrm{MD}_z(\mathbf{x_i}) = \sqrt{(\mathbf{x_i} - \bar{\mathbf{x}})^T \cdot \mathbf{C_X}^{-1} \cdot (\mathbf{x_i} - \bar{\mathbf{x}})} \tag{2.36}$$

Die Berücksichtigung der Korrelation der Daten durch die Mahalanobis-Distanz ist in Abb. 2.6 veranschaulicht. Hier liegt Punkt B bezüglich des Zentroiden des Datensatzes in einer Richtung geringer Korrelation. Die Wahrscheinlichkeit, daß ein weiterer zum Datensatz hinzugefügter Meßpunkt in der Region um B läge, ist geringer als die Wahrscheinlichkeit, daß er in der Region um Punkt A läge. Punkt A nämlich liegt in Richtung hoher Korrelation.

Diese unterschiedliche Wahrscheinlichkeit wird durch die Mahalanobis-Distanz berücksichtigt: Punkt B besitzt eine höhere MD (2.5 Einheiten, angezeigt durch die äquidistanten Ellipsen) zum Zentroiden als Punkt A (1 Einheit). Die euklidischen Distanzen von A und B zum Zentroiden sind dagegen identisch (jeweils 1 Einheit), da die ED die Korrelation der Daten nicht berücksichtigt.

2.3.2.1 Mahalanobis-Distanz im PC-Raum

Im Hauptkomponenten-Raum gelten für die Mahalanobis-Distanz einige nützliche Zusammenhänge, die hier nur kurz erwähnt seien. So entspricht die MD im normalisierten PC-Raum bis auf einen konstanten Faktor der ED,

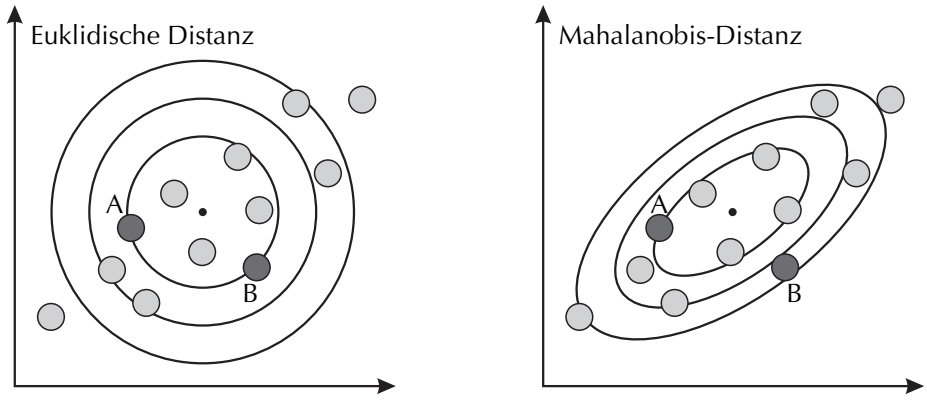


Abbildung 2.6 Euklidische Distanz *vs.* Mahalanobis-Distanz. Die beiden Objekte A und B besitzen dieselbe euklidische Distanz zum Zentroiden. Ihre Mahalanobis-Distanzen sind dagegen unterschiedlich (A: 1 Einheit, B: 2.5 Einheiten), da A der Korrelation der Daten folgt, B hingegen nicht.

sofern alle Hauptkomponenten berücksichtigt werden:

$$\mathrm{MD}_{\mathrm{PC,alle}} = \sqrt{n-1} \cdot \mathrm{ED}_{\mathrm{PC,alle}} \tag{2.37}$$

Weiterhin ist — wiederum unter Berücksichtigung aller Hauptkomponenten — die MD des Original-Datenraums identisch mit der MD des PC-Raums; daraus ergibt sich eine Äquivalenz der MD im Originaldatenraum und der ED im vollständigen PC-Raum

2.3.3 Dimensionsabhängigkeit

Aufgrund der in der QSAR meist verwendeten hohen Anzahl von Deskriptoren müssen die eingesetzten chemometrischen Techniken hochdimensionale Daten verarbeiten. Distanzen müssen also gegebenenfalls nicht nur im gewohnten zwei- oder dreidimensionalen Raum gemessen werden, sondern in Räumen mit oft 100 und mehr Dimensionen. Dies führt jedoch zu einem „verfluchten“ Problem, das als *The Curse of Dimensionality* bekannt ist.[57–59]

Eine Behandlung der Geometrie n-dimensionaler Hyperräume würde weit über die Intention dieser Arbeit hinausgehen; eine interessante Einleitung in dieses Thema findet sich beispielsweise bei KÖPPEN[57] oder VERLEYSEN[58]. Als wichtiger Aspekt dieses Phänomens bleibt festzuhalten, daß der maxima-

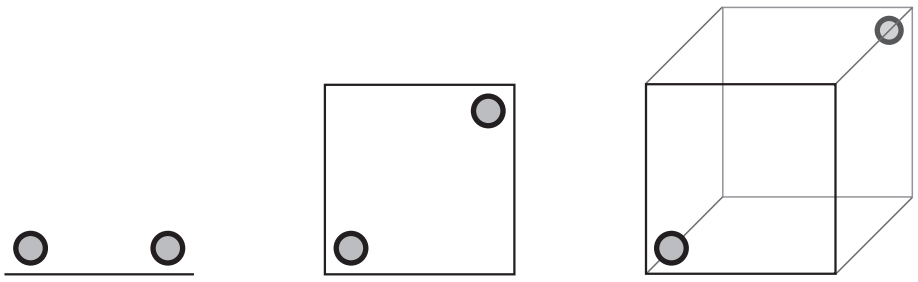


Abbildung 2.7 Dimensionsabhängigkeit von Distanzen (auch bekannt als *The Curse of Dimensionality*): Mit steigender Dimensionalität des Raums wird die maximale Distanz $d_{\max}$ zwischen zwei Punkten immer größer. Hier gezeigt ist der ein-, zwei- und dreidimensionale Fall im Einheitsraum mit Kantenlänge eins.

le Abstand zwischen zwei Punkten eines Raums mit steigender Dimensionalität anwächst. Dies ist für den ein- bis dreidimensionalen Fall in Abb. 2.7 anschaulich gezeigt.

Gegeben sei ein euklidischer n-dimensionaler Einheitsraum, also ein Hypercubus $[0,1]^n$. Dieser entspricht im eindimensionalen Fall einer Strecke der Länge $a = 1$, im zweidimensionalen Fall einem Quadrat der Kantenlänge $a = 1$ und im dreidimensionalen Fall einem Würfel ebenfalls der Kantenlänge $a = 1$. Der maximale Abstand $d_{\max}$ zwischen zwei Punkten des jeweiligen n-dimensionalen Raums beträgt für $n = 1$: $d_{\max} = 1$, für $n = 2$: $d_{\max} = \sqrt{2} = 1.41$ und für $n = 3$: $d_{\max} = \sqrt{3} = 1.73$ usw. Ein hochdimensionaler QSAR-Datensatz spannt also einen praktisch „leeren“ Hyperraum auf, da die Abstände zwischen den Objekten sehr groß sind.

Zwar sind von diesem „Fluch der Dimensionalität“ alle Objekte eines Datensatzes gleichermaßen betroffen. Die Beurteilung von Distanzen ist jedoch nur unter der Voraussetzung sinnvoll, daß alle Operationen und vergleichenden Funktionen dieselbe Dimensionsabhängigkeit zeigen wie die Distanzen selbst.

2.3.4 Distanzen als Ähnlichkeitsmaß

Wird ein Molekül, wie in 2.1.1 dargestellt, mit Hilfe von Deskriptoren charakterisiert, so erhält man einen Eigenschaftsvektor des Moleküls, der die be-

rechneten Werte der unterschiedlichen Deskriptoren enthält. Für einen Vektor $\mathbf{v} = [v_1, v_2, v_3]$ mit drei Elementen ist die geometrische Interpretation geläufig, seine Elemente als Koordinaten eines Punkts im dreidimensionalen Raum aufzufassen. Auch wenn Punkte in höherdimensionalen Räumen, wie sie analog von Vektoren mit mehr als drei Elementen beschrieben würden, unsere geometrische Vorstellungskraft überfordern, so ist diese Sichtweise dennoch mathematisch korrekt. Auch der euklidische Abstand zweier Punkte zueinander ist für den dreidimensionalen Fall ebenso definiert wie für den n-dimensionalen. Es spricht also nichts dagegen, ein Molekül als ein Objekt im n-dimensionalen Hyperraum anzusehen, dessen Koordinaten durch seine n Deskriptorenwerte definiert sind.

Offensichtlich weisen also Moleküle, die im Hyperraum eng beieinander liegen, eine hohe Ähnlichkeit bezüglich ihrer Deskriptorenwerte auf. Da die Deskriptoren die physikalisch-chemischen Eigenschaften des Moleküls charakterisieren, besitzen chemisch ähnliche Moleküle eine geringe Distanz zueinander, während chemisch nicht verwandte Verbindungen weit voneinander entfernt liegen. Die Aussagekraft bezüglich der *chemischen* Ähnlichkeit wird dabei durch die verwendeten Deskriptoren bestimmt: Sind zwei Moleküle nur durch ihr Molekulargewicht und die Anzahl der Einfachbindungen charakterisiert, liefert ihre Distanz keine Information über ihre Ähnlichkeit hinsichtlich der Fähigkeit, Wasserstoffbrückenbindungen auszubilden.

Die euklidische Distanz zweier Moleküle im n-dimensionalen Deskriptorraum stellt also ein anschauliches Maß ihrer Ähnlichkeit dar,[55] das zudem gemäß Gl. 2.33 (Seite 34) leicht zu berechnen ist. Dabei ist die Distanz natürlich ein „umgekehrtes“ Ähnlichkeitsmaß, d. h. eine geringe Distanz bedeutet eine hohe Ähnlichkeit.

2.3.5 Tanimoto-Koeffizient

Der Tanimoto-Koeffizient* T_c ist das am häufigsten eingesetzte Maß zur Bestimmung der Ähnlichkeit von Fingerprints. Er mißt den relativen Anteil gemeinsamer Eigenschaften bezogen auf die Variablen. Dazu wird die Anzahl der Eigenschaften, die beide Objekte besitzen, durch die Anzahl der Eigen-

*Auch als „Jacard-Distanzfunktion“ bekannt.

schaften geteilt, die jeweils nur ein Objekt besitzt.[55] Demnach folgt der Tanimoto-Koeffizient zweier Moleküle A und B der allgemeinen Formel

$$T_c(A,B) = \frac{\sum x_{A,i} \cdot x_{B,i}}{\sum x_{A,i}^2 + \sum x_{B,i}^2 - \sum x_{A,i} \cdot x_{B,i}} \tag{2.38}$$

Gegeben seien beispielsweise zwei Fingerprints A und B, die zehn Eigenschaften kodieren; sie bestehen also aus zehn Bits, die jeweils den Wert 0 oder 1 annehmen können (Eigenschaft vorhanden bzw. Eigenschaft nicht vorhanden). Zur Berechnung des Tanimoto-Koeffizienten wird nun in A und B die Anzahl a bzw. b von Bits mit dem Wert 1 gezählt; dasselbe geschieht für die logische UND-Verknüpfung (Multiplikation in Gl. 2.38) von A und B.

$$\begin{aligned} \text{A} &= 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1 \quad &\Rightarrow \quad a = 6 \\ \text{B} &= 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 1 \quad &\Rightarrow \quad b = 4 \\ \text{A UND B} &= 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 1 \quad &\Rightarrow \quad c = 3 \end{aligned}$$

Der Tanimoto-Koeffizient T_c für dieses Beispiel ist also

$$T_c = \frac{c}{a+b-c} = \frac{3}{6+4-3} = 0.43 \tag{2.39}$$

Der Tanimoto-Koeffizient kann demnach Werte zwischen $0 \leq T_c \leq 1$ annehmen, wobei $T_c = 1$ eine exakte Übereinstimmung der vorhandenen Eigenschaften anzeigt, während bei $T_c = 0$ keine gemeinsamen Eigenschaften existieren. Es sei allerdings darauf hingewiesen, daß der Tanimoto-Koeffizient nur *vorhandene* Eigenschaften (1-Bits) berücksichtigt, übereinstimmende 0-Bits also nicht zu einem höheren Tanimoto-Ähnlichkeitswert führen.

Gelegentlich werden zur Beschreibung von Molekülen auch Fingerprints eingesetzt, die nicht nur aus 0 und 1 bestehen, sondern beliebige ganzzahlige Werte annehmen können. Es handelt sich also nicht im eigentlichen Sinne um Bitstrings. So könnte z. B. nicht nur das (Nicht)-Vorhandensein eines Fragments kodiert werden (1/0), sondern eine konkrete Zählung erfolgen, wie oft das Fragment im Molekül vorkommt. Um derartige „Integer"-Fingerprints* zu vergleichen, muß die ausführliche Form (Gl. 2.38) verwendet werden.

*Der Variablentyp für ganzzahlige Werte wird in der Informatik als „Integer" bezeichnet.

2.4 Docking

Die Wirkung eines Medikaments beruht in den meisten Fällen auf der Bindung des Wirkstoffmoleküls an die biologische Zielstruktur, etwa ein Enzym. Dessen Struktur weist dafür eine spezielle Region auf, die als Bindetasche oder *Active site* bezeichnet wird. Oft wird die das Wechselspiel zwischen Substrat und Enzym mit dem von EMIL FISCHER geprägten Begriff des Schlüssel-Schloß-Prinzips[60–62] beschrieben: Ebenso wie in ein Schloß nur ganz bestimmte Schlüssel mit der richtigen Anzahl und Anordnung von Zacken hineinpassen, so können auch nur Wirkstoffmoleküle mit einer bestimmten chemischen Struktur in der Bindetasche des Enzyms gebunden werden.

Das Computermodell des gezielten Einpassens eines Ligandmoleküls in die Bindetasche, also die virtuelle Suche nach einem passenden Schlüssel für ein gegebenes Schloß, wird als *Docking* bezeichnet.[63] Es handelt sich also gewissermaßen um die computergestützte Simulation des komplexen Vorgangs der molekularen Erkennung. Dazu müssen prinzipiell zwei voneinander getrennte Probleme gelöst werden: Erstens gilt es, die Orientierung und Konformation des Liganden in der Bindetasche vorherzusagen. Diese werden im wesentlichen bestimmt durch die Wechselwirkungen zwischen dem Liganden und dem Rezeptor. Zweitens müssen die gefundenen möglichen Plazierungen im Hinblick darauf beurteilt werden, wie „fest“ der Ligand in der jeweiligen Position an den Rezeptor gebunden ist. Für dieses *Scoring* ist folglich die Abschätzung der Bindungsaffinität notwendig.

2.4.1 Dockingalgorithmen

Zur Adressierung des ersten Problems existieren im wesentlichen drei unterschiedliche Ansätze: Programme wie DOCK[64] oder FRED[65] verfolgen den sogenannten *Rigid-body approach*, bei dem für jeden Liganden eine ganze Bibliothek unterschiedlicher Konformationen auf ihre sterische Komplementarität bezüglich der Bindetasche überprüft wird; die einzelnen Konformationen selbst bleiben dabei starr. Dagegen versuchen beispielsweise AUTODOCK[66] und GOLD[67], die Konformation des Liganden in der Bindetasche mit Hilfe von genetischen Algorithmen zu optimieren. Der dritte An-

satz schließlich besteht im inkrementellen Aufbau des Liganden in der Bindetasche. Programme wie FLEXX[68,69], PHDOCK[70] oder HAMMERHEAD[71] zerteilen den Liganden dazu zunächst in einzelne Fragmente und setzen ihn anschließend ausgehend von einem Basisfragment in der Bindetasche wieder flexibel zusammen.

2.4.2 Scoringfunktion

Das Scoring, also die Abschätzung der Bindungsaffinität des in der Bindetasche plazierten Liganden, ist die zweite große Herausforderung des Dockings. Neben der Bewertung der möglichen Ligand-Rezeptor-Interaktionen und ihren Beitragen zur Bindungsenergie müssen eine Reihe weiterer Faktoren berücksichtigt werden. Insbesondere entropische Effekte durch Änderungen der Solvatationszustände, dem Verlust von Freiheitsgraden bei der Bindung des Liganden usw. spielen eine wichtige Rolle.

Die Parametrisierung der ersten Scoringfunktionen, beispielsweise der von BÖHM für das Programm LUDI entwickelten Funktion[72], erfolgte meist empirisch durch den Versuch, experimentell bekannte Freie Bindungsenthalpien ΔG von Ligand-Rezeptor-Komplexen zu reproduzieren. Der berechnete Dockingscore sollte also möglichst gut mit den bekannten ΔG-Werten korrelieren. Das im folgenden vorgestellte Dockingprogramm FLEXX etwa verwendet eine modifizierte Version der von BÖHM für das Programm LUDI[72] entwickelten Scoringfunktion zur Berechnung der Freien Bindungsenthalpie. Berücksichtigt werden dabei u. a. der Entropieverlust bei der Bindung des Liganden (Einschränkung der Anzahl rotierbarer Bindungen), die paarweisen Ligand-Rezeptor-Interaktionen (Wasserstoffbrückenbindungen, ionische und aromatische Wechselwirkungen), lipophile Interaktionen sowie unerlaubt nahe Atom-Atom-Kontakte zwischen Ligand und Rezeptor (engl. *Clashes*).

Das Ziel der exakten Vorhersage von Bindungsenthalpien wird jedoch in der Praxis des Dockings und zumal im virtuellen Screening (siehe 2.5, Seite 48) nicht verfolgt. Hier werden die berechneten Dockingscores keineswegs direkt als ΔG-Werte interpretiert, sondern vielmehr als abstrakte Größen zum Vergleich der an einem Target gedockten Liganden untereinander benutzt.

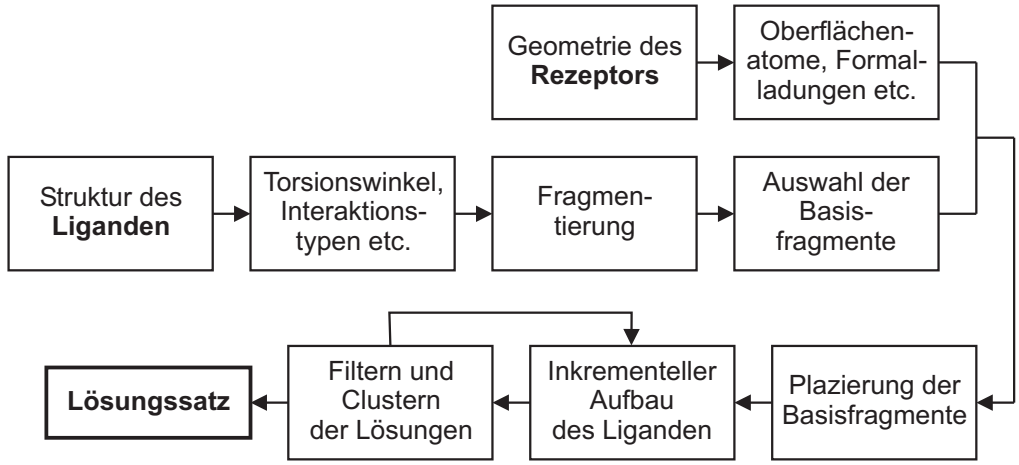


Abbildung 2.8 Schematische Darstellung des FlexX-Algorithmus.

In der vorliegenden Dissertation wurde für alle Dockingexperimente ausschließlich das Programm FLEXX verwendet. Daher wird auf eine detaillierte Beschreibung der Algorithmen anderer Ansätze und alternativer Dockingprogramme verzichtet; in der Literatur finden sich dazu entsprechende Übersichtsartikel[73].

2.4.3 FlexX

FLEXX ist ein Dockingprogramm, das den Ansatz des inkrementellen Ligandaufbaus verfolgt. Der zugrundeliegende Algorithmus ist schematisch in Abb. 2.8 dargestellt.

Zunächst werden sowohl für den Liganden als auch für die Bindetasche des Rezeptors alle für eine potentielle Ligand-Rezeptor-Interaktion verantwortlichen Parameter bestimmt. Nach der Fragmentierung des Liganden durch Aufbrechen aller frei drehbaren Einfachbindungen werden zunächst mehrere Plazierungen eines Basisfragments (bestehend aus ein bis drei Einzelfragmenten) evaluiert. Dabei hat der Algorithmus zum Ziel, jeweils mindestens drei Ligand-Rezeptor-Interaktionen herzustellen. Ausgehend von diesen Basisfragment-Plazierungen werden dann weitere Fragmente an die bestehenden Lösungen angefügt und der Ligand dadurch schrittweise wieder aufgebaut. Am Ende jeder Iteration wird der Lösungssatz durch Zusam-

menfassen ähnlicher und Eliminierung chemisch nicht sinnvoller Lösungen verkleinert, um die Gesamtzahl von Zwischenlösungen handhabbar zu halten.

Zur Abschätzung der Bindungsaffinität einer gefundenen Plazierung und damit zur Bewertung der Qualität einer Lösung ist in FLEXX die in 2.4.2 (Seite 41) genannte Scoringfunktion implementiert.

2.4.3.1 Ligand-Rezeptor-Interaktionen

Jede mögliche Wechselwirkung zwischen Ligand und Rezeptor wird von FLEXX durch ein Interaktionszentrum sowie die Form einer dieses Zentrum umgebenden Wechselwirkungsoberfläche beschrieben (siehe Abb. 2.9). Für eine Methylgruppe etwa ist dies eine volle Kugeloberfläche, für aromatische Ringe jeweils eine Kappe (kreisförmiger Ausschnitt einer Kugeloberfläche) ober- und unterhalb der Ringebene, für eine Carboxylatgruppe zwei sphärische Rechtecke (Überschneidung zweier Kugelzweiecke, deren Achsen senkrecht aufeinanderstehen) entsprechend den freien Elektronenpaaren an den Sauerstoffatomen (siehe Abb. 2.9). Damit eine Wechselwirkung ausgebildet werden kann, müssen Akzeptor- und Donorzentrum gegenseitig auf den Wechselwirkungsoberflächen des Interaktionspartners liegen.

In FLEXX haben Donor- und Akzeptor-Interaktionen sowie ionische Wechselwirkungen die höchste Priorität. Danach folgen räumlich gerichtete hydrophobe Wechselwirkungen (etwa zwischen aromatischen Zentren) und schließlich unspezifische hydrophobe Interaktionen (beispielsweise zwischen aliphatischen Kohlenstoffatomen).

2.4.4 Structural Interactions Fingerprint (SIFt)

Im Rahmen dieser Arbeit kam eine Modifikation einer Analysemethode von Protein-Ligand-Interaktionen zum Einsatz, die von SINGH *et al.* unter dem Namen *Structural interactions fingerprints* (SIFt) publiziert wurde.[74,75] Dabei werden die im Dockingexperiment ermittelten Interaktionen zwischen Ligand und Rezeptor in Form eines Fingerprints (siehe 2.1.2, Seite 8) gespeichert, der die dreidimensionale Bindungsinformation in einen eindimensionalen Bitstring überführt. Dies ermöglicht das einfache Clustern, Filtern und

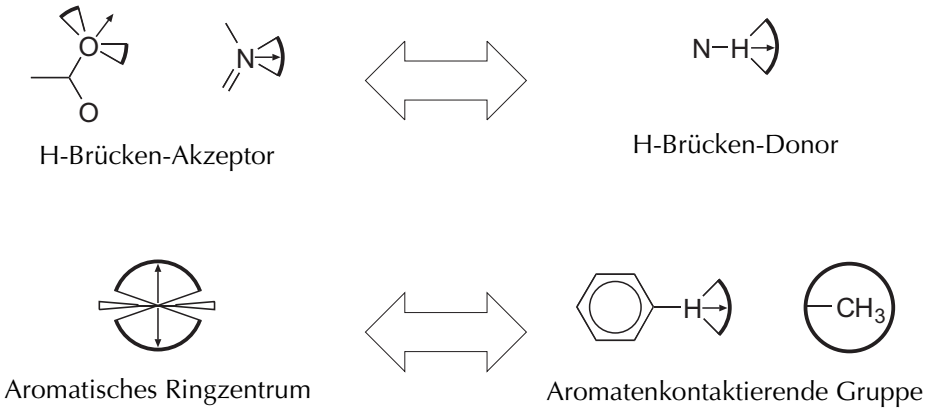


Abbildung 2.9 Ausgewählte Ligand-Rezeptor-Interaktionen in FlexX: Jedem Interaktionszentrum wird eine entsprechende Wechselwirkungsoberfläche zugewiesen. Donor-Akzeptor-Interaktionen besitzen eine höhere Priorität als etwa aromatische Wechselwirkungen.

Vergleichen von komplexen geometrischen Ligandplazierungen (Dockinglösungen).

Für die direkte Verwendung des FLEXX-Interaktionsmodells wurde der ursprünglich publizierte Algorithmus leicht modifiziert: Zur Erzeugung eines SIFts wurde zunächst für jede Dockinglösung ein Bitstring angelegt, in dem jedes Bit einen prinzipiell möglichen Interaktionsort und -typ der Bindetasche des Rezeptors repräsentiert. Trat genau diese Interaktion in der ermittelten Dockinglösung tatsächlich auf, so wurde das entsprechende Bit gleich Eins gesetzt; andernfalls verblieb es im Zustand Null. Diese Vorgehensweise ist schematisch in Abb. 2.10 dargestellt. Damit genügen der angepaßten Methode allein die beim Docking mit FLEXX sowieso erzeugten Daten; der Einsatz zusätzlicher Software und Rechenzeit ist nicht notwendig.

2.4.5 Kovalentes Docking an Cysteinproteasen

Die korrekte Ermittlung der Plazierung eines kovalent bindenden Inhibitors und die Abschätzung seiner Bindungsaffinität stellen eine besondere Herausforderung dar, da die verfügbaren Dockingprogramme üblicherweise nur nicht-kovalente Interaktionen in Betracht ziehen. Die Simulation einer kovalenten Inhibition, also einer chemischen Reaktion mit der Knüpfung einer neuen chemischen Bindung zwischen Rezeptor und Ligand ist dagegen nicht

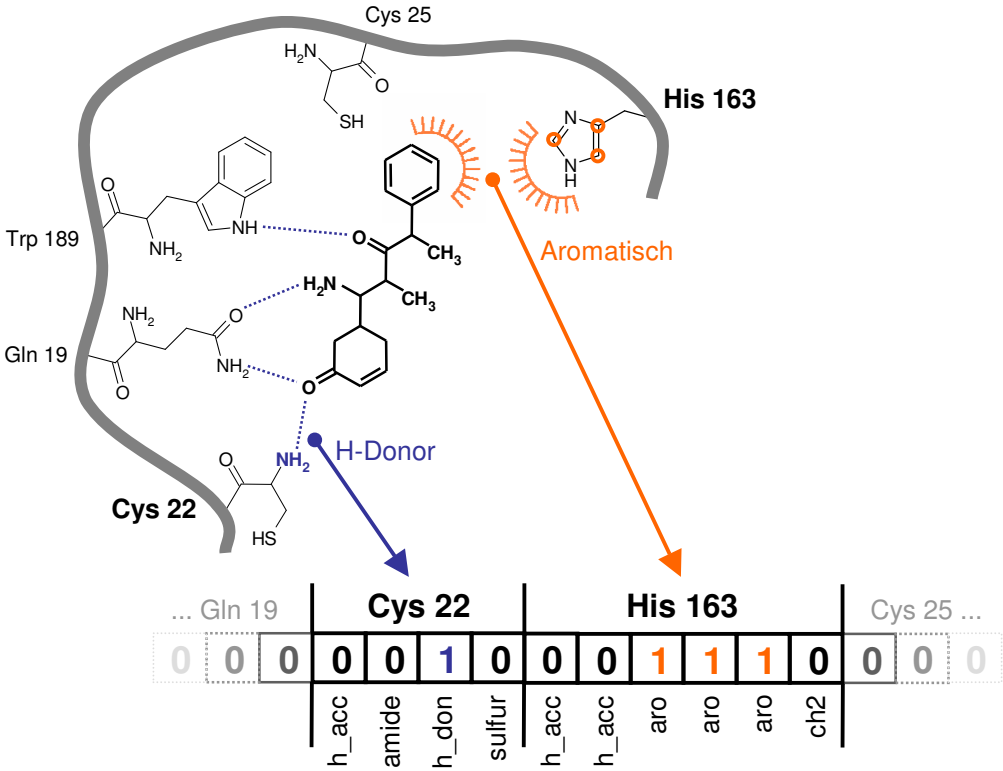


Abbildung 2.10 Schematische Darstellung der Erzeugung eines SIFt (*Structural interactions fingerprint*). Der Fingerprint enthält zunächst für alle prinzipiell möglichen Interaktionszentren und -typen des Rezeptors je ein 0-Bit. Wird beim Docking eine Wechselwirkung tatsächlich gefunden, so wird das entsprechende Bit auf 1 gesetzt (hier exemplarisch für drei aromatische Interaktionen des Liganden mit His163 und eine H-Donor-Interaktion mit Cys22).

oder nur stark eingeschränkt vorgesehen. Zudem führt die Ausbildung einer neuen chemischen Bindung zwangsläufig zu konstitutionellen Änderungen des Liganden, die gegenüber der Eingangsstruktur berücksichtigt werden müssen.

Im vorliegenden Fall von Aziridinen als kovalente Inhibitoren der Cysteinproteasen Cathepsin B und Cathepsin L[76–79] (siehe Abb. 2.11) stellen die beiden Kohlenstoffatome des Aziridinrings **1** ein elektrophile Zentren dar. Dieses können vom Schwefelatom des aktiven Cysteinrests **2** des Enzyms kovalent angegriffen werden. In einer S_N2-Reaktion wird eine neue kovalente S–C-Bindung zwischen Enzym und Ligand geknüpft, während die entsprechende C–N-Bindung des Aziridinrings gelöst wird (**3**).

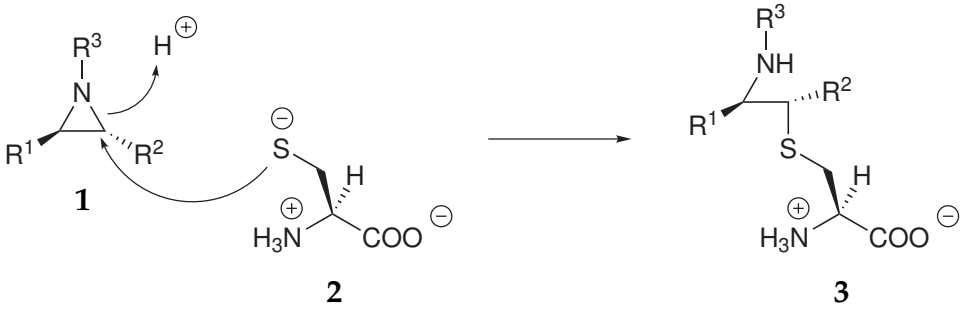


Abbildung 2.11 Bindung des kovalenten Inhibitors: Der Ringkohlenstoff des Aziridins (**1**) wird vom nukleophilen Schwefelatom des Cysteins (**2**) angegriffen. In einer S_N2-Reaktion entsteht der kovalente Ligand-Rezeptor-Komplex (**3**).

Als Grundvoraussetzung für diese Reaktion muß natürlich das elektrophile Zentrum zunächst in die Reichweite des Cystein-Nukleophils gelangen. Der wesentliche Teil des Inhibitions- und damit auch des Dockingvorgangs unterscheidet sich also nicht von dem eines nicht-kovalenten Inhibitors. Erst wenn der Ligand nicht-kovalent (reversibel) entsprechend seiner übrigen Ligand-Rezeptor-Interaktionen (Wasserstoffbrückenbindungen, hydrophobe Wechselwirkungen etc.) in der Bindetasche plaziert und somit der elektrophile Aziridinring vorfixiert ist, kommt es zur Ausbildung der kovalenten S–C-Bindung.[80,81]

Zur Simulation des entscheidenden ersten Schritts — der Vorfixierung des Liganden in der Bindetasche — wurde ein angepaßtes Dockingprotokoll unter Verwendung von FLEXX entworfen. Die Details sowie die Validierung und Anwendung dieses Protokolls werden im Ergebnisteil in 3.5.1 vorgestellt.

2.4.5.1 Bedeutung von Cysteinproteasen

Eine der größten und bedeutendsten Gruppen von Enzymen bilden die Proteasen, die selektiv die Hydrolyse von Peptidbindungen katalysieren. Sie werden in fünf Gruppen eingeteilt: Serin-, Cystein-, Aspartat-, Metallo- und Threonin-Proteasen.[82] Bei den Cysteinproteasen, werden sieben Superfamilien unterschieden, sogenannte Clans. Die meisten bislang bekannten Cysteinproteasen gehören dem Papain-Clan CA an, so auch Cathepsin B (CB) und Cathepsin L (CL), die in einem Teilprojekt dieser Dissertation bearbei-

tet wurden. Letztere kommen in beinahe jedem menschlichen Körpergewebe vor und werden, neben anderen Cysteinproteasen, mit Erkrankungen wie rheumatoider Arthritis, muskulärer Dystrophie und der Metastasenbildung von Tumoren in Verbindung gebracht.[83–85]

In 3.5 werden Dockingexperimente an CB und CL vorgestellt, die zur Unterstützung der Entwicklung CL-selektiver Inhibitoren dienten. Diese beiden Enzyme unterscheiden sich im wesentlichen durch den sogenannten *Occluding loop* mit zwei positiv geladenen His-Resten (His110 und His111), den nur CB aufweist; außerdem sind zwei Teilbereiche der Bindetasche bei CB enger als bei CL. Beide Merkmale können zur Erzielung der gewünschten Selektivität ausgenutzt werden.

2.5 Virtuelles Screening

Das virtuelle Screening (VS) oder auch virtuelle High-Throughput-Screening (vHTS), also die Suche nach aktiven Substanzen in großen virtuellen Molekülbibliotheken, ist eine der zentralen Technologien moderner Wirkstoffentwicklung.[86–88] Nach einer Untersuchung der BOSTON CONSULTING GROUP im November 2001[89] eröffnen *in-silico*-Technologien das Potential, die Entwicklungskosten eines neuen Medikaments um durchschnittlich 130 Millionen US-Dollar zu senken und die Entwicklungsdauer um etwa zehn Monate zu verkürzen.* Unter ökonomischen Gesichtspunkten ist also eine effiziente Wirkstoffentwicklung ohne den Einsatz computergestützter Technologien und dabei insbesondere des virtuellen Screenings kaum denkbar.[90]

Die Beschreibung des virtuellen High-Throughput-Screenings ist von zwei verschiedenen Standpunkten aus möglich: Aus der Sicht der Informatik handelt es sich beim vHTS um eine sehr ausgefeilte Form der Datenbanksuche; es gilt, virtuelle Moleküle mit einer gewünschten Eigenschaft in einer großen Menge von virtuellen Substanzen identifizieren, die diese Eigenschaft nicht aufweisen. Vom biologisch-chemischen Standpunkt aus stellt das vHTS eine stark vereinfachte Simulation etablierter Screeningassays dar, bei denen aus einer Bibliothek von „echten“ Substanzen etwa mit Hilfe von Enzymassays diejenigen Verbindungen ausgewählt werden, die im Experiment die gewünschte biologische Aktivität aufweisen.

Das Ziel des vHTS liegt im möglichst schnellen und effizienten Auffinden chemischer Strukturen, die eine bestimmte Eigenschaft besitzen, etwa eine Wirkung als Inhibitor eines gegebenen Enzyms zeigen. Aus einer großen Screeningdatenbank mit oft mehreren Millionen von Molekülen sollen also die wenigen aktiven Verbindungen herausgesucht werden. Dies erfordert eine Bewertung jedes einzelnen Moleküls hinsichtlich seiner Eignung, etwa seiner biologischen Aktivität am gewünschten Zielprotein. Als Ergebnis liegt schließlich eine Rangliste aller gescreenten Verbindungen vor, in der die am besten geeigneten Moleküle die vordersten Plätze einnehmen.

*Dieselbe Untersuchung ermittelt als durchschnittliche Entwicklungskosten eines Medikaments 880 Millionen US-Dollar und als durchschnittliche Entwicklungsdauer einen Zeitraum von 15 Jahren.

In der Praxis existiert jedoch keine absolut zuverlässige vHTS-Methode. Vielmehr enthalten die Ergebnisse immer auch eine gewisse Anzahl von Verbindungen, die fälschlicherweise als geeignet identifiziert wurden, sogenannte *False positives*. Auch der oberste Bereich der Rangliste ist also durchsetzt mit Molekülen, die — entgegen den Erwartungen — im nachfolgenden Experiment (z. B. einem Enzymassay) keine Aktivität zeigen. Ebenso werden nicht alle tatsächlich aktiven Verbindungen als solche identifiziert (*False negatives*), so daß auch eigentlich aussichtsreiche Moleküle auf die hinteren Plätze gelangen und damit der experimentellen Überprüfung womöglich nie zugänglich werden. Daher kann die realistische Erwartung an einen vHTS-Ansatz lediglich eine Anreicherung von geeigneten Strukturen unter den bestplazierten Ergebnissen des virtuellen Screenings sein, nicht aber die absolut zuverlässige Vorhersage von tatsächlich geeigneten Verbindungen.[91]

2.5.1 Datenbankanreicherung

Die Effizienz eines virtuellen Screenings kann anhand der Datenbankanreicherung (engl. *Database enrichment*) gemessen werden. In der Wirkstoffentwicklung besteht nämlich die übliche Vorgehensweise darin, die z. B. ersten 1000 Verbindungen der Rangliste auch experimentell im Labor auf ihre Aktivität zu überprüfen.[92] Das vHTS war also dann besonders nützlich, wenn sich besonders viele dieser 1000 getesteten Substanzen als tatsächlich aktiv herausstellen und damit die Vorhersage des virtuellen Screenings bestätigen. Folglich steht und fällt das Leistungsvermögen eines VS-Ansatzes mit seiner Fähigkeit, möglichst viele der zufällig über die gesamte Screeningdatenbank verteilten Aktiven in den ersten 1000 Plazierungen der Rangliste anzureichern und damit für die experimentelle Testung vorzuschlagen (siehe Abb. 2.12).

2.5.1.1 Maßzahlen

Bei retrospektiven Studien zur Datenbankanreicherung wird überprüft, welchen Prozentsatz von aktiven Verbindungen die Screeningtechnologie in einer größeren Menge von Inaktiven wiederzufinden vermag. Dazu werden beispielsweise einige hundert als aktiv bekannte Strukturen in einer großen

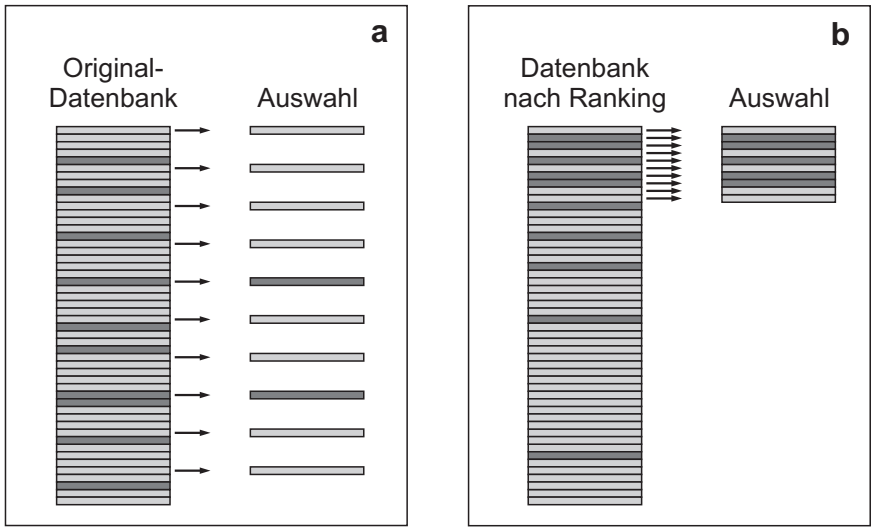


Abbildung 2.12 Schematische Darstellung der Datenbankanreicherung (engl. *Database enrichment*). In der Originaldatenbank sind die aktiven Verbindungen (dunkel dargestellt) zufällig unter den inaktiven (hell) verteilt; eine willkürliche Auswahl führt hier zu zwei Hits (a). Dagegen resultiert das virtuelle Screening in einer Sortierung der Datenbank, die zu einer Anreicherung der Aktiven unter den bestplazierten Verbindungen führt; hier ergibt die gezielte Auswahl der ersten Ränge fünf Hits (b).

Screeningdatenbank (Hunderttausende oder sogar Millionen von Molekülen) „versteckt". Zur praxisnahen Beurteilung der Leistungsfähigkeit des vHTS-Ansatzes kann dann die Information herangezogen werden, wieviele der aktiven Verbindungen sich unter den ersten 1000 oder im ersten Prozent der vorgeschlagenen Kandidaten wiederfinden. Der sogenannte *Recall* R_x für die ersten x Prozent der Rangliste ist definiert als der Prozentsatz der wiedergefundenen Aktiven a gegenüber der Gesamtzahl aller Aktiven A:[93]

$$R_x = \frac{100 \cdot a}{A} \tag{2.40}$$

Ein ebenfalls gebräuchliches Maß für die Performance eines Screeningverfahrens ist der Anreicherungsfaktor (engl. *Enrichment factor*) EF_x.[94] Er beschreibt die Anzahl der wiedergefundenen Aktiven a in den ersten x Prozent der Rangliste im Verhältnis zur erwarteten Anzahl e von Aktiven innerhalb

desselben Prozentsatzes x bei zufälliger Sortierung der Datenbank.

$$EF_x = \frac{a}{e} = \frac{a}{A \cdot \frac{x}{100}} \tag{2.41}$$

Der Anreicherungsfaktor EF_x gibt also an, wievielfach besser das vHTS gegenüber einer Zufallsauswahl abschneidet (bezogen auf die Anzahl der gefundenen Aktiven).

2.5.2 Strukturbasierter Ansatz

Steht eine 3D-Struktur des Zielmoleküls (engl. *Target*) zur Verfügung, so wird für das virtuelle Screening zumeist ein strukturbasierter* Ansatz wie Docking gewählt.[95] Dazu werden alle Moleküle der Screeningdatenbank in die Targetstruktur gedockt und anschließend gemäß dem resultierenden Dockingscore sortiert. Diese Vorgehensweise folgt der Hypothese, daß Moleküle mit einem hohen Dockingscore auch tatsächlich gute Inhibitoren des gewählten Targets sind, also auch im biologischen Experiment etwa einen niedrigen IC_{50}-Wert aufweisen sollten.

Eine der größten Hürden des strukturbasierten Dockingansatzes liegt in der mangelnden Berücksichtigung der Flexibilität des Proteins. Während in den meisten Dockingverfahren zwar der konformellen Flexibilität des Liganden Rechnung getragen wird, muß die Bindetasche weiterhin als starr angenommen werden. Die ansonsten geradezu explodierende Zahl von Freiheitsgraden des Ligand-Rezeptor-Systems wäre auch mit modernen Ressourcen nicht zu bewältigen.

Ein weiteres Problem erwächst aus den notwendigen Näherungen der Scoringfunktion.[96] In den letzten Jahren hat sich entgegen früherer Annahmen herausgestellt, daß nicht die Vorhersage der korrekten Konformation und Orientierung des Liganden die Hauptschwierigkeit des Dockings darstellt, sondern der nachfolgende Schritt des Scorings bzw. Rankings.[97–99] Die Beurteilung einer Dockinglösung durch nicht nur eine einzelne, sondern durch mehrere Scoringfunktionen (engl. *Consensus scoring*) kann zwar die Zuverlässigkeit der Ergebnisse erhöhen, räumt das Problem jedoch nicht grund-

*Auch als „rezeptorbasiert“ bezeichnet.

sätzlich aus.[100,101] Weitergehenden Ansätzen wie etwa aufwendigen Kraftfeldoptimierungen steht — gerade im industriellen Umfeld — die Erfordernis einer möglichst schnellen und bezüglich der benötigten Rechenzeit kostengünstigen Technologie entgegen.

Die auch nur annähernd exakte Berechnung der Bindungsenthalpie als Kriterium für das anschließende Sortieren der gedockten Verbindungen ist also nicht das Ziel des strukturbasierten virtuellen Screenings. Es kann vielmehr als ein sehr hochentwickelter Filter angesehen werden, der prinzipiell geeignete Verbindungen identifiziert und ungeeignete Verbindungen eliminiert — und somit in der Tat zum gewünschten Ergebnis der Anreicherung von Aktiven führt.

2.5.3 Ligandbasierter Ansatz

Der Beweggrund für die Wahl eines ligandbasiertes Screeningansatzes ist meist das Fehlen einer 3D-Struktur des Targets. Gerade bei Targets wie etwa G-Protein-gekoppelten Rezeptoren (GPCRs), für die keine Röntgenkristallstruktur existiert,* werden häufig ligandbasierte Techniken angewandt.[102] Hinter diesem Ansatz steht die der gesamten Wirkstoffentwicklung immanente Struktur-Wirkungs-Hypothese, aus der sich die Annahme ableiten läßt, daß chemisch bzw. strukturell ähnliche Moleküle auch ähnliche biologische Wirkungen aufweisen.[103,104]

2.5.3.1 Alignment

Die Technik des *Alignments* nutzt die virtuelle 3D-Struktur eines Referenzmoleküls als Vorlage, auf die andere Moleküle überlagert werden.[105] Die Superpositionierung orientiert sich dabei an übereinstimmenden Merkmalen zwischen den beiden Strukturen wie etwa der räumlichen Gestalt, funktionellen Gruppen oder Wechselwirkungsmöglichkeiten. Ein anschauliches Beispiel insbesondere für den letztgenannten Punkt der analogen Interaktionsmöglichkeiten liefert die in Abb. 2.13 zweidimensional gezeigte Überlagerung von Methotrexat (**4**) und Dihydrofolat (**5**).

*Im Fall der GPCRs wurde bislang nur die 3D-Struktur von Rhodopsin aufgeklärt.

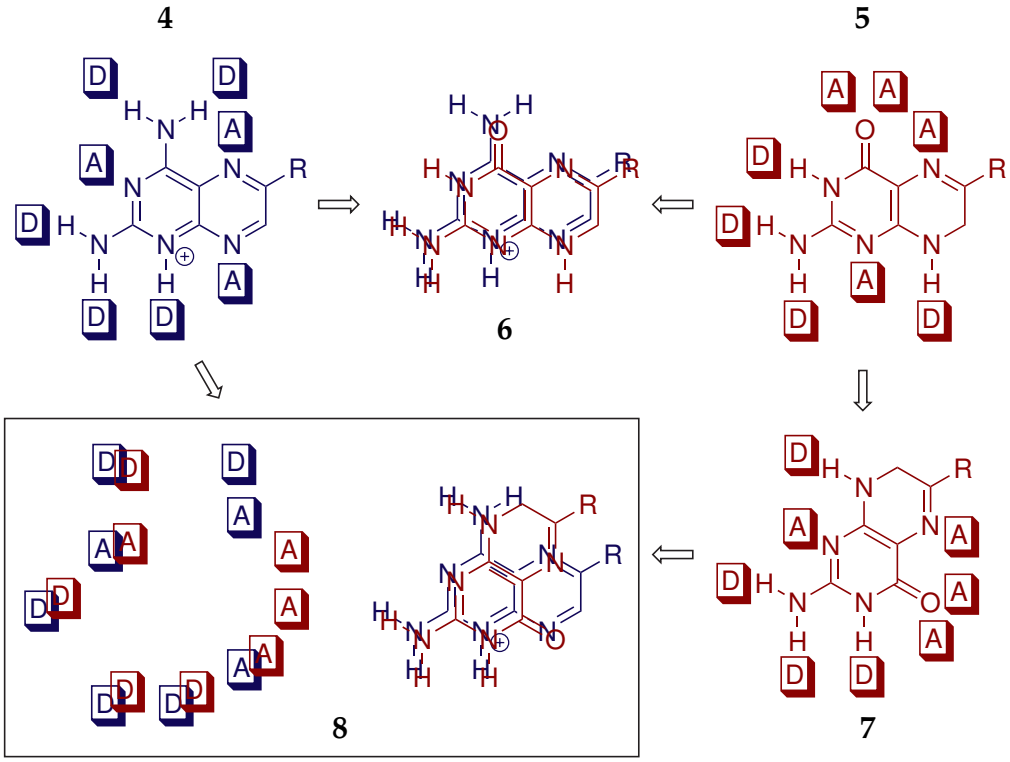


Abbildung 2.13 Überlagerung (Alignment) von Methotrexat (**4**) und Dihydrofolat (**5**). „A“ bezeichnet eine Akzeptorfunktion, „D“ eine Donorfunktion. Bei der zunächst intuitiven Überlagerung **6** stimmen zwar die Gestalt der Moleküle sowie die Position der Heteroatome überein. Eine weitgehende Übereinstimmung der für die Ligand-Rezeptor-Interaktion entscheidenden Interaktionsmöglichkeiten („A“ und „D“) liegt jedoch erst für die Überlagerung **8** des gedrehten/gespiegelten Dihydrofolatmoleküls **7** mit der unveränderten Methotrexatstruktur vor.

2.5.3.2 Ähnlichkeitssuche und Substruktursuche

Eine Datenbank kann auch mit Hilfe eines Ähnlichkeitsmaßes gescreent werden[106,93]. Das Erstellen der Rangliste erfolgt dann anhand der berechneten Ähnlichkeit bezüglich der gegebenen Referenzstruktur(en). Dazu kann beispielsweise die Distanz zur Referenzstruktur im Deskriptorraum oder die Tanimoto-Ähnlichkeit der durch Fingerprints repräsentierten Moleküle verwendet werden (siehe 2.3.4 und 2.3.5, Seite 38).

Für die Fingerprint-Ähnlichkeitssuche mit mehr als einer Referenzstruktur beschreiben HERT *et al.* drei verschiedene mögliche Herangehensweisen:[107]

Erstens die Verwendung eines kombinierten Fingerprints, in dem ein Bit immer dann auf Eins gesetzt wird, wenn es in der Mehrzahl der einzelnen Referenzen eingeschaltet ist. Zweitens die separate Ähnlichkeitssuche mit jeder einzelnen der Referenzen und die anschließende Zusammenführung der Ergebnisse (engl. *Data fusion*); für letzteren Schritt sind zwei Verfahren möglich, nämlich die Bildung der Summe über alle einzelnen Ähnlichkeitswerte oder die Betrachtung nur des Maximums der einzelnen Werte. Schließlich drittens die Sortierung der Datenbank anhand von Wichtungsfaktoren; diese ergeben sich aus der Anzahl der Aktiven, in den ein bestimmtes Bit eingeschaltet ist, geteilt durch die Anzahl aller Moleküle, in denen dieses Bit gleich Eins ist.

Gemäß der oben zitierten Studie[107] ist für die in dieser Arbeit durchgeführten Studien die Methode der separaten Ähnlichkeitssuchen mit anschließender Betrachtung des maximalen Ähnlichkeitswerts am effektivsten (Data fusion, MAX).

Ebenfalls um eine Art der Ähnlichkeitssuche handelt es sich bei der Substruktursuche. Mit Hilfe dieser Methode kann eine Screeningdatenbank nach allen Molekülen durchsucht werden, die eine bestimmte Substruktur enthalten. Ist beispielsweise eine Serie von aktiven Verbindungen bekannt, die alle ein bestimmtes chemisches Motiv besitzen, so können alle Moleküle aus der Datenbank extrahiert werden, die ebenfalls dieses oder ein ähnliches Motiv aufweisen. Eine solche Abfrage erfordert eine geeignete Art der Speicherung bzw. Konvertierung der chemischen Strukturen. Für Substruktursuchen hat sich das SMILES-Format[108] (*Simplified Molecular Input Line Entry Specification*) bzw. dessen Erweiterung SMARTS[109] (*Smiles Arbitrary Target Specification*) bewährt, in dem chemische Strukturen als eine einzelne Zeile von ASCII-Zeichen abgelegt werden. So lautet etwa der SMILES-String für Essigsäure `CC(=O)O` oder für Triethylamin `CN(C)C`. Durch dieses überaus handliche Format wird die Substruktursuche auf den bloßen Vergleich von Zeichenketten reduziert.

2.5.3.3 Pharmakophorsuche

Der 1909 von PAUL EHRLICH geprägte Begriff Pharmakophor bezeichnet im Wortsinn den Träger (griech. *phoros*) der biologischen Aktivität eines Arznei-

9 **10**

Abbildung 2.14 Das Histamin-Molekül und sein Pharmakophor, bestehend aus der positiv geladenen Gruppe (P), dem Wasserstoffbrücken-Akzeptor (A) und dem Wasserstoffbrücken-Donor (D).

stoffs (griech. *pharmacon*).[110] Er beschreibt eine definierte Anordnung chemischer Funktionalitäten im Wirkstoffmolekül, die mehreren Wirkstoffen gemeinsam ist, vom Rezeptor erkannt wird und demnach als Ursache der biologischen Wirkung anzusehen ist.[111,63] In vielen Fällen können konkrete Winkel und Abstände zwischen den Komponenten des Pharmakophors (also den relevanten funktionellen Gruppen) bestimmt werden. In Abb. 2.14 ist exemplarisch der Pharmakophor (**10**) des Botenstoffs Histamin (**9**) gezeigt, der sich hier aus der räumlichen Anordnung von positiv geladener Gruppe (P), Donor- (D) und Akzeptor-Funktionalität (A) zusammensetzt.

Um eine Datenbank anhand eines Pharmakophors zu screenen,[112] werden die enthaltenen Moleküle dahingehend untersucht, ob auch sie einen entsprechenden Pharmakophor enthalten, also den spezifizierten geometrischen Anforderungen genügen. Dazu werden aus der 3D-Struktur der zu screenenden Verbindungen zunächst verschiedene energetisch günstige Konformere generiert. Anschließend werden die Distanzen und Winkel zwischen den funktionellen Gruppen berechnet und mit dem Pharmakophormodell verglichen.

Die Struktur des Pharmakophors kann entweder durch den Vergleich und die Überlagerung mehrerer bekannter Liganden eines Targets bestimmt werden oder aber direkt aus der 3D-Struktur des Targets abgeleitet werden. Genaugenommen ist das virtuelle Screening per Pharmakophorsuche also prinzipiell gleichermaßen dem ligand- wie dem strukturbasierten Ansatz zuzuordnen.

2.5.3.4 QSAR-Modell

Auch ein QSAR-Modell kann zum virtuellen Screening verwendet werden. Zur Kalibrierung des Modells ist zunächst eine Serie von am betrachteten

Target aktiven Verbindungen notwendig. Anschließend können durch Anwendung der QSAR-Gleichung auf die deskriptor-kodierten Moleküle der Screeningdatenbank deren biologische Aktivitäten vorhergesagt werden.

Bei hochdiversen Screeningbibliotheken muß jedoch sehr sorgfältig die Existenz von Outliern überprüft werden (siehe 2.2.4, Seite 22), für die die Gültigkeit des QSAR-Modells nicht gewährleistet ist. Bei virtuellen kombinatorischen Bibliotheken, die sich eng an der Struktur der zur Kalibrierung verwendeten Moleküle orientieren, mag dieses Problem noch überschaubar bleiben. Werden jedoch beispielsweise Datenbanken mit mehreren Millionen kommerziell verfügbaren und hochdiversen Verbindungen gescreent, ist davon auszugehen, daß es sich bei der Mehrzahl um Outlier handelt und das QSAR-Modell gar nicht angewendet werden darf.

Kapitel 3

Methoden und Ergebnisse

Eine neue wissenschaftliche Wahrheit pflegt sich nicht in der Weise durchzusetzen, daß ihre Gegner überzeugt werden und sich als bekehrt erklären, sondern vielmehr dadurch, daß die Gegner allmählich aussterben und daß die heranwachsende Generation von vornherein mit der Wahrheit vertraut gemacht wird.
—Max Planck, Physiker und Begründer der Quantentheorie, † 1947

Im Rahmen dieser Dissertation wurden Forschungsarbeiten an verschiedenen Problemen im Bereich der *in-silico*-Wirkstoffentwicklung durchgeführt. Die Vorgehensweise in den einzelnen Projekten, die entwickelten Methoden und deren Validierung sowie die erzielten Ergebnisse werden in den folgenden Abschnitten vorgestellt.

3.1 Identifizierung von Outliern

Ebenso wie der Bergwetterbericht für die Alpen nicht für die Vorhersage des Wetters an der Nordsee taugt, kann ein mit einer bestimmten Verbindungsklasse kalibriertes QSAR-Modell nicht auf gänzlich andersartige Strukturen angewandt werden. Solche Objekte, die weit außerhalb des Kalibrierdatenraums liegen und für die deshalb prinzipiell keine Vorhersagen des Modells sinnvoll sind, heißen *Prediction Outlier*.

In den folgenden Abschnitten wird zunächst der Einfluß von Outliern auf den Vorhersagefehler des Modells deutlich gemacht, dann eine neue Metho-

de zur Identifizierung von Outliern vorgestellt, und schließlich werden die damit erzielten Ergebnisse präsentiert.

3.1.1 Distanzabhängigkeit des Vorhersagefehlers

Aus der Literatur über lineare Regression ist bekannt, daß der Erwartungswert des Vorhersagefehlers eines Objekts mit seiner Distanz zum Kalibrierdatenraum anwächst.[113,36,114] Der durch ein QSAR-Modell vorhergesagte Wert für die biologische Aktivität eines Moleküls ist also umso weniger vertrauenswürdig, je weiter entfernt vom chemischen Raum der Trainingsdaten dieses Molekül liegt.[115] Mathematisch läßt sich zeigen, daß der mittlere erwartete Vorhersagefehler $\overline{\text{MSEP}}$ im Fall der multiplen linearen Regression proportional zur Mahalanobis-Distanz der Testdaten MD_{Test} ist:[116]

$$\overline{\text{MSEP}} = \sigma^2 \cdot \left(1 + \frac{1}{n_{\text{Train}}} + \frac{p-1}{n_{\text{Train}}} \cdot \frac{\text{MD}_{\text{Test}}}{\text{MD}_{\text{Train}}}\right) \tag{3.1}$$

Das Regressionsmodell ist demnach nicht in der Lage, weit über seinen Strukturraum hinaus zu extrapolieren, ohne daß große Fehler wahrscheinlich werden. Mit anderen Worten ist es nicht möglich, zuverlässige Vorhersagen für Verbindungen zu treffen, die sich deutlich von den Trainingsmolekülen unterscheiden. Es gilt also, solche Outlier zu identifizieren und zu diesem Zweck einen Schwellenwert zu definieren, anhand dessen entschieden werden kann, ob ein Molekül „weit entfernt“ und damit „deutlich verschieden“ ist.

Um den Effekt des wachsenden Vorhersagefehlers bei steigender Distanz der Testobjekte zu zeigen, wurden Simulationen auf der Grundlage von drei verschiedenen Datensätzen (Sol, logP, DHODH, siehe Anhang A.1, Seite 139) berechnet. Jeder dieser Datensätze wurde zunächst zufällig in 2/3 Trainingsdaten und 1/3 Testdaten aufgeteilt. Aus den Trainingsdaten wurde ein PCR-Modell erstellt, wobei zur Validierung und Bestimmung der Modellparameter die L50%O-Kreuzvalidierung verwendet wurde. Der optimale Rang des Modells wurde als diejenige Anzahl von Hauptkomponenten gewählt, die im niedrigsten kreuzvalidierten Vorhersagefehler ($\text{RMSEP}_{\text{CV-50\%}}$) resultierte.

Die Testdaten wurden zunächst entsprechend den Trainingsdaten zentriert und skaliert, dann in den Hauptkomponentenraum des erstellten Modells projiziert. Anschließend wurden die Aktivitätswerte der Testdaten durch das Modell vorhergesagt. Das Residuum $\hat{e}_{\text{Test}}$, also die Abweichung zwischen Vorhersagewert und experimentellem Wert, wurde für jedes Testobjekt wie folgt berechnet:

$$\hat{e}_{\text{Test}} = \sqrt{(y_{\text{Test}} - \hat{y}_{\text{Test}})^2} \tag{3.2}$$

Außerdem wurde zwischen allen Testobjekten und allen Trainingsobjekten eine euklidische Distanzmatrix im Datenraum des Modells berechnet. Im allgemeinen wird die Distanz zwischen einem Objekt und seinem nächstgelegenen Nachbarobjekt als Nächste-Nachbar-Distanz (engl. *Nearest neighbour distance*, NND) bezeichnet. Aus der berechneten Distanzmatrix kann für jedes Testobjekt die Distanz zu seinem nächstgelegenen Trainingsobjekt abgelesen werden; diese spezielle Nächste-Nachbar-Distanz wird im folgenden als $\text{NND}_{\text{Te/Tr}}$ bezeichnet. Für jedes Testobjekt lag also schließlich ein Wertepaar aus der Distanz $\text{NND}_{\text{Te/Tr}}$ und dem Residuum $\hat{e}_{\text{Test}}$ vor.

Zur Überprüfung der angenommenen Korrelation zwischen Distanz und Residuum wurden die $\text{NND}_{\text{Te/Tr}}$ zunächst mittels einer Histogrammfunktion in zehn gleich große Kategorien (engl. *Bins*) eingeteilt. Anschließend wurde für jede Kategorie der Mittelwert der enthaltenen Residuen bestimmt. Als Ergebnis erhält man ein mittleres Residuum $\overline{\hat{e}}_{\text{Test}}$ für ein gewisses Distanzintervall (Bin).

Die gesamte Prozedur wurde 5000 Mal wiederholt, wobei sich durch die zufällige Unterteilung in Test- und Trainingsdaten jeweils leicht unterschiedliche Resultate ergaben; darüber wurde erneut der Mittelwert gebildet. In Abb. 3.1 ist das mittlere Residuum pro Bin gegen die Anzahl der Bins (10) aufgetragen. Das Ergebnis zeigt, daß die beschriebene Simulation die erwartete Korrelation des Vorhersagefehlers mit der Distanz zwischen Test- und Trainingsdaten $\text{NND}_{\text{Te/Tr}}$ auch für die PCR bestätigt.

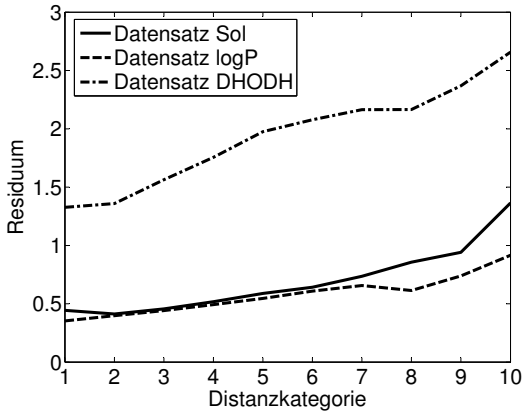


Abbildung 3.1 Der Vorhersagefehler (mittleres Residuum) eines Testobjekts korreliert mit dessen Distanz bezüglich der Kalibrierobjekte $NND_{Te/Tr}$. Der Graph zeigt das gemittelte Ergebnis von 5000 Simulationen.

3.1.2 ODD (*Outlier Detection by Distance towards training data*)

Ausgehend von der Erkenntnis, daß die Distanz zwischen einem Testobjekt und seinem nächstgelegenen Trainingsobjekt offenbar direkten Einfluß auf den Erwartungswert des Residuums hat, wurde ein neues Verfahren zur Outlier-Identifizierung entwickelt: Die Methode ODD (*Outlier Detection by Distance towards training data*) benutzt eben dieses Maß der Nächste-Nachbar-Distanz $NND_{Te/Tr}$ zur Entscheidung, ob ein Testobjekt „zu weit“ außerhalb des Trainingsdatenraums liegt, als daß verläßliche Vorhersagen möglich wären.

3.1.2.1 Grundannahmen bei der Entwicklung von ODD

Die Berechnung von $NND_{Te/Tr}$ liefert ein Maß dafür, wie weit ein Testobjekt im Datenraum vom nächstgelegenen Trainingsobjekt entfernt ist. Anhand dieses Werts allein kann jedoch noch nicht entschieden werden, ob das Testobjekt ein Outlier ist oder nicht. Dafür ist vielmehr die Definition eines Grenzwerts (engl. *Cut-off*) notwendig.

Die Methode ODD folgt hier der Idee, daß die Nächste-Nachbar-Distanz der Trainingsobjekte untereinander (im folgenden $NND_{Tr/Tr}$) auch ein Maß für die maximal zulässige Nächste-Nachbar-Distanz zwischen Test- und Trai-

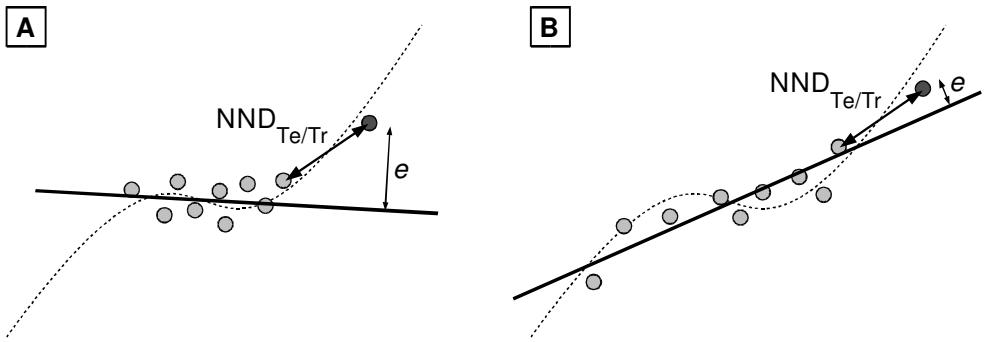


Abbildung 3.2 (A) Die auf einen engen Bereich begrenzten Trainingsobjekte (hell) folgen einem zugrundeliegenden funktionellen Zusammenhang (gestrichelte Kurve) und können mit einem linearen Modell (fette Linie) beschrieben werden. Ein Testobjekt (dunkel) mit Abstand $NND_{Te/Tr}$ zum nächstgelegenen Trainingsobjekt weist hier bereits ein hohes Residuum e auf; das Modell ist also nicht imstande zu extrapolieren. Spannen die Trainingsobjekte dagegen einen weiteren Raum auf (B), so ist das Modell weniger stark begrenzt; ein im selben Abstand $NND_{Te/Tr}$ gelegenes Testobjekt kann nun mit guter Genauigkeit vorhergesagt werden, sein Residuum e ist klein.

ningsobjekten ($NND_{Te/Tr}$) ist. Wird nämlich ein Modell auf Grundlage von zueinander sehr ähnlichen Trainingsobjekten erstellt, so umfaßt es nur einen sehr begrenzten Datenraum. Die erfolgreiche Validierung des Modells bestätigt somit lediglich seine lokal stark begrenzte Gültigkeit, während die funktionellen Zusammenhänge für den darüber hinausgehenden Datenraum nicht abschätzbar sind. Das Modell läuft dann Gefahr, mit seinem vereinfachenden linearen Ansatz nur einen engen lokalen Zusammenhang des tatsächlichen Systems zu modellieren.

Zeigt der Trainingsdatensatz dagegen eine hohe Diversität, so spannen die Objekte einen weiten Datenraum auf, d. h. die $NND_{Tr/Tr}$ ist groß. Ist dennoch mit all diesen Objekten die erfolgreiche Kalibrierung und Validierung eines Modells möglich, so kann von seiner entsprechend breiteren Gültigkeit ausgegangen werden; das Modell besitzt also eine größere Fähigkeit zur Extrapolation (siehe Abb. 3.2).

Der ODD-Algorithmus geht bei der Definition des Grenzwerts davon aus, daß für ein Testobjekt eine prinzipielle Vorhersagbarkeit dann gewährleistet ist, wenn seine $NND_{Te/Tr}$ nicht größer ist als die maximale* $NND_{Tr/Tr}$ innerhalb des Trainingsdatensatzes. Liegt nämlich das Trainingsobjekt mit der

*genauer: 90. Perzentil aller $NND_{Tr/Tr}$, siehe 3.1.2.2

höchsten $NND_{Tr/Tr}$ etwas abseits der übrigen Trainingsdaten und wird durch die Validierung dennoch die Gültigkeit des Modells auch für dieses Objekt bestätigt, so sollte auch für ein ebenso weit abseits liegendes Testobjekt eine Vorhersage möglich sein.

3.1.2.2 ODD-Algorithmus

Im folgenden wird der ODD-Algorithmus zur Identifizierung von Outliern für die Anwendung im Rahmen eines PCR-Modells beschrieben. Es wird eine Repräsentation der Trainingsdaten im Hauptkomponentenraum sowie eine entsprechende Projektion der Testdaten in dasselbe Koordinatensystem vorausgesetzt. Das beschriebene Verfahren ist jedoch unabhängig von der Modellierungstechnik und kann deshalb in analoger Weise auch beispielsweise bei PLS-Modellen angewendet werden. Eine schematische Darstellung des Algorithmus zeigt Abb. 3.3.

Im ersten Schritt berechnet der Algorithmus eine euklidische Distanzmatrix innerhalb des Trainingsdatensatzes. Gehen in das betrachtete QSAR-Modell die ersten q Hauptkomponenten ein, so erfolgt auch die Berechnung der Distanzmatrix im q-dimensionalen Datenraum. Die Distanzmatrix beschreibt dann für jedes Molekül die Abstände zu allen anderen Molekülen in diesem Raum.

Der jeweils kleinste Abstandswert eines Moleküls (also das Minimum innerhalb einer Zeile der Distanzmatrix, abgesehen vom Diagonalelement) ist seine Nächste-Nachbar-Distanz $NND_{Tr/Tr}$. Diese wird für jedes Molekül ermittelt, so daß bei einer Datensatzgröße von n Trainingsobjekten schließlich n $NND_{Tr/Tr}$-Werte vorliegen. Das 90. Perzentil* aller $NND_{Tr/Tr}$ wird als Cut-off c gespeichert.

Anschließend erfolgt die Berechnung einer Distanzmatrix zwischen allen Test- und allen Trainingsobjekten, wiederum im q-dimensionalen Hauptkomponentenraum. Auch hier wird für jedes Testobjekt das Minimum seiner Distanzen bestimmt, also seine Nächste-Nachbar-Distanz $NND_{Te/Tr}$. Dieser Wert wird mit dem zuvor definierten Cut-off c verglichen. Auf diese Weise

*Definition Perzentil: x Prozent aller Werte sind kleiner als das x-te Perzentil; beispielsweise ist das 50. Perzentil unter der Bezeichnung Median bekannt.

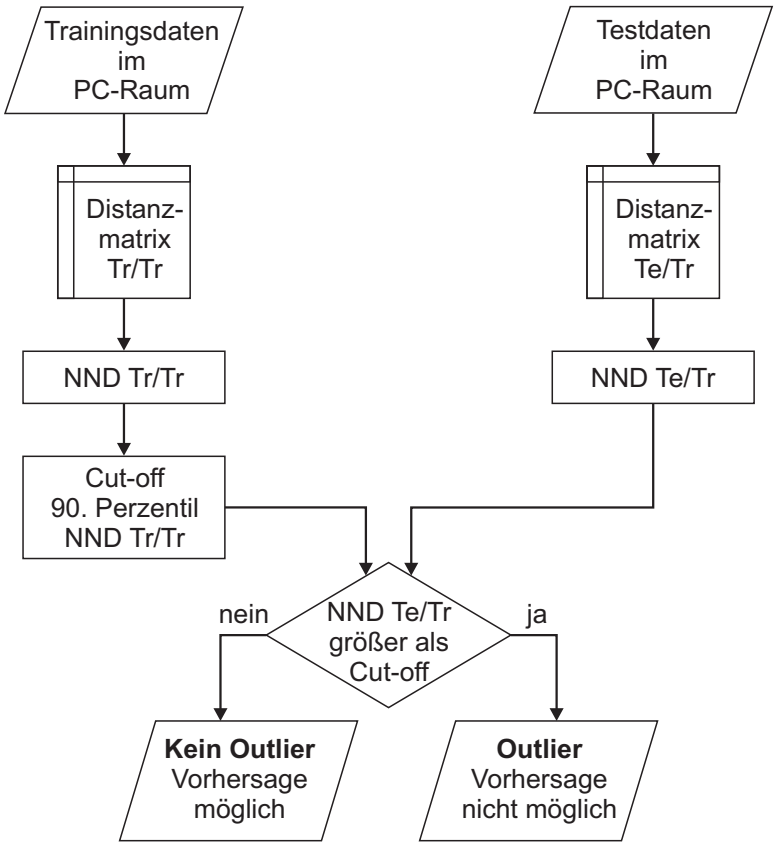


Abbildung 3.3 Schematische Darstellung des ODD-Algorithmus. Aus der Distanzmatrix der Trainingsdaten untereinander wird der Cut-off c bestimmt. Mit Hilfe der Distanzmatrix zwischen Test- und Trainingsobjekten wird der $\mathrm{NND}_{\mathrm{Te/Tr}}$-Wert eines jeden Testobjekts ermittelt und mit dem Cut-off c verglichen. Dieser Vergleich entscheidet darüber, ob das Objekt als Outlier angesehen wird oder nicht.

wird für jedes Testobjekt überprüft, ob es ein Outlier ist; es gilt:

$$\mathrm{NND}_{\mathrm{Te/Tr}} > c \quad \Rightarrow \quad \text{Testobjekt ist ein Outlier.} \tag{3.3}$$

$$\mathrm{NND}_{\mathrm{Te/Tr}} \leq c \quad \Rightarrow \quad \text{Testobjekt ist kein Outlier.} \tag{3.4}$$

Wie Abb. 3.4 zeigt, kann der Cut-off c im zweidimensionalen Fall ($q = 2$) als der Radius eines Kreises aufgefasst werden, der um jedes Trainingsobjekt geschlagen wird. Innerhalb aller von diesen Kreisen eingeschlossenen Berei-

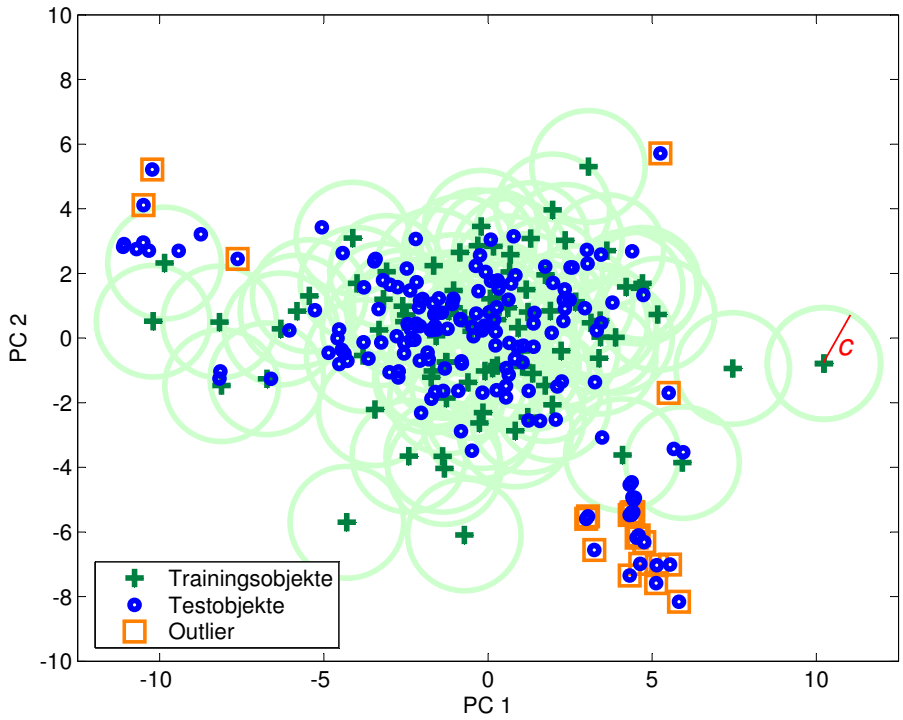


Abbildung 3.4 Darstellung der Methode ODD an einem exemplarischen Datensatz im zweidimensionalen Hauptkomponentenraum (PC1, PC2: erste bzw. zweite Hauptkomponente). Um jedes Trainingsobjekt wird ein Kreis mit Radius c (siehe Text) geschlagen. Die von den Kreisen eingeschlossene Fläche beschreibt den Datenraum, innerhalb dessen Vorhersagen möglich sind. Alle Testobjekte, die außerhalb dieser Region liegen, gelten als Outlier.

che sind Vorhersagen möglich; dagegen sind Testobjekte, die nicht innerhalb eines Kreises liegen, als Outlier anzusehen.

Das als Cut-off c verwendete Perzentil aller $\mathrm{NND}_{\mathrm{Tr/Tr}}$ ist der einzige justierbare Parameter der Methode. Die Verwendung nicht der maximalen $\mathrm{NND}_{\mathrm{Tr/Tr}}$ als Cut-off (wie eingangs in 3.1.2.1 vereinfachend beschrieben), sondern eines etwas kleineren Werts mindert den übergroßen Einfluß von Trainingsobjekten, die abseits der Hauptmenge der Trainingsdaten liegen. Als empirischer Standardwert hat sich hier das 90. Perzentil ($z = 0.9$) bewährt. Je kleiner z gewählt wird, umso geringer wird der Radius der in Abb. 3.4 dargestellten Kreise um die einzelnen Kalibrierobjekte, d. h. umso empfindlicher wird die Outlier-Detektion. In gewissem Sinn entspricht $1 - z$ $\{z = (0, 1)\}$ der Irrtumswahrscheinlichkeit der klassischen mathematischen Statistik.

Die Implementierung von ODD erfolgte in MATLAB R13[117].

Vorbedingung Aufgrund der Definition des Cut-offs aus dem Trainingsdatensatz selbst heraus muß natürlich die Bedingung erfüllt sein, daß das Modell für alle Trainingsobjekte Gültigkeit besitzt. Dies kann (und wird in jeder gewissenhaften Validierung) jedoch leicht überprüft werden, etwa anhand der kreuzvalidierten Residuen: Jedes Trainingsobjekt, das ein kreuzvalidiertes Residuum $\hat{e}_{CV} > 3 \cdot \text{RMSEP}_{CV}$ aufweist, sollte in der Regel aus dem Trainingsdatensatz eliminiert werden. Offensichtlich kann nämlich für dieses Trainingsobjekt keine zuverlässige Vorhersage getroffen werden; das Modell ist also für dieses Objekt nicht gültig. Da somit die Information dieses Objekts nicht in relevantem Maße in das Modell eingeht, trägt es auch nicht zur Charakterisierung des ihn umgebenden Datenraums bei. Folglich kann auch für nahegelegene Testobjekte nicht davon ausgegangen werden, daß sie ausreichend vom Modell erfaßt und damit vorhersagbar sind.

Verwandte Methode Eine Methode, die eine gewisse Ähnlichkeit zu ODD aufweist, wurde von TROPSHA für nichtlineare Nearest-Neighbour-QSPR-Modelle publiziert[41]. Ein wichtiger Unterschied liegt jedoch in der Definition des Cut-offs. Während sich dieser Grenzwert bei ODD der Charakteristik der Trainingsdaten intelligent anpaßt, verwendet TROPSHA lediglich einen konstanten Wert. Neben einer feineren Anpassung an den gegebenen Datensatz gewinnt ODD dadurch den Vorteil, für hochdimensionale Datensätze besser geeignet zu sein (siehe 3.1.3.3, Seite 70).

3.1.2.3 Validierung anhand des Vorhersagefehlers

Wenn der Erwartungswert des Vorhersagefehlers, wie in 3.1.1 (Seite 58) gezeigt, für Outlier höher ist als für „normale“ Testobjekte, dann muß umgekehrt die Eliminierung von vorhandenen Outliern zu einer Reduzierung des mittleren Vorhersagefehlers des Modells führen. Eben diese Hypothese wurde zur Validierung der Methode ODD überprüft.

Dazu wurde Datensatz logP (siehe Anhang A.1, Seite 139) zufällig in $^2/_3$ Trainingsdaten und $^1/_3$ Testdaten aufgeteilt. Auf Grundlage der Trainingsdaten wurde ein L50%O-CV-validiertes PCR-Modell erstellt; als optimaler Rang wurde diejenige Anzahl von Hauptkomponenten gewählt, die im niedrigsten kreuzvalidierten Vorhersagefehler ($\text{RMSEP}_{\text{CV-50\%}}$) resultierte. Die ent-

sprechend den Trainingsdaten zentrierten und skalierten Testdaten wurden in den Modell-PC-Raum projiziert und ihre abhängige Größe vorhergesagt. Aus der Differenz zwischen Vorhersagewert und wahrem Wert wurde dann der mittlere Vorhersagefehler RMSEP des Modells berechnet.

Anschließend wurden mit Hilfe der Methode ODD eventuell vorhandene Outlier aus dem Testdatensatz eliminiert. Aus den Vorhersagefehlern der verbliebenen Testobjekte wurde schließlich der mittlere Vorhersagefehler nach Outlier-Eliminierung $RMSEP_{el}$ berechnet.

Das Ergebnis von 50 solchen Simulationen zeigt Abb. 3.5. Es wird deutlich, daß für eine große Zahl von Experimenten der $RMSEP_{el}$ nur eine sehr geringe oder keine Verbesserung gegenüber dem RMSEP bringt. Die Outlier-Eliminierung hat also auf den Vorhersagefehler „normaler“ Datensätze kaum Einfluß. Bei Test-/Trainingsdaten-Aufteilungen dagegen, in denen sich offenbar extreme Outlier ergaben und die wiederum zu einem auffällig hohen RMSEP führten, konnte der Vorhersagefehler deutlich reduziert werden. Entsprechend der oben formulierten Hypothese ist der Ansatz, den ODD verfolgt, also valide: Die Outlier-Eliminierung ist ein probates Mittel, um hohe mittlere Vorhersagefehler bei Anwesenheit extremer Outlier zu verhindern.

3.1.3 Vergleich von ODD und Mahalanobis-Distanz

Die Mahalanobis-Distanz ist in der Chemometrie eines der etabliertesten Verfahren zur Identifizierung von Prediction Outliern.[34] Die Vorteile der neuen Methode ODD sollen in den folgenden Abschnitten durch einen Vergleich mit der MD deutlich gemacht werden. Es zeigt sich, daß ODD der MD sowohl bei der Anpassung an die Charakteristik des Trainingsdatensatzes als auch in der Effizienz der Outlier-Eliminierung überlegen ist; ebenso spricht die bessere Eignung für die Bearbeitung hochdimensionaler Datensätze für die Verwendung von ODD.

3.1.3.1 Details der Outlier-Identifizierung

Der offensichtlichste Vorteil von ODD gegenüber der Mahalanobis-Distanz ist die bessere Anwendbarkeit von ODD auf Datensätze mit eher ungewöhnlicher Gestalt (bezogen auf den Datenraum). Bei streng normalverteilten Da-

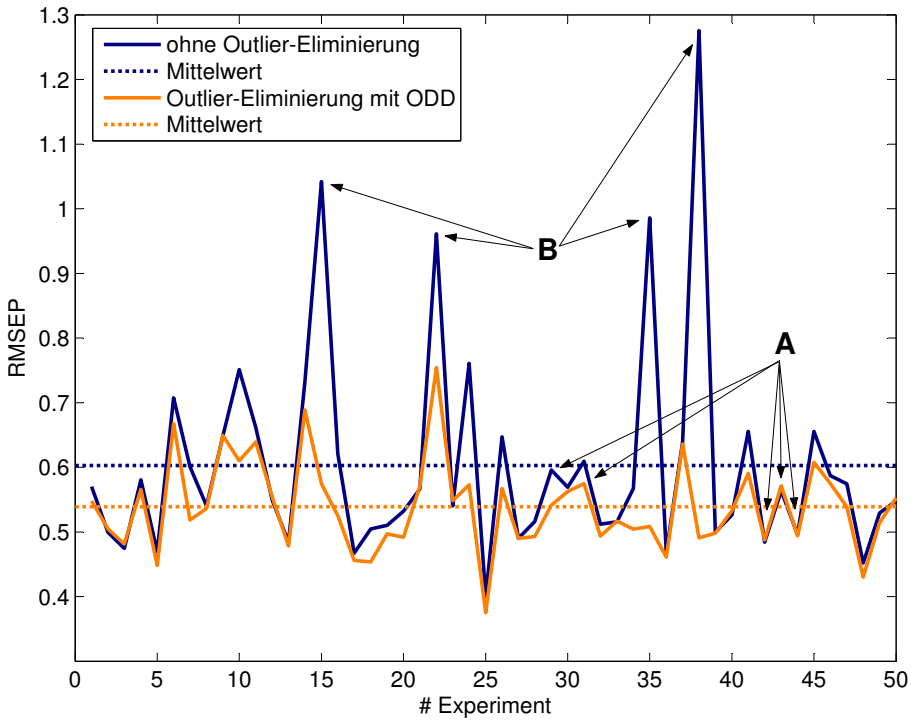


Abbildung 3.5 Reduzierung des mittleren Vorhersagefehlers (RMSEP) durch Eliminierung von Prediction Outliern mit Hilfe der Methode ODD. Gezeigt ist der RMSEP vor (blau) sowie der $RMSEP_{el}$ nach (orange) der Outlier-Eliminierung für 50 Simulationen. Die gestrichelten horizontalen Linien geben den Mittelwert über alle 50 Experimente an. Für „normale“ Datensätze (A) bringt die Outlier-Eliminierung nur eine geringe oder keine Verbesserung; sind jedoch offensichtlich extreme Outlier vorhanden (B), so führt deren Eliminierung zu einer deutlichen Reduzierung des Vorhersagefehlers.

ten, die im Hauptkomponentenraum zu einem Kalibrierdatenraum von hyperellipsoider Form führen, liefert auch die Outlier-Eliminierung per MD gute Ergebnisse. Darüber hinausgehend ist ODD aber in der Lage, auch Objekte korrekt zu behandeln, die außerhalb dieses auf den Schwerpunkt des Kalibrierdatenraums zentrierten Hyperellipsoids liegen. ODD geht nämlich davon aus, daß auch ein einzelnes Trainingsobjekt, das zwar abseits der Hauptmenge der Kalibrierdaten liegt, aber in der Kreuzvalidierung kein auffällig hohes Residuum aufweist, den umgebenden Datenraum zu charakterisieren vermag.

Diese Eigenschaft ist in Abb. 3.6 dargestellt: Das Urteil der MD, ob ein Testobjekt ein Outlier ist, stützt sich ausschließlich auf dessen Distanz zum Zen-

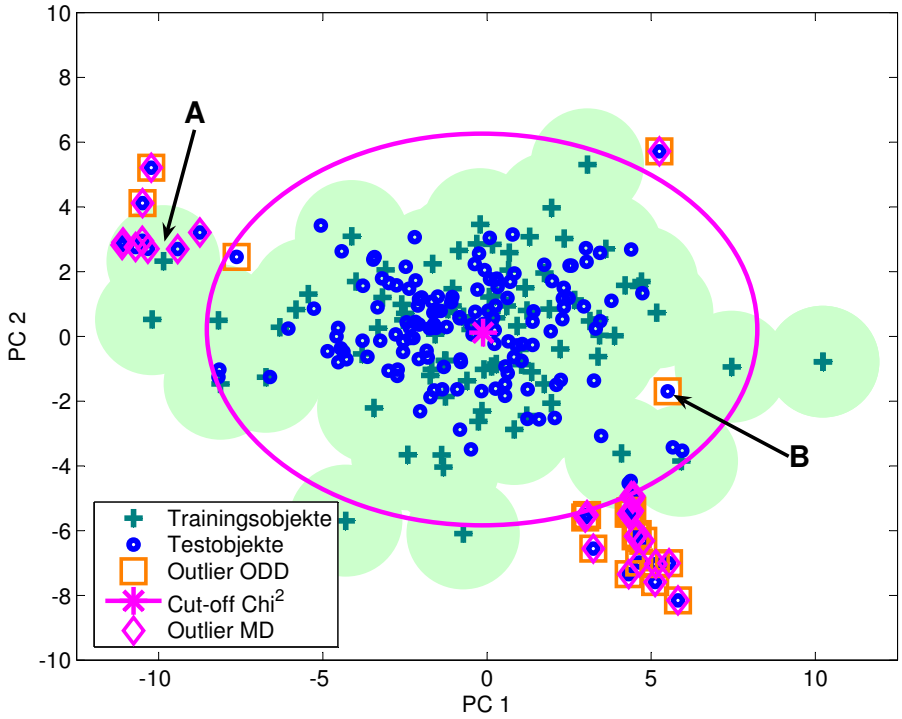


Abbildung 3.6 Methoden zur Outlier-Identifizierung: Vergleich zwischen ODD und Mahalanobis-Distanz (MD). ODD vermag sehr detailliert auf die Gestalt des Kalibrierdatenraums einzugehen. Dagegen orientiert sich die MD ausschließlich am Zentroiden ($*$) des Trainingsdatensatzes und des ihn umgebenden Hyperellipsoids, dessen Größe durch die χ^2-Verteilung definiert ist. Außerdem ist ODD in der Lage, auch Inlier (B) zu identifizieren.

troiden des Trainingsdatensatzes. Dadurch werden auch solche Objekte als Outlier identifiziert, die zwar abseits der Hauptmenge der Kalibrierdaten liegen, aber doch in unmittelbarer Nähe eines einzelnen oder ein kleinen Gruppe von Trainingsobjekten (Region A in Abb. 3.6). ODD dagegen geht sehr viel detaillierter auf die Gestalt des Trainingsdatensatzes ein und läßt sich nicht allein von der Distanz zu dessen Zentroiden leiten.

Identifizierung von Inliern Ebenfalls aus Abb. 3.6 ist weiterhin ersichtlich, daß ODD auch in der Lage ist, Inlier zu identifizieren (siehe 2.2.4, Seite 23). Dieser spezielle Fall tritt auf, wenn ein Testobjekt zwar innerhalb der äußersten Grenzen des Kalibrierdatenraums liegt, aber dennoch in einem schlecht charakterisierten Bereich — gewissermaßen inmitten eines weißen Flecks auf

der „Kalibrierdaten-Landkarte“ (Punkt B in Abb. 3.6). Aufgrund ihres zentroid-bezogenen Ansatzes ist dagegen die MD prinzipiell nicht in der Lage, Inlier zu erkennen.

3.1.3.2 Effizienz

Die positiven Auswirkungen der Outlier-Eliminierung auf den Vorhersagefehler wurden bereits in 3.1.2.3 beschrieben und in Abb. 3.5 graphisch dargestellt. Die Reduzierung des RMSEP allein ist jedoch kein ausreichendes Maß zur Beurteilung einer Methode zur Outlier-Eliminierung und zum Vergleich verschiedener Methoden untereinander.

Vielmehr kommt zum Kriterium der Minderung des Vorhersagefehlers noch die Forderung hinzu, daß nur tatsächliche Outlier eliminiert werden. Die Methode soll also möglichst wenige *False positives* (Objekt ist in Wirklichkeit gar kein Outlier) eliminieren. Umgekehrt würde nämlich ein Verfahren, das fast alle Testobjekte als Outlier eliminiert und nur die ganz zentral im Datenraum gelegenen beibehält, mit hoher Wahrscheinlichkeit zu einer sehr deutlichen RMSEP-Reduzierung führen; gleichwohl wäre ein solches Ergebnis in der Praxis natürlich irrelevant.

Es liegt also nahe, als Maß zur Bestimmung der Effizienz einer Outlier-Identifizierung sowohl die Reduzierung des Vorhersagefehlers als auch die Anzahl der eliminierten Objekte zu erfassen. Eine Methode erweist sich als umso besser, je stärker sie den Vorhersagefehler mindert *und* je weniger Testobjekte sie gleichzeitig als Outlier eliminiert. Die Effizienz E ist demnach als

$$E = \frac{\Delta\text{RMSEP}}{N_{\text{el}}} \tag{3.5}$$

zu definieren, wobei ΔRMSEP die Differenz zum RMSEP ohne Outlier-Eliminierung angibt und N_{el} die Anzahl der als Outlier eliminierten Testobjekte bezeichnet.

Analog zu 3.1.2.3 und der dort gezeigten Abb. 3.5 (Seite 67) wurde eine weitere Simulation durchgeführt, in der die Effizienz der Outlier-Eliminierung gemäß Gl. 3.5 berechnet wurde. Zum Vergleich wurde zusätzlich auch die Effizienz der MD bestimmt sowie die einer Adaption der von TROPSHA

beschriebenen Methode (siehe 3.1.2.2, Seite 65; im folgenden als „TkNN“* bezeichnet).

Gewissermaßen als Gegenprobe und zur Überprüfung des prinzipiellen Nutzens der Outlier-Identifizierung wurde auch die Eliminierung zufälliger Testobjekte simuliert. Dabei wurden jeweils genau so viele Testobjekte eliminiert, wie im selben Experiment auch die MD entfernt hatte; die Auswahl der eliminierten Objekte erfolgte jedoch rein zufällig. Demnach ist zu erwarten, daß die Effizienz dieser zufälligen Eliminierung gegen Null geht, da mit einer hohen Anzahl von False positives zu rechnen ist. Diese führen jedoch nicht zu einer Verbesserung von ΔRMSEP, sondern lediglich zu einer Erhöhung von N_{el} und mindern damit gemäß Gl. 3.5 die Effizienz.

Die Ergebnisse dieser Simulationen sind in Abb. 3.7 gezeigt. Es ist deutlich zu erkennen, daß die Methode ODD unter Berücksichtigung der genannten Kriterien die effizienteste Methode zur Outlier-Eliminierung darstellt, gefolgt von TkNN und MD. Die Effizienz der Eliminierung zufälliger Testobjekte liegt wie erwartet bei Null, was umgekehrt den Einsatz aufwendiger Techniken zur Outlier-Identifizierung rechtfertigt.

3.1.3.3 Dimensionsabhängigkeit

In 2.3.3 (Seite 36) wurde das Phänomen *The Curse of Dimensionality* vorgestellt, das die besonderen Eigenschaften hochdimensionaler Räume beschreibt. Ein für die hier präsentierten Arbeiten wichtiger Aspekt ist dabei die Dimensionsabhängigkeit von Distanzen: Mit steigender Dimensionalität des Datenraums wachsen die Distanzen zwischen benachbarten Objekten stark an. Im konkreten Fall hat dies zur Folge, daß sich beispielsweise die NND eines Objekts vergrößert, wenn statt des D-dimensionalen Hauptkomponentenraums der (D+1)-dimensionale betrachtet wird.

Die Änderung der Distanzen selbst stellt dabei jedoch nicht das eigentliche Problem dar, da sie für alle Objekte gleichermaßen gilt. Vielmehr muß gewährleistet sein, daß der Cut-off derselben Dimensionsabhängigkeit folgt wie die Distanz. Wächst nämlich der Cut-off mit steigender Dimensionalität langsamer als die $\mathrm{NND}_{\mathrm{Te/Tr}}$ eines Objekts, so wird dieses früher oder später allein

* „TROPSHA k-Nearest-Neighbours“

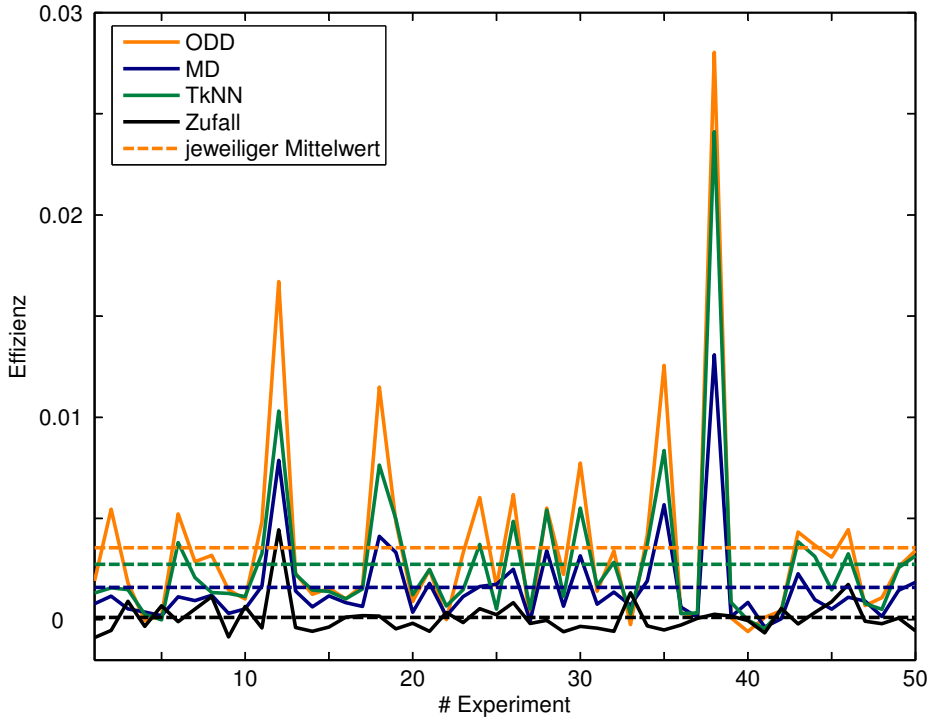


Abbildung 3.7 Effizienz der Outlier-Eliminierung: Vergleich der Methoden ODD, MD und TkNN. Die Effizienz beschreibt den Quotienten aus Verringerung des Vorhersagefehlers RMSEP einerseits und Anzahl eliminierter Outlier andererseits. Demnach ist eine Methode besonders effizient, wenn sie zu einer deutlichen Reduzierung des RMSEP führt und gleichzeitig wenige False positives liefert. Hier gezeigt sind 50 Simulationen für den Datensatz logP.

durch den Effekt sich vergrößernder Distanzen zum Outlier. In einem genügend hochdimensionalen Datenraum werden dann also auch durchschnittliche Objekte als Outlier identifiziert, d. h. die Zahl der False positives steigt.

Simulation Um diese Dimensionsabhängigkeit von Cut-off und Distanzmaß zu untersuchen, wurden auf Grundlage des Datensatzes Sol Objekte unterschiedlicher Dimensionalität simuliert. Dazu wurde für jede Dimensionalität $D = [1, \ldots, 45]$ folgende Vorgehensweise angewandt:

Der Datensatz wurde zunächst zufällig in $2/3$ Trainings- und $1/3$ Testdaten unterteilt. Nach Zentrierung und Autoskalierung wurden die Trainingsdaten durch Berechnung der SVD und Auswahl der ersten D Hauptkomponenten

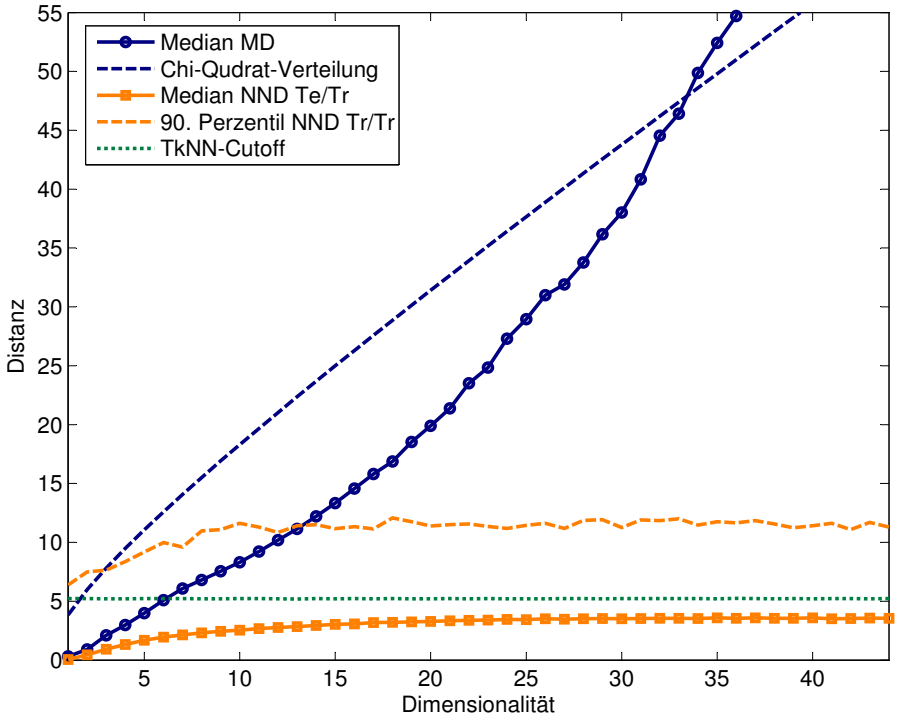


Abbildung 3.8 Dimensionsabhängigkeit der verschiedenen Distanzmaße und Cut-off-Definitionen, hier simuliert auf Grundlage des Datensatzes Sol. Besonders auffällig ist, daß die Kurve für die mittlere Mahalanobis-Distanz (Median MD) bei etwa D=34 von ihrem Cut-off (χ^2-Verteilung) geschnitten wird; ab dieser Dimensionalität werden also auch ganz gewöhnliche Objekte als Outlier identifiziert. Bei ODD tritt dieses Problem dagegen nicht auf; hier zeigen Distanzmaß und Cut-off einen übereinstimmenden Verlauf.

in den D-dimensionalen Hauptkomponentenraum transformiert. Die Testdaten wurden entsprechend in denselben PC-Raum projiziert. Dann wurde für jedes Testobjekt sowohl die MD als auch die $\mathrm{NND}_{\mathrm{Te/Tr}}$ bestimmt und der Median dieser Distanzen gespeichert. Zusätzlich wurden die jeweiligen Cut-offs berechnet, also der entsprechende Wert der χ^2-Verteilung und der Cut-off c gemäß der Beschreibung in 3.1.2.2 (Seite 62).

Diese Prozedur wurde für jede Dimensionalität D 200 Mal wiederholt (mit jeweils zufälliger Test-/Trainingsdatenaufteilung), so daß schließlich 200 Median-Werte für jedes Distanzmaß bzw. jede Cut-off-Definition vorlagen; über diese wurde der Mittelwert gebildet. Eine Auftragung dieser Mittelwerte gegen die Dimensionalität D zeigt Abb. 3.8.

Interpretation Wie erwartet ist für die MD ein exponentieller Anstieg der mittleren Distanzen bei steigender Dimensionalität zu erkennen. Der zugehörige Cut-off (der Wert der χ^2-Verteilung) wächst dagegen lediglich annähernd linear. Der kritische Punkt liegt bei einer Dimensionalität von etwa D=34, denn hier schneiden sich die Distanz- und die Cut-off-Kurve. Für den gezeigten Datensatz besitzen also ab D=34 selbst durchschnittliche (genau dem Median entsprechende) Testobjekte eine MD, die oberhalb des Cut-offs liegt. Demzufolge werden selbst diese eigentlich unkritischen Testobjekte fälschlicherweise als Outlier identifiziert.

Der Graph der NND, wie sie ODD benutzt, zeigt dagegen einen anderen, eher asymptotischen Verlauf. Die Nächste-Nachbar-Distanz verhält sich bei steigender Dimensionalität also offenbar anders als die auf den Zentroiden bezogene MD. Während letztere kontinuierlich anwächst, stabilisiert sich die NND allmählich. Die Erklärung für diese Beobachtung liegt im Grad der Ähnlichkeit der Trainingsobjekte untereinander: Bestünde der Trainingsdatensatz nur aus Paaren zueinander sehr ähnlicher Objekten, so würde jedes dieser Paare im Datenraum durch zwei eng benachbarte Punkte repräsentiert. Bei steigender Dimensionalität würde die Distanz zwischen den Paaren deutlich schneller wachsen als die Distanz der beiden Punkte innerhalb des Paares. Daraus ergibt sich automatisch eine Stabilisierung der NND, die in dieser Überlegung durch die kleinste Distanz innerhalb eines Paares (und nicht zwischen zwei Paaren) gegeben ist.

Der entscheidende Faktor für die Dimensionsabhängigkeit der Outlier-Identifizierung besteht jedoch nicht im Verhalten des Distanzmaßes, sondern in dem des Cut-offs. Hier zeigt Abb. 3.8 für ODD und den bei dieser Methode verwendeten Grenzwert (90. Perzentil der $\mathrm{NND}_{\mathrm{Tr/Tr}}$) eine gute Übereinstimmung des qualitativen Verlaufs der Graphen. Der Cut-off folgt also in etwa derselben Gesetzmäßigkeit wie das Distanzmaß selbst. Damit ist die Voraussetzung für die dimensionsunabhängige Identifizierung von Outliern erfüllt.

Die TkNN-Methode nutzt — ebenso wie ODD — die NND als Distanzmaß. Als Cut-off ist jedoch eine Konstante vorgesehen. Daraus ergibt sich ebenfalls ein mit steigender Dimensionalität deutlich weniger kritischer Verlauf als bei der mit der MD verknüpften χ^2-Verteilung. Gerade bei niederdimensionalen Datensätzen paßt sich der TkNN-Cut-off jedoch dem Distanzmaß deutlich schlechter an als der für ODD definierte Grenzwert.

Die Verschiebung des Cut-offs in y-Richtung spielt für die angestellten Betrachtungen zur Dimensionsabhängigkeit keine Rolle. Sie beschreibt lediglich die Empfindlichkeit der Methode, also wie eng die Grenzen der Extrapolationsmöglichkeiten des Modells gezogen werden. Durch eine veränderte Parametrisierung des Cut-offs kann diese Empfindlichkeit bei allen drei dargestellten Methoden leicht angepaßt werden, etwa durch Wahl eines anderen als des neunzigsten Perzentils bei ODD.

3.1.4 Beurteilung der Methode ODD

Die hier vorgestellte Methode ODD bietet ein zuverlässiges Verfahren zur Identifizierung von Outliern. Die Abgrenzung zu etablierten Verfahren liegt hauptsächlich in der auf Nächster-Nachbar-Distanzen basierenden Erkennung sowie des sich aus den Trainingsdaten selbst ergebenen Cut-offs. Die neue Methode bietet vor allem die Vorteile

- der besseren Berücksichtigung der Charakteristik („Gestalt“) des Trainingsdatensatzes,
- der hohen Effizienz bezüglich der Verbesserung des Vorhersagefehlers einerseits und der geringen Anzahl von False positives andererseits sowie
- der weitgehenden Unabhängigkeit von der Dimensionalität des betrachteten Datenraums.

Dabei ist der Rechenaufwand für die Bereitstellung der Distanzmatrizen zwar höher als beispielsweise der für die Berechnung der Varianz-Covarianz-Matrix (MD), bleibt jedoch auch für umfangreichere Datensätze noch im akzeptablen Bereich. Dies ermöglicht den Einsatz von ODD als Routineinstrument zur Outlier-Eliminierung in der QSAR-Analyse.

3.2 Modellstabilisierung durch Ensemble-Techniken

Ein weiterer Ansatz zur Verbesserung der Vorhersageergebnisse ist die Verwendung von Ensembles. Statt aus einem einzelnen Trainingsdatensatz ein einzelnes Modell zu kalibrieren, wird dabei ein ganzes Ensemble von Datensätzen und Modellen verwendet. Der Mittelwert über alle damit getroffenen Vorhersagen erweist sich häufig als stabiler als die Vorhersage der separat betrachteten Einzelmodelle.

In der Literatur finden sich verschiedene Studien, die den Einfluß einzelner Ensemble-Methoden auf die Vorhersagegenauigkeit beschreiben; dort wurden sowohl lineare und nichtlineare Regressionstechniken als auch Klassifizierer untersucht.[118–124] Darüber hinausgehend werden in den nachfolgend vorgestellten Untersuchungen u. a. die Ensemble-Techniken Konvexe Pseudodaten und Noise addition anhand realer QSAR-Datensätze untersucht. Die in dieser Arbeit gezeigten Studien bieten damit einen praxisrelevanten Vergleich eines breiten Spektrums von Methoden zur Ensemble-Generierung.

Wie schon bei der Validierung der Methode ODD zur Outlier-Eliminierung diente auch hier die Reduzierung des mittleren Vorhersagefehlers RMSEP als Maßstab. Alle Berechnungen wurden mit MATLAB R13[117] durchgeführt.

3.2.1 Anwendung auf PCR-Modelle

Um den Nutzen der Ensemble-Techniken zu untersuchen, wurde zunächst ein einzelnes PCR-Modell kalibriert und damit anhand eines Testdatensatzes direkt der mittlere Vorhersagefehler $\mathrm{RMSEP}_{\mathrm{dir}}$ des Einzelmodells bestimmt. Anschließend wurde zum Vergleich der Vorhersagefehler bei Anwendung einer bestimmten Ensemble-Technik $\mathrm{RMSEP}_{\mathrm{ens}}$ berechnet. Nach bisherigem Kenntnisstand sollte die Verwendung des Ensembles zu einer Stabilisierung der Vorhersage führen, d. h. es wird erwartet, daß $\mathrm{RMSEP}_{\mathrm{ens}} < \mathrm{RMSEP}_{\mathrm{dir}}$. Konkret folgten die Simulationen folgendem Ablauf:

Der Originaldatensatz wurde zunächst zufällig in 1/3 Trainingsdaten und 2/3 Testdaten unterteilt. Der geringe Anteil an Trainingsdaten (üblicherweise wird ein genau umgekehrtes Verhältnis verwendet) wurde bewußt gewählt, um einerseits die Anforderungen an die Modellkalibrierung zu erhöhen und andererseits den Vorhersagefehler anhand eines größeren Testdatensatzes be-

stimmen zu können. Beides trägt dazu bei, den Effekt der Ensembles sowie die Unterschiede zwischen den verschiedenen Techniken zur Ensemble-Generierung deutlicher hervorzuheben. Anschließend wurde mit den Trainingsdaten ein PCR-Modell erstellt, zu dessen Validierung eine L50%O-CV mit $3 \cdot n$ Splits in Konstruktions- und Validierdaten ($^2/_3 : {}^1/_3$) verwendet wurde; der optimale Rang wurde auf Grundlage des minimalen $\mathrm{RMSEP}_{\mathrm{CV}}$ bestimmt. Mit diesem Einzelmodell wurden die Y-Werte des Testdatensatzes vorhergesagt und daraus der mittlere quadrierte Vorhersagefehler $\mathrm{MSEP}_{\mathrm{dir}}$ berechnet.

Im zweiten Teil der Simulation wurde aus den Trainingsobjekten ein Ensemble von k Trainingsdatensätzen erzeugt. Dazu wurde jeweils eine der Methoden Noise addition, Konvexe Pseudodaten, Bagging oder Subsampling angewendet. Auf Grundlage der so gestörten Trainingsdaten wurden insgesamt k PCR-Modelle (wie oben) erstellt. Mit jedem dieser k Modelle wurden die Y-Daten des Testdatensatzes vorhergesagt, so daß schließlich k Vektoren $\hat{\mathbf{y}}_i$ $(i = 1 \ldots k)$ mit Vorhersagewerten vorlagen. Aus dem Ensemble-Mittelwert

$$\hat{y}_{i,\mathrm{ens}} = \frac{1}{k} \cdot \sum_{j=1}^{k} \hat{y}_{i,j} \tag{3.6}$$

($\hat{y}_{i,j}$ ist hier der mit dem j-ten Modell des Ensembles generierte Vorhersagewert des i-ten Testobjekts) wurde dann die Fehlerquadratsumme

$$\mathrm{PRESS}_{\mathrm{ens}} = \sum_{i=1}^{n_{\mathrm{Test}}} (y_i - \hat{y}_{i,\mathrm{ens}})^2 \tag{3.7}$$

(y_i und $\hat{y}_{i,\mathrm{ens}}$ beziehen sich hier auf die Testdaten) sowie der mittlere quadrierte Vorhersagefehler des Ensembles $\mathrm{MSEP}_{\mathrm{ens}}$ berechnet:

$$\mathrm{MSEP}_{\mathrm{ens}} = \frac{1}{n_{\mathrm{Test}}} \cdot \mathrm{PRESS}_{\mathrm{ens}} \tag{3.8}$$

Die gesamte Prozedur wurde s Mal wiederholt und über die Ergebnisse (jeweils s Werte für $\mathrm{MSEP}_{\mathrm{dir}}$ und $\mathrm{MSEP}_{\mathrm{ens}}$) der Mittelwert gebildet. Daraus

können die Vorhersagefehler $\mathrm{RMSEP}_{\mathrm{dir}}$ und $\mathrm{RMSEP}_{\mathrm{ens}}$ berechnet werden:

$$\mathrm{RMSEP}_{\mathrm{dir}} = \sqrt{\frac{1}{s} \cdot \sum_{i=1}^{s} \mathrm{MSEP}_{\mathrm{dir,i}}} \tag{3.9}$$

$$\mathrm{RMSEP}_{\mathrm{ens}} = \sqrt{\frac{1}{s} \cdot \sum_{i=1}^{s} \mathrm{MSEP}_{\mathrm{ens,i}}} \tag{3.10}$$

Der Rechenaufwand der beschriebenen Simulation wird dabei durch die beiden Parameter Ensemblegröße k und Anzahl Simulationen s bestimmt, da insgesamt $s \cdot (k \cdot 4 + 1)$ (vier Ensemble-Methoden plus direktes Einzelmodell) Modelle erstellt werden müssen.

3.2.1.1 Voruntersuchungen zu Ensemblegröße und Anzahl Simulationen

Die benötigte Rechenzeit steigt bei einer Erhöhung der Ensemblegröße und der Anzahl der Simulationen stark an. Deshalb wurde der Einfluß dieser beiden Parameter zunächst in einer Voruntersuchung anhand zweier Datensätze überprüft. Für den Datensatz logP (siehe A.1, Seite 139) wurde eine Simulation mit s=100, k=100 und s=100, k=250 durchgeführt, für den Datensatz DHFR mit s=40, k=100 und s=100, k=100 sowie s=100, k=250.

Abb. 3.9 zeigt, daß die Erhöhung der Ensemblegröße bei logP von k=100 auf k=250 keinen Einfluß auf den resultierenden RMSEP hat. Beim Datensatz DHFR weichen die Ergebnisse der beiden Simulationen im Mittel ebenfalls nur um etwa ein Prozent voneinander ab. Auch wenn diese stichprobenartige und damit nicht systematische Voruntersuchung den Einfluß der Parameter s und k nicht abschließend klären kann, legt sie im Sinne einer Kosten-Nutzen-Rechnung dennoch nahe, daß sich übertrieben hohe Werte für Ensemblegröße und Simulationszahl nicht lohnen. Für die Simulationen mit den übrigen Datensätzen wurde deshalb s=100, k=100 gewählt.

3.2.1.2 Ergebnisse der Simulationen

Die Simulationen wurden wie beschrieben für die fünf Datensätze logP, Sol, DHFR, ACE und HEPT durchgeführt. Wie die Darstellung der Ergebnisse in

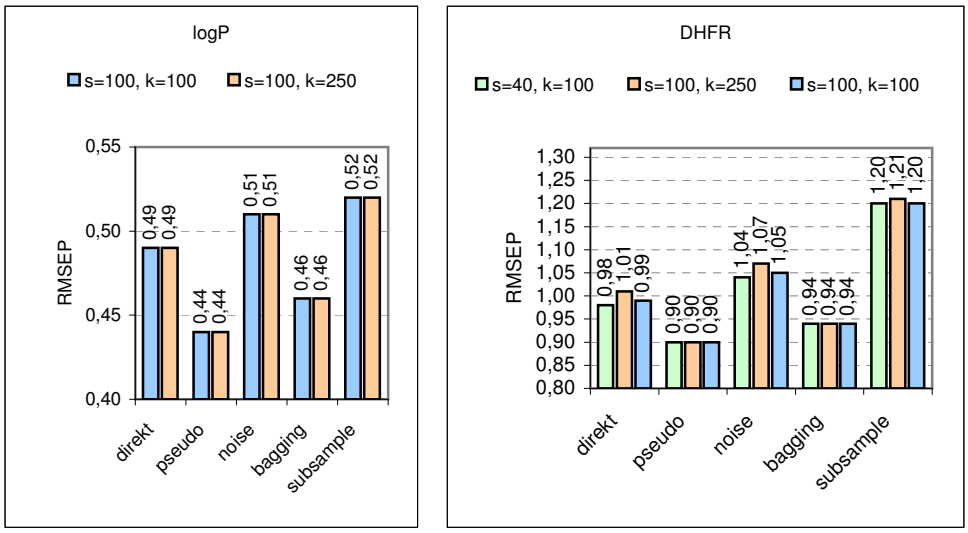


Abbildung 3.9 Einfluß der Ensemblegröße k und der Anzahl der Simulationen s auf den mittleren Vorhersagefehler RMSEP des Ensembles. Hier gezeigt ist eine Voruntersuchung an den Datensätzen logP und DHFR. Innerhalb des gewählten Parameterbereichs ergeben sich keine maßgeblichen Änderungen.

Tab. 3.2 (Seite 81) bzw. Abb. 3.10 (Seite 80) zeigt, wird für ACE praktisch keine Verbesserung des RMSEP erzielt, für Sol nur eine geringe. Die anderen drei Datensätze können jedoch vom Ensemble-Ansatz deutlich profitieren — allerdings nur bei der Verwendung bestimmter Methoden zur Generierung des Ensembles. Die prozentualen Reduzierungen des Vorhersagefehlers mit der jeweils besten Methode finden sich in Tab. 3.1.

Die Methode der konvexen Pseudodaten führt mit deutlichem Abstand zu den besten Ergebnissen. Auch durch Bagging ist eine Reduzierung des Vorhersagefehlers möglich. Die Noise addition dagegen zeigt kaum Auswirkungen, das Subsampling schließlich wirkt sich deutlich negativ auf den RMSEP aus.

Letzteres kann zum Teil durch die geringere Größe des Ensemble-Trainingsdatensatzes erklärt werden: Nur ein Drittel der Objekte des Originaldatensatzes bilden den Trainingsdatensatz des Einzelmodells; daraus werden beim Subsampling zufällige zwei Drittel für den Ensemble-Trainingsdatensatz entnommen. Somit stehen nur etwa 22 Prozent der Originaldaten für die Modellbildung zur Verfügung, gegenüber 33 Prozent beim Einzelmodell.

Tabelle 3.1 Prozentuale Verbesserung des Vorhersagefehlers RMSEP (gegenüber dem $RMSEP_{dir}$ des Einzelmodells) durch den Einsatz von Ensembles bei PCR-Modellen; angegeben ist jeweils die stärkste erzielte Verbesserung und die zugehörige(n) Ensemble-Methode(n) (bei ACE: keine nennenswerte Unterschiede zwischen den Methoden).

Datensatz	logP	HEPT	Sol	ACE	DHFR
Verbesserung (%)	10	15	5	1	9
Methode(n)	pseudo	pseudo	pseudo bagging	−/−	pseudo

Die Simulationen wurden daher für die Methode Subsampling wiederholt, diesmal jedoch mit einer Aufteilung der Originaldaten in 2/3 Trainings- und 1/3 Testdaten. Durch den vergrößerten Anteil von Trainingsobjekten fallen die Ergebnisse hier nicht mehr ganz so schlecht aus wie bei der ursprünglichen Datenaufteilung. Dennoch führt die Ensemble-Methode Subsampling weiterhin bei keinem der fünf Datensätze zu einer Verbesserung des RMSEP gegenüber dem Einzelmodell; im Durchschnitt verschlechtert sich der Vorhersagefehler um 1 Prozent.

Die Methode Noise addition wurde ebenfalls noch einmal mit einem Anteil von 2/3 Trainingsdaten untersucht. Hier bleiben die Ergebnisse gegenüber der ursprünglichen Aufteilung (1/3 Trainingsdaten) weitestgehend unverändert.

3.2.2 Anwendung auf PCR-Modelle mit Variablenselektion

Der Einsatz einer Variablenselektion im Zuge einer QSAR-Analyse führt häufig zu einer besseren Interpretierbarkeit und höheren Vorhersagekraft des Modells. Nichtsdestotrotz birgt ein solches Verfahren auch die Gefahr der Zufallskorrelation und Überoptimierung.[52] Unter diesen Rahmenbedingungen erweisen sich stabilisierende Ansätze wie die Ensemble-Techniken als besonders nützlich.

Um den Effekt von Ensemble-Techniken auf PCR-Modelle mit Variablenselektion zu untersuchen, wurden die im vorigen Abschnitt vorgestellten Simulationen wiederholt; vor der Modellerstellung wurde jedoch eine Variablenselektion durchgeführt. Wie in 2.2.6 beschrieben, wurde dabei als Suchalgorithmus die Tabu-Suche verwendet, als Gütefunktion diente eine L50%O-CV.

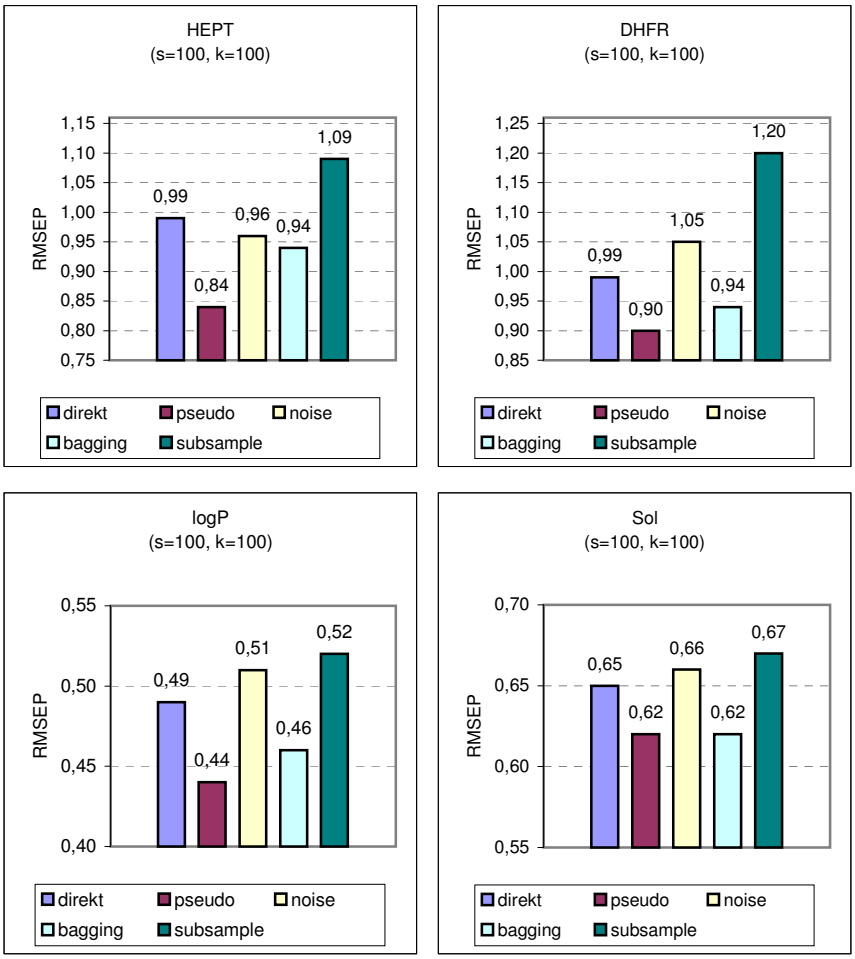


Abbildung 3.10 Darstellung des direkten Vorhersagefehlers (RMSEP) des Einzelmodells sowie der Ensemble-Vorhersagefehler der jeweiligen Methoden. Bei den Datensätzen HEPT, DHFR und logP führt die Verwendung bestimmter Ensemble-Techniken zur Reduzierung des RMSEP, bei Sol fällt die Verbesserung dagegen recht gering aus. Durch die Anwendung der Methoden Subsampling und Noise addition verschlechtert sich dagegen der Vorhersagefehler sogar. Bei allen vier Datensätzen liefert die Methode Konvexe Pseudodaten die besten Resultate. Für ACE (hier nicht gezeigt) ergab sich keine nennenswerte Differenz zwischen $RMSEP_{dir}$ und den einzelnen $RMSEP_{ens}$.

Tabelle 3.2 Vorhersagefehler $RMSEP_{dir}$ und $RMSEP_{ens}$ für die verschiedenen Datensätze. Der obere Teil der Tabelle zeigt die Ergebnisse für die Simulationen ohne Variablenselektion, der untere Teil bezieht sich auf die Simulationen mit Variablenselektion.

ohne Variablenselektion (s=100, k=100)					
Datensatz	logP	HEPT	Sol	ACE	DHFR
Direkt	0.49	0.99	0.65	1.68	0.99
Pseudo	0.44	0.84	0.62	1.67	0.90
Noise	0.51	0.96	0.66	1.66	1.05
Bagging	0.46	0.94	0.62	1.66	0.94
Subsampling	0.52	1.09	0.67	1.66	1.20
mit Variablenselektion (s=50, k=50)					
Datensatz	logP	HEPT	Sol	ACE	DHFR
Direkt	0.49	0.98	0.62	1.72	1.13
Pseudo	0.38	0.72	0.58	1.52	1.00
Noise	0.46	0.71	0.59	1.55	1.08
Bagging	0.41	0.72	0.58	1.51	1.00
Subsampling	0.57	0.73	0.61	1.57	1.10

Die Variablenselektion führt zu einer immensen Erhöhung des für die Simulation notwendigen Rechenaufwands, da zur Auswahl der Variablensubmenge eine Vielzahl von Modellen erstellt und validiert werden muß. Aus diesem Grund war eine entsprechende Reduzierung der Ensemblegröße auf k=50 sowie der Anzahl der Simulationen auf s=50 nötig. Die Rechenzeit betrug mit diesen Vorgaben auf einem Pentium4-Prozessor mit 2.4 GHz Taktfrequenz je nach Datensatz zwischen etwa 16 und 115 Stunden pro Simulation.

3.2.2.1 Ergebnisse der Simulationen

Die mittleren Vorhersagefehler der Simulationen mit Variablenselektion sind in Tab. 3.2 (unterer Teil) angegeben. Wie im vorangegangenen Abschnitt erfolgt eine graphische Darstellung in Abb. 3.11 sowie die numerische Angabe der prozentualen Verbesserung für die beste Methode in Tab. 3.3 (Seite 83).

Aufgrund der durch die Variablenselektion eingeführten Instabilität tritt die Wirkung der Ensembles erwartungsgemäß deutlicher hervor: Die prozentualen Verbesserungen, die mit einer Ensemble-Technik gegenüber dem Einzelmodell erreicht wurden, fallen stärker aus als bei den Modellen ohne

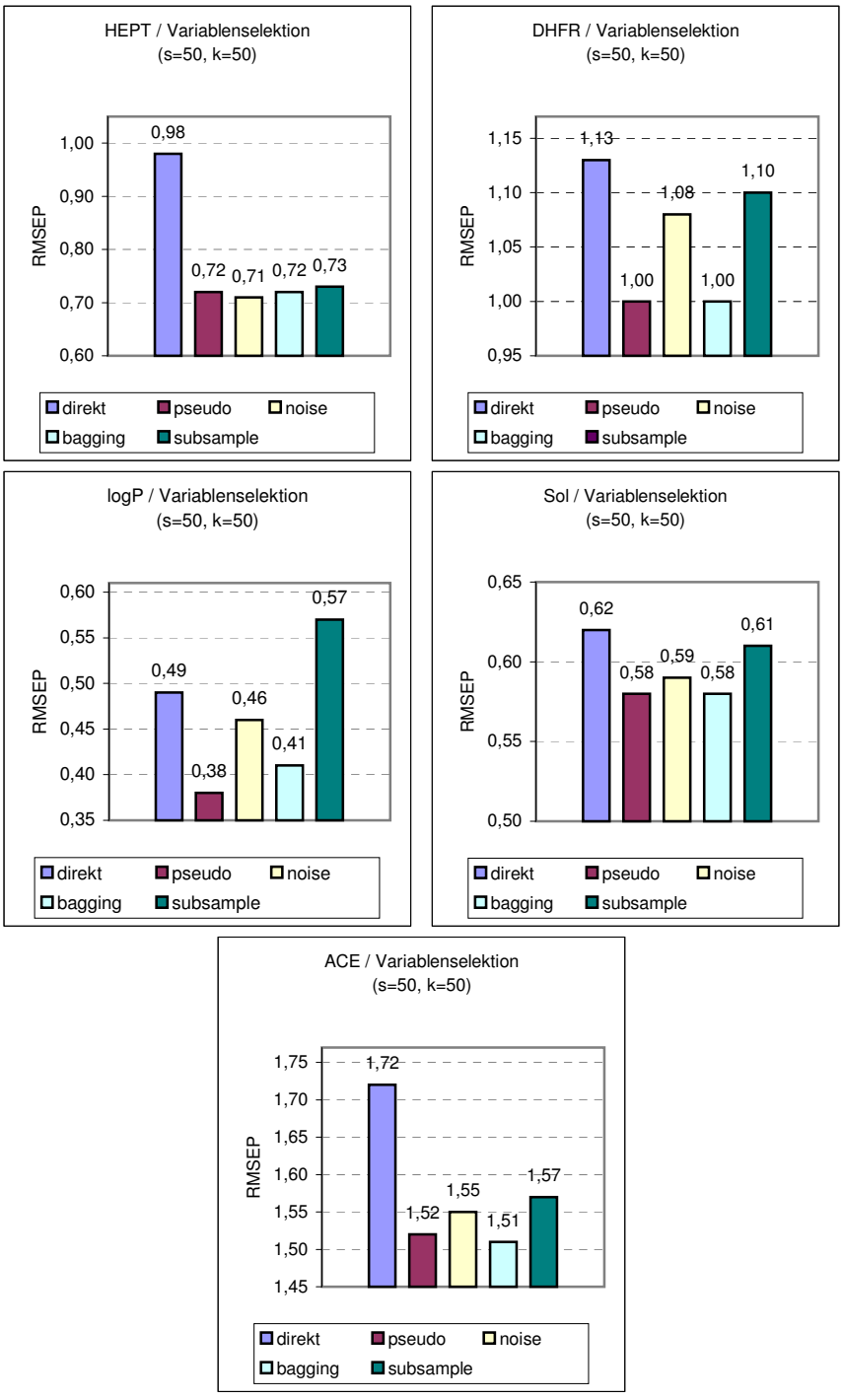


Abbildung 3.11 Vorhersagefehler (RMSEP) des Einzelmodells und der verschiedenen Ensembles für die Modelle mit Variablenselektion. Für alle fünf Datensätze wurde eine Reduzierung des Vorhersagefehlers gegenüber dem $RMSEP_{dir}$ erzielt. Die Methoden Konvexe Pseudodaten und Bagging führten dabei im Durchschnitt zu den deutlichsten Verbesserungen.

Tabelle 3.3 Prozentuale Verbesserung des Vorhersagefehlers RMSEP (gegenüber dem $RMSEP_{dir}$ des Einzelmodells) durch den Einsatz von Ensembles bei PCR-Modellen mit Variablenselektion; angegeben ist jeweils die stärkste erzielte Verbesserung und die zugehörige(n) Ensemble-Methode(n).

Datensatz	logP	HEPT	Sol	ACE	DHFR
Verbesserung (%)	22	28	7	12	12
Methode(n)	pseudo	noise pseudo bagging	pseudo bagging noise	bagging pseudo	pseudo bagging

Variablenselektion (vgl. Tab. 3.2 oberer vs. unterer Teil sowie Abb. 3.10 vs. Abb. 3.11). Dabei liegt die höchste erzielte Reduktion des mittleren Vorhersagefehlers bei deutlichen 28 Prozent (HEPT, Noise addition).

Im Durchschnitt erweist sich — wie schon bei den Simulationen ohne Variablenselektion — die Methode der konvexen Pseudodaten als das zuverlässigste und effizienteste Ensemble-Verfahren, wobei nun allerdings die Ergebnisse des Baggings gleichauf liegen. Mit diesen beiden Verfahren konnte für alle fünf Datensätze eine deutliche Verbesserung des Vorhersagefehlers gegenüber dem $RMSEP_{dir}$ erzielt werden. Lediglich in einem Fall (HEPT, Noise addition) schnitt eine andere Ensemble-Methode ebenso gut ab.

Durch den Einsatz der Noise addition wird in allen Fällen eine Senkung des RMSEP erzielt, die im Mittel allerdings geringer ausfällt als bei den konvexen Pseudodaten und beim Bagging. Eine Ausnahme bilden die Datensätze HEPT und Sol, bei denen die Noise addition etwa gleich gut abschneidet wie Bagging und Konvexe Pseudodaten.

Das Subsampling schneidet erneut am schlechtesten ab, auch wenn die Ergebnisse uneinheitlicher sind als bei den Simulationen ohne Variablenselektion. Während für HEPT eine beinahe ebenso deutliche Verbesserung erzielt wird wie mit den Methoden Bagging und Konvexe Pseudodaten, verschlechtert sich der Vorhersagefehler bei logP um 16 Prozent; bei Sol und DHFR ist keine nennenswerte Differenz zu beobachten.

Die Methoden Subsampling und Noise addition wurden — wie schon bei den Modellen ohne Variablenselektion (siehe 3.2.1.2, Seite 78) — zusätzlich noch einmal mit einer Datensatzaufteilung von 2/3 Trainingsdaten und 1/3 Testdaten untersucht. Auch hier bleiben die Ergebnisse der Noise addition

im wesentlichen stabil. Das Subsampling führt nun beim Datensatz logP nicht mehr zu einer Verschlechterung des Vorhersagefehlers, sondern dieser bleibt gegenüber dem Einzelmodell unverändert.

3.2.3 Folgerungen für den Einsatz von Ensembles

Die hier vorgestellten Simulationen an fünf realen Datensätzen zeigen, daß sich der Einsatz von Ensemble-Techniken durchaus positiv auf den Vorhersagefehler auswirken kann. Allerdings werden große Unterschiede zwischen den einzelnen Methoden zur Ensemble-Generierung deutlich.

Die zuverlässigsten Verbesserungen wurden mit der Methode Konvexe Pseudodaten erzielt. Nach den hier gewonnenen Erfahrungen kann dieses Verfahren als universell einsetzbar gelten, da es in allen Fällen zu einer Reduzierung des RMSEP führte. Auch das Bagging lieferte gute Ergebnisse und ist kaum schlechter als die konvexen Pseudodaten.

Der Nutzen der Verwendung von Ensembles ist für jeden Datensatz sehr individuell. Während eine Reduzierung des Vorhersagefehlers um 15–28 Prozent (Datensatz HEPT) den erhöhten Aufwand durchaus rechtfertigt, hat eine Verbesserung von 5–7 Prozent (Datensatz Sol) für den praktischen Einsatz nur geringe Relevanz. Insofern ist es für die routinemäßige Implementierung des Ensemble-Ansatzes umso wichtiger, durch die Wahl einer zuverlässigen Methode zur Erzeugung der Ensembles zumindest Verschlechterungen des Vorhersagefehlers auszuschließen.

Wie erwartet zeigt der Einsatz von Ensembles bei Modellen mit vorangehender Variablenselektion die deutlichste Wirkung. Allerdings stellt gerade diese Art der Modellierung ohnehin schon hohe Anforderungen an die Rechenleistung, so daß die zusätzliche Verwendung von Ensembles gleich in zweifacher Hinsicht zu einer erheblichen Verlängerung der benötigten Rechenzeit führt. Zusammenfassend bleibt festzuhalten, daß

- die Wahl der Ensemble-Methode entscheidenden Einfluß auf die Höhe der RMSEP-Verbesserung besitzt,
- die Methoden Konvexe Pseudodaten und Bagging die zuverlässigsten Ergebnisse erzielen und stets zu einer signifikanten Reduzierung des mittleren Vorhersagefehlers führen und

- Subsampling und Noise addition im allgemeinen nicht als Methoden zur Ensemble-Generierung zu empfehlen sind und insbesondere das Subsampling bei (zu) kleinen Trainingsdatensätzen leicht in einer Verschlechterung des Vorhersagefehlers resultieren kann.

Bei ausreichender Rechenkapazität erscheint die Verwendung von mittels konvexen Pseudodaten oder Bagging erzeugten Ensembles also durchaus attraktiv für den Routineeinsatz. Diese beiden Methoden eröffnen die Möglichkeit zur signifikanten Verbesserung des Vorhersagefehlers und bergen andererseits nach den hier gewonnenen Erfahrungen nicht die Gefahr von negativen Auswirkungen.

Die guten Resultate der konvexen Pseudodaten sind umso interessanter, als die Verwendung dieser Methode zur Modellstabilisierung im Rahmen der QSAR-Analyse bislang in der Literatur noch nicht vergleichend untersucht wurde.

3.3 Distanzbasierte Ähnlichkeitssuche DIBSI

Die Nutzung der *k*-Nearest-Neighbour-Distanz zur Identifizierung von Outliern wurde bereits in der Beschreibung der Methode ODD (siehe 3.1.2, Seite 60) vorgestellt. Aus dem Ansatz heraus, die Distanzen von Verbindungen im Deskriptorraum zum Auffinden *andersartiger* Moleküle (nämlich der Outlier) zu verwenden, ergibt sich der Gedanke, dasselbe Maß umgekehrt auch zur Suche nach *gleichartigen* Verbindungen einzusetzen. Bei ODD wird ein Testobjekt als Outlier identfiziert, wenn es eine hohe Distanz zum nächstgelegenen Trainingsobjekt aufweist ($NND_{Te/Tr}$), d. h. eine hohe Distanz wird als starke Unähnlichkeit interpretiert. Umgekehrt kann für die Anwendung zur Ähnlichkeitssuche eine niedrige Distanz zweier Verbindungen im Deskriptorraum als Hinweis auf ihre hohe Ähnlichkeit verstanden werden.

Eben diesen Ansatz verfolgt die hier vorgestellte Methode DIBSI (*DIstance-Based SImilarity Search*). Sie nutzt den etablierten Ansatz der distanzbasierten Ähnlichkeitsbestimmung,[125,126] wobei sich die hier verwendete Kombination aus MOE-Deskriptoren und euklidischer Distanz als besonders einfach implementierbar, anschaulich und zugleich nutzbringend erwies.

3.3.1 Anforderungen

Die Suche nach Verbindungen, die zu einem gegebenen Referenzmolekül ähnlich sind, ist eine alltägliche Aufgabe in der computergestützten Wirkstoffentwicklung. Die allgemeine Hypothese, daß ähnliche chemische Strukturen auch ähnliche biologische oder pharmakologische Wirkungen hervorrufen,[104,127] legt diese Vorgehensweise nahe: Die Kenntnis einer aktiven Substanz, einer sogenannten Leitstruktur (engl. *Lead structure*), ermöglicht das Auffinden weiterer aktiver Moleküle durch Bestimmung ihrer Ähnlichkeit.

Da in der Wirkstoffentwicklung oft sehr große virtuelle Molekülbibliotheken durchsucht werden, sollte ein Verfahren zur Ähnlichkeitssuche möglichst schnell sein. Darüber hinaus muß es der Anforderung genügen, auch auf große Datenmengen anwendbar zu sein — in der Praxis wird häufig ein Maß für die Ähnlichkeit von mehreren Millionen Verbindungen benötigt. Liegt nicht nur eine einzelne Referenz vor, sondern eine ganze Serie von bekannten

Aktiven, so muß der dann schnell ansteigende Aufwand der Ähnlichkeitssuche dennoch handhabbar bleiben. In diesem Fall stellt sich außerdem die Frage, wie die Ähnlichkeit zu mehreren Vergleichsmolekülen bestimmt werden soll, etwa über den Durchschnitt aller Referenzen, als gewichteter Mittelwert, als minimale Distanz bezüglich einer einzelnen Referenz usw. Hier sollte das Verfahren möglichst flexibel einsetzbar sein.

Dabei darf der eigentliche Zweck der Ähnlichkeitssuche nicht außer acht gelassen werden: Ist beispielsweise die Leitstruktur an einer bestimmten Position des Moleküls chlor-substituiert, so weist die analoge brom-substituierte Struktur natürlich eine sehr hohe Ähnlichkeit auf. Trotzdem besitzt dieses Ergebnis nur einen geringen Wert, weil es offensichtlich ist und keine neuen Impulse für den Entwicklungsprozeß liefert. Die paradoxe Anforderung an eine für die Wirkstoffentwicklung besonders nützliche Ähnlichkeitssuche besteht also oft darin, gerade solche Verbindungen ausfindig zu machen, die auf den ersten Blick gar keine hohe Ähnlichkeit mit der Referenz aufweisen und sich erst auf den zweiten Blick als gleichartig herausstellen.

Ein Aspekt dieser Forderung ist etwa das *Scaffold hopping*[128] (auch *Lead hopping*[129]), also der Sprung zu einem alternativen chemischen Grundgerüst mit dennoch isofunktionellen bzw. bioisosteren Funktionalitäten. Dies ist häufig schon aus allein praktischen Erwägungen heraus wünschenswert, etwa um den Bereich eines bereits patentierten Chemotyps zu verlassen oder einer chemisch unlösbaren Synthese auszuweichen.[130] Eine einfache Substruktursuche beispielsweise kann diese Aufgabe jedoch nicht erfüllen.

3.3.2 Berechnung der DIBSI-Ähnlichkeit

Für die Ähnlichkeitssuche mit DIBSI wird die euklidische Distanz zweier Verbindungen im Raum der autoskalierten Deskriptoren berechnet. Diese Berechnung ist prinzipiell mit jedem beliebigen Satz von Deskriptoren möglich. In der vorliegenden Arbeit wurden dazu die 181 translations- und rotationsinvarianten MOE-Deskriptoren (siehe A.2, Seite 141) verwendet. Diese Deskriptoren haben sich in der alltäglichen Arbeit bewährt und sind allen Erfahrungen nach gut geeignet, die relevanten Eigenschaften eines Moleküls umfassend zu charakterisieren. Außerdem stand eine virtuelle Bibliothek zur Verfügung, in der mehrere Millionen Moleküle bereits inklusive ihrer MOE-

Deskriptoren gespeichert sind, so daß ihre Verwendung auch unter praktischen Gesichtspunkten nahelag.

Natürlich muß die gewünschte Definition von Ähnlichkeit in jedem Fall mit den verwendeten Deskriptoren übereinstimmen. Soll sich zum Beispiel die Ähnlichkeit auch auf HOMO- und LUMO-Energien beziehen, so sind die MOE-Deskriptoren ungeeignet, weil sie diese Eigenschaften nicht explizit beschreiben. Der Deskriptor-Satz müßte also entsprechend angepaßt werden.

Die euklidische Distanz stellt ein Distanzmaß dar, das eine gewisse Anschaulichkeit besitzt und vor allem auch für große Molekülbibliotheken mit geringem Aufwand zu berechnen ist. In der praktischen Umsetzung von DIBSI (mit OCTAVE 2.1.40[131] implementiert) wurde dabei auf die Berechnung der Wurzel gemäß Gl. 2.33 (Seite 34) verzichtet und stattdessen das Abstandsquadrat ED^2 zweier Objekte $\mathbf{x}$ und $\mathbf{y}$ verwendet.

$$ED^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{3.11}$$

Dies hat auf den Vergleich der Moleküle innerhalb der Datenbank keinen Einfluß, bringt jedoch einen deutlichen Vorteil bei der benötigten Rechenzeit.

3.3.3 Anwendung

Die Anwendung einer Ähnlichkeitssuche zum virtuellen Screening (VS) wird eingehend in 3.4 beschrieben. Es kann darüber hinaus aber auch sinnvoll sein, beispielsweise die Ergebnisse eines dockingbasierten Screenings (2.5.2) nachträglich mit einer Ähnlichkeitssuche zu analysieren.

Die aus dem Docking erhaltene Rangliste stellt nämlich nur eine recht grobe Sortierung der gescreenten Verbindungen dar. Für die biologische Testung der Substanzen werden aus dieser anhand des Dockingscores sortierten Liste z. B. die ersten 500 Einträge ausgewählt. Häufig unterscheidet sich jedoch der Score der 500. Verbindung nur minimal von dem der 2000., da gewissermaßen die „Meßgenauigkeit“ des vereinfachenden Docking-/Scoringalgorithmus zu einer so feinen Unterscheidung gar nicht in der Lage ist. Bei ausschließlicher Betrachtung der obersten Plätze der Rangliste werden also viele im Grunde gleichwertige Ergebnisse verworfen. Eine Ähnlichkeitssu-

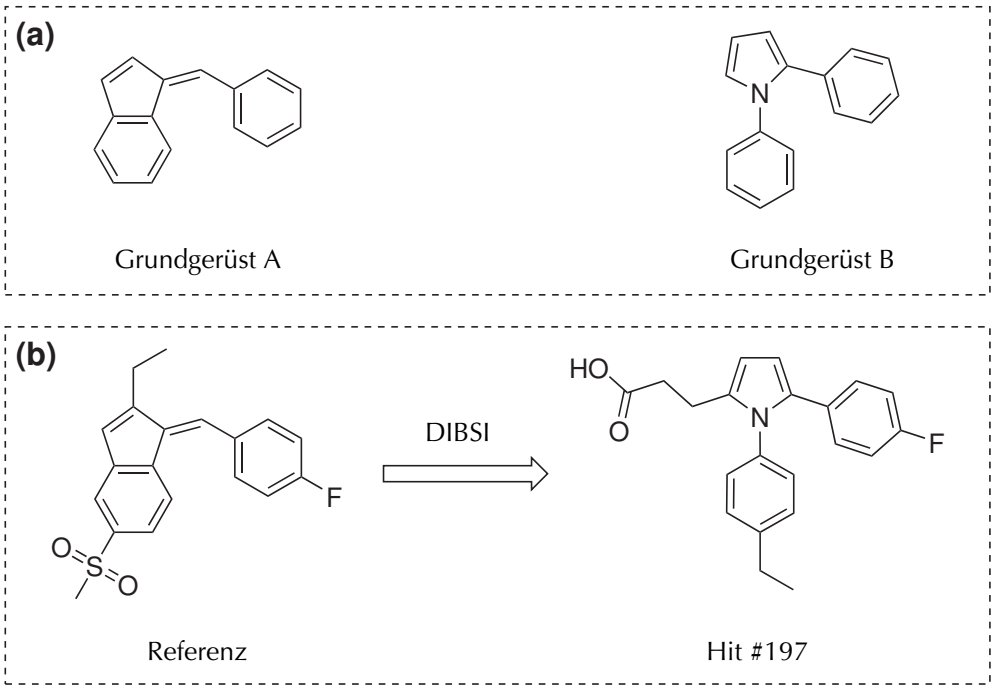


Abbildung 3.12 Beispiel eines mit DIBSI erreichten Scaffold-Hoppings. (a) Grundgerüste zweier bekannter Klassen von COX-2-Inhibitoren; (b) DIBSI findet ausgehend von der Referenz mit Grundgerüst A auch ähnliche Moleküle mit Grundgerüst B.

che auf Grundlage einiger hochrangiger, strukturell verschiedener Moleküle kann hier weitere interessante Substanzen zu Tage fördern, die sonst übersehen worden wären.

3.3.3.1 Beispiel einer DIBSI-Ähnlichkeitssuche

In Abb. 3.12 ist eine exemplarische Ähnlichkeitssuche mit DIBSI dargestellt, bei der auch das häufig geforderte Scaffold-Hopping erreicht wird. Ausgangspunkt war ein Referenzmolekül, welches dasselbe Grundgerüst besitzt wie eine Klasse bekannter COX-2-Inhibitoren (Grundgerüst A).[130] Die DIBSI-Ähnlichkeitssuche in einer virtuellen Bibliothek mit mehreren Millionen Verbindungen* findet bereits an 197. Stelle der nach Ähnlichkeit sortierten Rangliste ein Molekül, das einer zweiten bekannten Klasse von COX-2-Inhi-

*mehr als 5 Millionen kommerziell erhältlich (4SC AG, firmenintern)

bitoren angehört (Grundgerüst B). DIBSI ist also — im Gegensatz etwa zu einer Substruktursuche — in der Lage, den Strukturraum der Eingangsverbindung zu verlassen und anhand der deskriptorbezogenen Ähnlichkeit neue Substanzklassen zu erschließen.

3.4 Ligand- und strukturbasiertes virtuelles Screening

Die Grundzüge des struktur- und ligandbasierten virtuellen Screenings wurden bereits in 2.5.2 und 2.5.3 erläutert. Wie dort beschrieben, entscheidet häufig die Existenz entweder einer Rezeptor-Kristallstruktur oder einer Serie von bekannten Inhibitoren darüber, welcher der beiden Ansätze für das Screening verwendet wird. In der Praxis ist jedoch hin und wieder auch der Fall anzutreffen, daß sowohl Struktur- als auch Ligandinformation vorliegt, also beide Wege verfolgt werden können. Gerade bei schon weiter fortgeschrittenen Entwicklungsprojekten oder der Bearbeitung bereits gut erforschter Targets tritt diese Situation auf.

Der Computerchemiker steht dann vor der Frage, welche Vorgehensweise die besten Ergebnisse liefert, am effizientesten, schnellsten, kostengünstigsten arbeitet — der ligandbasierte oder der strukturbasierte Ansatz. Oder mehr noch: Wie die unterschiedliche Information aus der Proteinstruktur einerseits und bekannten Inhibitoren andererseits gewinnbringend zusammengeführt werden kann, so daß die Kombination einen größeren Nutzen erbringt als die beiden Einzeltechniken. Schließlich sollte es das Ziel sein, möglichst die gesamte zur Verfügung stehende Information für das virtuelle Screening zu nutzen und nicht durch die Beschränkung auf nur einen Ansatz einen großen Teil vorhandener Erkenntnisse außer acht zu lassen.

In der nachfolgend präsentierten Studie wird ein virtuelles Screening an sechs verschiedenen Datensätzen simuliert. Dabei kommen struktur- und ligandbasierte Verfahren gleichermaßen zum Einsatz; außerdem wird ein Zugang zur kombinierten Anwendung beider Strategien aufgezeigt. Die Leistungsfähigkeit des VS wird dabei anhand der Datenbankanreicherung bewertet.

3.4.1 Durchführung verschiedener VS-Verfahren

Das virtuelle Screening wurde simuliert, indem eine große virtuelle Datenbank von etwa 93 000 inaktiven Verbindungen mit einigen hundert Molekülen vermischt wurde, die an einem bestimmten Target bekanntermaßen aktiv sind. Diese Aktiven sollten dann im VS wiedergefunden werden. Beim strukturbasierten Ansatz wurde dazu die gesamte Datenbank in das entsprechen-

de Target gedockt. Für die Untersuchung ligandbasierter Strategien wurden dem Datensatz der Aktiven zunächst einige wenige Moleküle entnommen und nur die verbleibenden mit den Inaktiven vermischt; anhand der Information der zurückbehaltenen Aktiven war dann beispielsweise eine Ähnlichkeitssuche möglich.

Jedes Screeningverfahren liefert als Ergebnis zunächst eine Rangliste aller Verbindungen, sortiert entsprechend der prognostizierten Aktivität. Die Datenbankanreicherung kann dann leicht berechnet werden, indem die Anzahl der wiedergefundenen Aktiven in einem bestimmten Bruchteil der Rangliste bestimmt wird (siehe 2.5.1, Seite 49).

Für die Praxis relevant ist hier maximal die Anreicherung im obersten Prozent der Datenbank. Wird nämlich in einem realen Projekt zum Beispiel eine Substanzbibliothek mit nur 1 000 000 Strukturen gescreent, entspricht das oberste Prozent bereits 10 000 Verbindungen — eine Anzahl, die für die experimentelle biologische Testung noch deutlich zu hoch liegt; dafür werden meist nur einige hundert oder tausend Verbindungen ausgewählt (im Beispiel also gerade einmal 0.1 Prozent). Zusammen mit der Tatsache, daß übliche Screeningbibliotheken oft mehrere Millionen Verbindungen enthalten, lautet die Anforderung an das VS also, die Aktiven bereits im obersten Promille der Datenbank oder sogar nur einem Bruchteil davon anzureichern.

3.4.1.1 Verwendete Datensätze

Die Simulationen wurden auf Grundlage der Datenbank *MDL Drug Data Report* (MDDR)[132] durchgeführt. Aus diesem Verzeichnis bekannter pharmazeutischer Wirkstoffe können Datensätze von Inhibitoren der verschiedensten Targets extrahiert werden. Die hier vorgestellten Studien greifen auf insgesamt sechs solcher Datensätze zurück. Vier davon wurden einer Publikation von HERT *et al.* entnommen[107,133]: HIV, Thrombin, Renin und COX. Der ebenfalls dort beschriebene große Datensatz von Substanzen ohne bekannte Aktivität (im folgenden InA, Inaktive*) wurde als Screeningdatenbank verwendet.

*Die Bezeichnung „Inaktive“ ist nicht exakt, da für diese Verbindungen zwar bislang keine Aktivität beschrieben, möglicherweise aber dennoch vorhanden ist.

In gleicher Weise wurden aus dem MDDR zwei weitere Datensätze extrahiert, nämlich ACE und AChE. Da HERT *et al.* ihren Datensatz InA ohne Berücksichtigung dieser beiden Targets erstellt haben, wurde jeweils die Schnittmenge aus InA entfernt, so daß sich für diese beiden Simulationen zwei geringfügig kleinere Screeningdatenbanken InA-ACE und InA-AChE ergaben. Alle neun Datensätze sind in Anhang A.1 (Seite 139) verzeichnet.

Aufbereitung der Molekülstrukturen Die im RDF-Format vorliegenden Daten des MDDR wurden zunächst mit einem Perl-Script ins SD-Format konvertiert. Anschließend wurde CORINA[134] zur Erzeugung dreidimensionaler Konformationen verwendet. Eventuell vorhandene Gegenionen wurden gelöscht (CORINA-Option `-d rs`). Alle Verbindungen, die in CORINA Fehler hervorriefen, wurden ohne weitere Überprüfung gelöscht.* Für das Docking mit FLEXX wurden diese Strukturen mit Hilfe eines regelbasierten TRIPOS-SPL-Scripts (`ionise.sh`, verwendet `dbtranslate`) protoniert und mit CORINA ins mol2-Format konvertiert. Für die distanzbasierte Ähnlichkeitssuche wurden (automatisiert durch ein MOE-SVL-Script) die 181 rotations- und translationsinvarianten MOE-Deskriptoren berechnet und im ASCII-Format gespeichert.

3.4.1.2 Vorgehensweise FlexX

Alle Moleküle der jeweiligen durchmischten Screeningdatenbank (InA plus bekannte Aktive) wurden in das zugehörige Target gedockt. Die entsprechenden Kristallstrukturen wurden der RCSB Protein Data Bank[135] entnommen; die PDB-Codes sind aus Tab. 3.4 (Seite 96) ersichtlich.

Das Docking wurde mit den FLEXX-Standardeinstellungen durchgeführt. Die Bindetaschen wurden je nach Target individuell definiert. So besitzt etwa die Acetylcholinesterase (Datensatz AChE, PDB-Code 1EVE) eine enge und tiefe Bindetasche mit einer Länge von fast 25 Ångström, während beispielsweise für 4PHV der FLEXX-Standard mit einem Radius von 6.5 Å ausreichend ist. Für jedes Molekül wurde der beste Dockingscore gespeichert. In den Fällen, in denen FLEXX gar keine Plazierung des Liganden finden und somit

*Dadurch ergibt sich teilweise eine geringfügig veränderte Datensatzgröße gegenüber der von HERT *et al.* publizierten.

auch keinen Score berechnen konnte, wurde der Verbindung willkürlich ein Score von 200 (also ein äußerst schlechter Wert) zugewiesen.

Reproduzierung der bekannten Bindungsmodi Um die Definitionen der Bindetaschen und die FLEXX-Einstellungen zu validieren, wurden die co-kristallisierten Liganden in ihr Target gedockt. Vor der Verwendung eines Dockings zum virtuellen Screening sollte es zunächst unter Beweis stellen, die Position und Konformation der Ligand-Kristallstruktur in akzeptabler Weise reproduzieren zu können.[136] Dies wurde einerseits durch die visuelle Inspizierung der Bindungsmodi der Dockinglösungen überprüft, andererseits durch die Berechnung des RMSD (engl. *Root mean squared deviation*), der mittleren Abweichung der Schweratome (alle Atome außer Wasserstoff).

Der Lösungssatz eines FLEXX-Dockings besteht meist aus einigen hundert möglichen Plazierungen des Liganden. Aufgrund der relativ groben Abschätzung der Bindungsaffinität durch den FLEXX-Score besteht die Möglichkeit, daß die strukturell beste Ligandplazierung fälschlicherweise nicht auch gleichzeitig diejenige mit dem besten Score ist. Daher ist es sinnvoll, für die o. g. RMSD-Berechnung nicht ausschließlich die Lösung mit dem besten Score zu berücksichtigen. Vielmehr sollte der minimale RMSD der z. B. ersten 20 oder 50 nach Score sortierten Plazierungen als Maßstab für die Güte des Dockings herangezogen werden.[68] Oft unterscheiden sich die Scores dieser „Topscorer“ nämlich nur in geringem Maße, während die zugehörigen Bindungsmodi weitaus deutlicher differieren.

Für alle Systeme konnte FLEXX die Kristallstruktur wenigstens zufriedenstellend reproduzieren: Die für 1DWD, 1CX2 und 1O86 gefundenen Plazierungen zeigen eine gute Übereinstimmung mit den kristallographischen Daten, auch der RMSD von 1EVE ist noch annehmbar (siehe Tab. 3.4, Seite 96). Für die Komplexe 4PHV und 1HRN liegt die mittlere Abweichung der Schweratome zwar deutlich höher. Die Überlagerung in Abb. 3.13 zeigt aber, daß das Docking dennoch keine unsinnigen Plazierungen erzeugt und der prinzipielle Bindungsmodus in akzeptablem Maße reproduziert wurde. So relativiert sich der mit 7.6 Å sehr hohe RMSD von 1HRN bei genauerer Betrachtung der Dockinglösung: Die Ligandstruktur besitzt eine „H“-förmige Gestalt mit symmetrisch angeordneten Funktionalitäten. Im Vergleich zur Po-

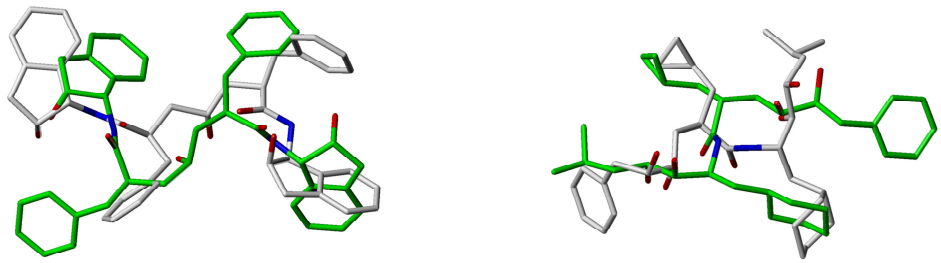


Abbildung 3.13 Überlagerung von gedockter (grau) und co-kristallisierter (grün) Ligandstruktur; links: 4PHV (HIV), rechts: 1HRN (Renin). Trotz der hohen mittleren Abweichung der Schweratome (RMSD) wurde der Bindungsmodus durch das Docking prinzipiell korrekt reproduziert. Dies wird insbesondere an der gleichsinnigen Ausrichtung der Donor- und Akzeptor-Funktionalitäten deutlich.

sition des Co-Kristallisats ist die gedockte Struktur lediglich um die Diagonale des „H"s gedreht, wobei die funktionellen Gruppen wieder an äquivalenter Stelle zu liegen kommen (siehe Abb. 3.13, rechts; der Cyclopropyl- und Cyclohexylrest liegen richtig, die Isobutylgruppe und der Phenylring jedoch sind gegeneinander vertauscht und verursachen den hohen RMSD-Wert).

Zudem stellen die co-kristallisierten Inhibitoren von 4PHV bzw. 1HRN sehr flexible Moleküle mit einer hohen Anzahl rotierbarer Bindungen dar. In solchen Fällen sinkt jedoch bekanntermaßen die Präzision des Dockings im allgemeinen[137] und insbesondere die der mit FLEXX erhaltenen Resultate[138]. Vor diesem Hintergrund sind die hohen RMSD-Werte bei 1HRN und 4PHV zwar als Warnung zu verstehen, die auf die generellen Unzulänglichkeiten heute verfügbarer Dockingmethoden hinweist; sie stellen jedoch die Validität der beiden Systeme für ein virtuelles Screening nicht in Frage.

3.4.1.3 Vorgehensweise FlexX/SIFt

Die Berechnung von Structural interaction fingerprints (SIFts) eröffnet die Möglichkeit, ligand- und strukturbasierte Information zu kombinieren. Für diesen Ansatz wurde folgende Vorgehensweise gewählt, die getrennt für jedes der sechs Targets durchgeführt wurde:

Für die Bindetasche wird eine Liste aller prinzipiell möglichen Interaktionszentren und -typen erzeugt. Dazu wird die Bindetasche in FLEXX mit der Option `set verbosity 10` geladen. In der vom Programm erzeugten detail-

Tabelle 3.4 Mittlere Abweichung (RMSD) der Schweratome zwischen co-kristallisiertem und gedocktem Liganden. Der Wert bezieht sich auf die Dockinglösung mit dem geringsten RMSD unter den ersten 20 nach FlexX-Score sortierten Plazierungen. Trotz der hohen Abweichungen sind auch die Ergebnisse für 4PHV und 1HRN zufriedenstellend (siehe Text).

PDB-Code	Datensatz	RMSD [Å]
1O86[139]	ACE	1.8
1EVE[140]	AChE	2.9
4PHV[141]	HIV	3.2
1CX2[142]	COX	1.6
1DWD[143]	Thrombin	1.4
1HRN[144]	Renin	7.6

lierten Ausgabe ist dann auch die gewünschte Auflistung der Wechselwirkungsmöglichkeiten enthalten.* Diese wurde mit Hilfe eines Perl-Scripts ausgewertet und bildet die Grundlage für den SIFt: Der Fingerprint enthält für jede in der Liste verzeichnete Interaktionsmöglichkeit ein Bit. Alle Bits sind zunächst auf Null gesetzt (siehe 2.4.4, Seite 43).

Anschließend werden dem Datensatz der Aktiven zehn zufällige Moleküle entnommen und nur die verbleibenden mit den Inaktiven (InA) gemischt. Die zehn bekannten Inhibitoren werden mit FLEXX in die Zielstruktur gedockt und für die ersten 30 gefundenen Plazierungen jeder Verbindung der SIFt erstellt. Dazu wird in FLEXX mit dem Befehl `listmat` ausgegeben, welche Interaktionen beim Docking gefunden wurden. Diese werden mit der Liste der prinzipiell möglichen Wechselwirkungen verglichen und die entsprechenden Bits im Fingerprint auf Eins gesetzt. Auch dieser Schritt wurde extern durch ein Perl-Script realisiert, das die zwischengespeicherten FLEXX-Ausgaben auswertet.

Am Ende liegt also für die besten 30 Plazierungen einer jeden der zehn Referenzverbindungen je ein SIFt vor. Diese Daten spiegeln die bevorzugten Interaktionen der Moleküle wider, d. h. es wird Information aus der Serie bekannter Inhibitoren für das Screening nutzbar gemacht.

Abschließend wird die Screeningdatenbank, also die Menge der Inaktiven und der verbliebenen Aktiven, gedockt. Auch hier werden für jedes Molekül die SIFts seiner (gemäß FLEXX-Score) besten 30 Plazierungen gespeichert.

*Tabelle mit der Bezeichnung „Identification of contact types (receptor)“.

Ein Vergleich des SIFTs einer Verbindung aus der Screeningdatenbank mit den SIFts der Referenzstrukturen erlaubt eine Beurteilung der Dockinglösung, die über den reinen FLEXX-Score hinausgeht. Weist der SIFt dieser Lösung nämlich eine hohe Ähnlichkeit mit den SIFts der Referenzen auf, so nimmt dieses Molekül einen ähnlichen Bindungsmodus wie die bekanntermaßen aktiven Verbindungen ein. Dies wiederum ist ein deutlicher Hinweis darauf, daß dieses Molekül auch eine ähnlich hohe Aktivität besitzt.

Auf diese Weise wird also die aus den Referenzstrukturen extrahierte Information mit dem strukturbasierten FLEXX-Docking verknüpft. Dabei bleibt der Schritt der Plazierung des Liganden unbeeinflußt. Die Sortierung (das Erzeugen der Rangliste) jedoch wird optimiert, indem diesem Vorgang die Bindungscharakteristik der aktiven Referenzen zur Verfügung gestellt wird. Dadurch basiert die Rangliste nicht mehr allein auf dem FLEXX-Score, sondern wird durch das Wissen darum unterstützt, welche Ligand-Rezeptor-Interaktionen offenbar für die Aktivität einer Verbindung essentiell sind.

Die Simulation wurde 100 Mal durchgeführt, d. h. es wurden 100 Mal zehn zufällige Aktive ausgewählt, die als Referenz dienten. Über die Ergebnisse dieser 100 Simulationen wurde der Mittelwert gebildet. Die genaue mathematische Auswertung der SIFts wird weiter unten in 3.4.2 beschrieben.

3.4.1.4 Vorgehensweise MOE

Als rein ligandbasiertes Screeningverfahren wurde eine Ähnlichkeitssuche mit DIBSI durchgeführt, also die euklidische Distanz im MOE-Deskriptorraum bestimmt. Dazu wurden aus dem Datensatz der bekannten Aktiven zunächst zehn Verbindungen entnommen, die als Referenzen für die Ähnlichkeitssuche dienten. Für die verbleibenden Aktiven sowie den InA-Datensatz wurde die DIBSI-Ähnlichkeit zu den Referenzen berechnet. Da die einzelnen zugrundeliegenden MOE-Deskriptoren auf gänzlich unterschiedlichen Skalen gemessen werden, wurde die Screeningdatenbank zuvor autoskaliert und die Referenzen ebenfalls entsprechend skaliert. Schließlich lag für jede Substanz der Screeningdatenbank ein Ähnlichkeitswert zu jeder der zehn Referenzen vor.

Die weitere Auswertung kann nun auf zwei verschiedene Weisen erfolgen: Entweder wird die durchschnittliche Distanz zu allen Referenzstrukturen be-

rechnet, oder aber die minimale Distanz zu nur einer der Referenzen; im letztgenannten Fall bleiben die übrigen neun Distanzen (die einen höheren Wert besitzen) unberücksichtigt. Die Entscheidung, welcher Weg verfolgt werden sollte, ist u. a. abhängig von der Diversität der Referenzstrukturen. Liegt eine hohe Diversität vor, besitzen also bereits die Referenzen untereinander eine hohe Distanz im Deskriptorraum, so kann sich die Methode mit der Berechnung des Durchschnitts nachteilig auswirken: Die untersuchte Verbindung ist dann zwar ähnlich zum gedachten „Durchschnittsmolekül", weist aber in ungünstigen Fällen nur noch sehr geringe Ähnlichkeit zu einer der konkreten Referenzen auf. Hier wäre also die Methode der minimalen Distanz vorzuziehen. Studien von HERT *et al.* legen ebenfalls nahe, daß die minimale Distanz zu nur einer der Referenzen das bessere Kriterium darstellt.[107]

Auch die auf den MOE-Deskriptoren basierenden Simulationen wurden je 100 Mal wiederholt, so daß sich durch die zufällige Auswahl der zehn Aktiven immer wieder eine andere Zusammensetzung des Referenzdatensatzes ergab.

3.4.2 Auswertung der Daten

Aus allen durchgeführten VS-Simulationen resultiert für jedes Molekül der Screeningdatenbank ein Score (entweder direkt der FLEXX-Score oder ein berechneter Ähnlichkeitsscore). Diesem Wert entsprechend wird die gesamte Datenbank sortiert, so daß die Moleküle mit den besten Scores die vordersten Plätze einnehmen. Anhand dieser Rangfolge kann dann der Recall (siehe Gl. 2.40, Seite 50) berechnet werden, also die Anzahl der wiedergefundenen Aktiven, wenn die Datenbank zu einem bestimmten Prozentsatz durchsucht wurde.

Zur Veranschaulichung der Ergebnisse wurde für jeden Datensatz der Recall graphisch dargestellt, indem der Anteil der wiedergefundenen Aktiven gegen den Anteil der bereits durchsuchten Screeningdatenbank aufgetragen wurde (siehe Abb. 3.14–3.16). Dies geschah getrennt für die Ergebnisse der SIFt-basierten Simulationen und die Auswertungen, die auf den MOE-Deskriptoren beruhen. In beiden Graphen ist zum Vergleich jeweils auch der von FLEXX erzielte Recall eingezeichnet. Weiterhin enthalten alle Abbildungen am unteren Ende der Y-Achse einen ansteigenden grauen Bereich, der

den Recall kennzeichnet, der bei einer Zufallsauswahl zu erwarten wäre. Alle Kurven, die innerhalb der grau dargestellten Fläche liegen, kennzeichnen demnach einen Recall, der nicht besser ist als die Zufallsauswahl (also zu einer „Abreicherung“ der Datenbank statt zu der gewünschten Anreicherung führt). Alle Berechnungen wurden mit MATLAB R14[145] angestellt.

3.4.2.1 SIFt

Für jede Verbindung wurden die SIFts der ersten 30 Dockingplazierungen gespeichert. Die Summe dieser 30 Fingerprints spiegelt gewissermaßen den „durchschnittlichen Bindungsmodus“ der besten Dockinglösungen wider. Beim Durchsuchen der Screeningdatenbank werden diese aufsummierten SIFts der zehn zufällig ausgewählten Referenzverbindungen mit den aufsummierten SIFts der gedockten Screeningverbindungen verglichen; als Ähnlichkeitsmaß diente der Tanimoto-Koeffizient.

Dabei sind zwei Varianten möglich: Die Summe der jeweils 30 SIFts kann direkt als Integer-Fingerprint verwendet werden; die Tanimoto-Ähnlichkeit berechnet sich dann gemäß Gl. 2.38 (Seite 39). Alternativ kann der Summen-Fingerprint in eine binäre Darstellung umgerechnet werden, d. h. alle Werte ungleich Null werden auf Eins gesetzt. Es ergibt sich also ein klassischer Bitstring, der nur aus 0/1-Informationen besteht.

Für jede dieser beiden Varianten stehen wiederum zwei Möglichkeiten zur Ähnlichkeitsberechnung offen: Entweder wird das betrachtete Molekül aus der Screeningdatenbank mit dem Durchschnitt aller zehn Referenz-Fingerprints verglichen, oder aber es wird nur das Minimum der Distanzen zu den zehn Referenzen verwendet.

Eine weitere Möglichkeit der Datenauswertung besteht in der Anwendung einer Klassifizierung. Die zehn aufsummierten Integer-SIFts der Referenzmoleküle definieren hier die Klasse der aktiven Moleküle, die inaktive Klasse wird durch 1000 zufällig ausgewählte Moleküle der Screeningdatenbank beschrieben.* Mit Hilfe einer Linearen Diskriminanzanalyse wird nun für jedes Molekül der Screeningdatenbank seine Klassenzugehörigkeit bestimmt.

*Die Klasse der Inaktiven enthält also auch eine gewisse Anzahl von aktiven Verbindungen; im vorliegenden Fall ist dieser Anteil von weniger als einem Prozent jedoch so gering, daß er die Klassifizierung nicht signifikant beeinträchtigt.

Die LDA wurde nicht mit den Original-SIFt-Daten durchgeführt, sondern mit den ersten 15 Hauptkomponenten. Diese optimale Anzahl von PCs wurde für den Renin-Datensatz empirisch ermittelt und dann auch bei allen anderen Datensätzen gewählt.

Im folgenden werden die fünf Varianten der Datenauswertung als Int/Avg (Integer, Durchschnitt), Int/Min (Integer, Minimum), Bin/Avg (Binär, Durchschnitt), Bin/Min (Binär, Minimum) und Int/LDA15 (Integer, Lineare Diskriminanzanalyse mit 15 Hauptkomponenten) bezeichnet.

3.4.2.2 MOE

Die MOE-Deskriptoren wurden im Sinne der DIBSI-Ähnlichkeitssuche mit Hilfe der euklidischen Distanz ausgewertet. Auch hier bieten sich wieder die beiden Möglichkeiten, die durchschnittliche Distanz zu allen zehn Referenzen oder die minimale Distanz zu nur einer Referenz zu betrachten. Diese werden nachfolgend als ED/Min bzw. ED/Avg bezeichnet. Da jedoch bei den ersten beiden untersuchten Datensätzen (ACE und AChE) ED/Avg bereits klar unterlegen war, wurde im weiteren Verlauf nur noch ED/Min berechnet.

Darüber hinaus kann der Vektor mit Deskriptorwerten, durch den ein jedes Molekül charakterisiert ist, auch analog einem Integer-Fingerprint aufgefaßt werden.* Als Ähnlichkeitsmaß wird dann der Tanimoto-Koeffizient entsprechend den oben beschriebenen Integer-SIFTs verwendet. Diese beiden Auswertungsvarianten tragen ebenfalls die Bezeichnung Int/Avg bzw. Int/Min; obwohl gleichlautend, bezeichnen die Kürzel also je nach Kontext (SIFt oder MOE) unterschiedliche Recalls.

Die Transformation in die binäre Darstellungsweise ist hier im Gegensatz zu der Auswertung der SIFts nicht sinnvoll; da praktisch alle MOE-Deskriptoren einen Wert $\neq 0$ annehmen, würde der resultierende Bitstring ebenfalls nur aus Einsen bestehen.

Auch mit den MOE-Deskriptoren ist, wie oben für die SIFts beschrieben, eine Klassifizierung mittels Linearer Diskriminanzanalyse möglich. Da der Informationsgehalt der MOE-Deskriptoren jedoch deutlich höher ist als der der SIFts, wurden hier die ersten 80 Hauptkomponenten berücksichtigt (Bezeich-

*Der Deskriptorvektor enthält zwar nicht nur Integer-, sondern auch Fließkommawerte; dies ist jedoch für die Berechnung des Tanimoto-Koeffizienten (Integer-Form) nicht relevant.

nung Int/LDA80). Diese Anzahl wurde am Beispiel des Renin-Datensatzes empirisch ermittelt und dann auch für alle anderen Datensätze verwendet.

3.4.3 Ergebnisse

Wie oben beschrieben wurde jedes simulierte Screening 100 Mal durchgeführt. Die Darstellung der Ergebnisse erfolgt durch Auftragung der Mittelwerte der dabei erzielten Recalls für die obersten 2.5 Prozent der Rangliste (Abb. 3.14–3.16, Seiten 102–104), jeweils getrennt für jede Screening- und Auswertungsmethode. Zusätzlich zeigen Abb. 3.17–3.18 (Seiten 105–106) den Recall bei 1 Prozent bzw. 1 Promille, also an zwei Punkten aus dem für die Praxis relevanten Bereich. Diese Werte sind auch numerisch in Tab. 3.5 und Tab. 3.6 (Seite 109) angegeben.

Insgesamt betrachtet wird beim ligandbasierten virtuellen Screening auf Grundlage der MOE-Deskriptoren die höchste Anreicherung erzielt, wenn die Daten mit der Methode ED/Min ausgewertet werden, also analog zur Ähnlichkeitssuche DIBSI. Vor allem der Recall bei 0.1 Prozent ist hier im Durchschnitt am größten. Bei 1 Prozent führt Int/Avg zu ebenso guten Ergebnissen wie ED/Min. Bei den Datensätzen HIV und Renin liegen die Anreicherungen mit Int/Avg etwas über denen mit ED/Min, allerdings nur bei 1 Prozent durchsuchter Datenbank. Bei allen übrigen Datensätzen ist jeweils ED/Min leicht im Vorteil.

Bei der Auswertung der SIFt-Daten reichert Int/Min die Aktiven am besten an, insbesondere im obersten Teil der Rangliste bei 0.1 Prozent. Die Klassifizierung Int/LDA15 zeigt in diesem Bereich einen deutlich niedrigeren Recall, der jedoch im weiteren Verlauf (also bei wachsendem Anteil durchsuchter Screeningdatenbank) steiler ansteigt als bei den übrigen Auswertungsmethoden. Bei 1 Prozent durchsuchter Datenbank ergibt sich im Mittel bereits ein Vorteil für Int/LDA15. Beispielsweise liegt für den Datensatz Thrombin der Recall der Methode Int/LDA bei 1 Prozent bereits doppelt so hoch wie der bei Anwendung von Int/Min, während bei 0.1 Prozent das Verhältnis noch genau umgekehrt ist (Recall von Int/Min beinahe doppelt so hoch wie der von Int/LDA15). Diese Tendenz der Recall-Kurven setzt sich auch für den weiteren, in den Abbildungen nicht mehr gezeigten Verlauf fort.

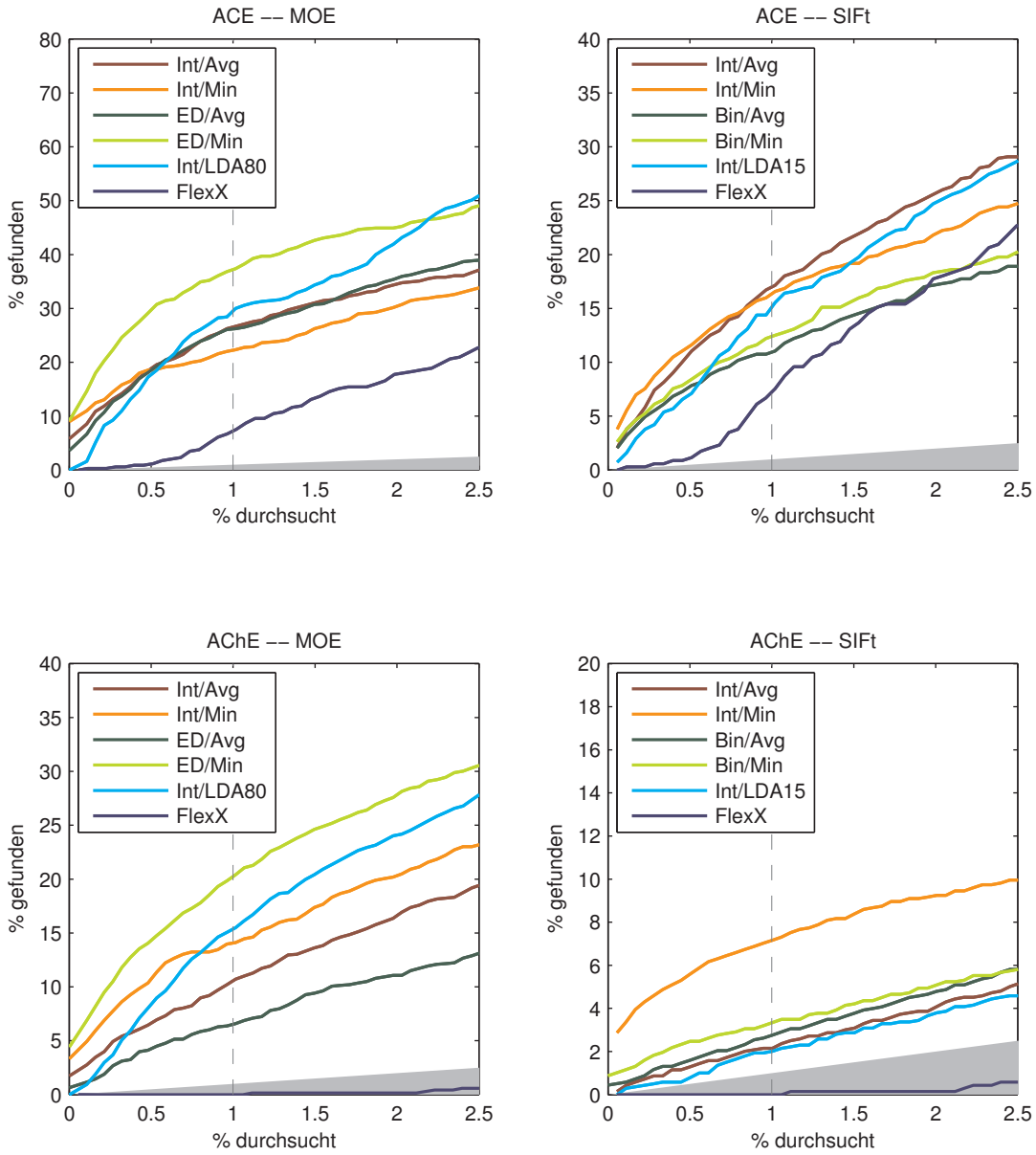


Abbildung 3.14 Recall im Bereich bis 2.5 Prozent gescreenter Datenbank. Links die Ergebnisse der Simulationen auf Grundlage der MOE-Deskriptoren, rechts die der SIFt-basierten Simulationen, jeweils zum Vergleich der mit FlexX erzielte Recall. Die Abbildungen zeigen die Werte für die beiden Datensätze ACE (oben) und AChE (unten).

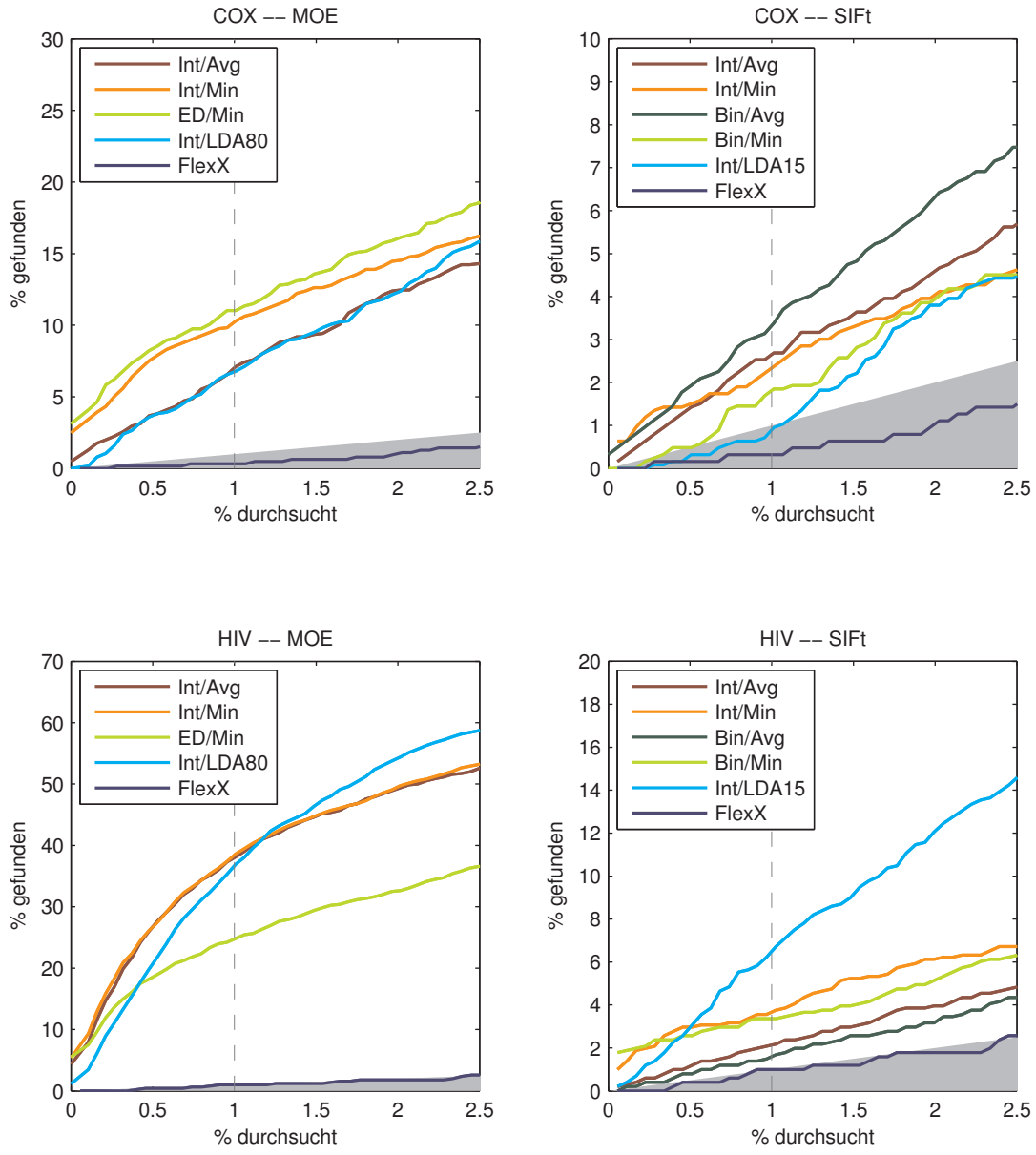


Abbildung 3.15 Recalls der VS-Simulationen für die Datensätze COX und HIV (Erläuterung siehe Abb. 3.14 und Text).

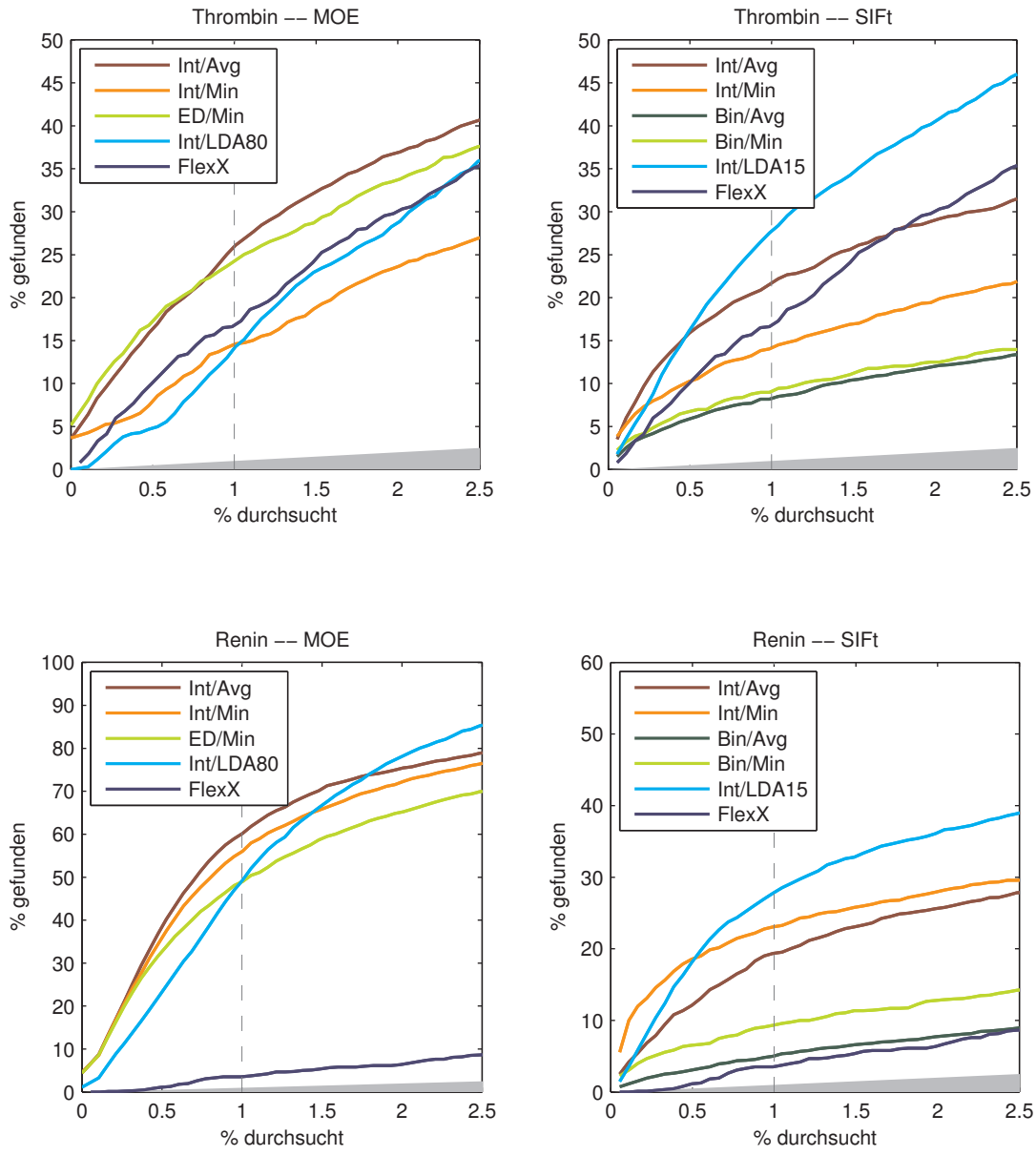


Abbildung 3.16 Recalls der VS-Simulationen für die Datensätze Thrombin und Renin (Erläuterung siehe Abb. 3.14 und Text).

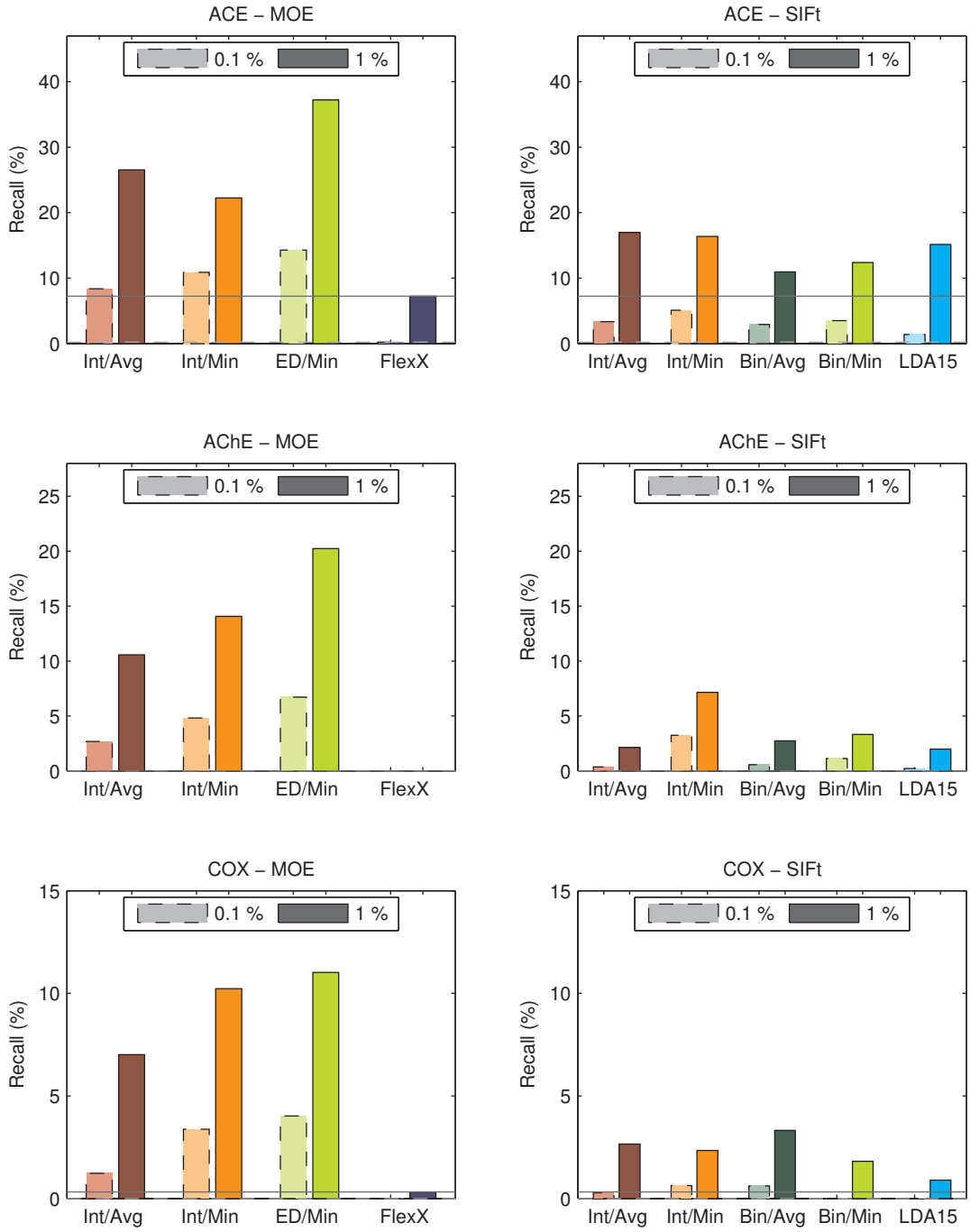


Abbildung 3.17 Recalls bei 1 Prozent (sattere Farben, durchgezogener Rand) und 0.1 Prozent (blassere Farben, gestrichelter Rand) gescreenter Datenbank. Die horizontale Linie gibt zum Vergleich den jeweiligen mit FlexX erzielten Recall an. Hier gezeigt sind die Ergebnisse für die Datensätze ACE, AChE und COX.

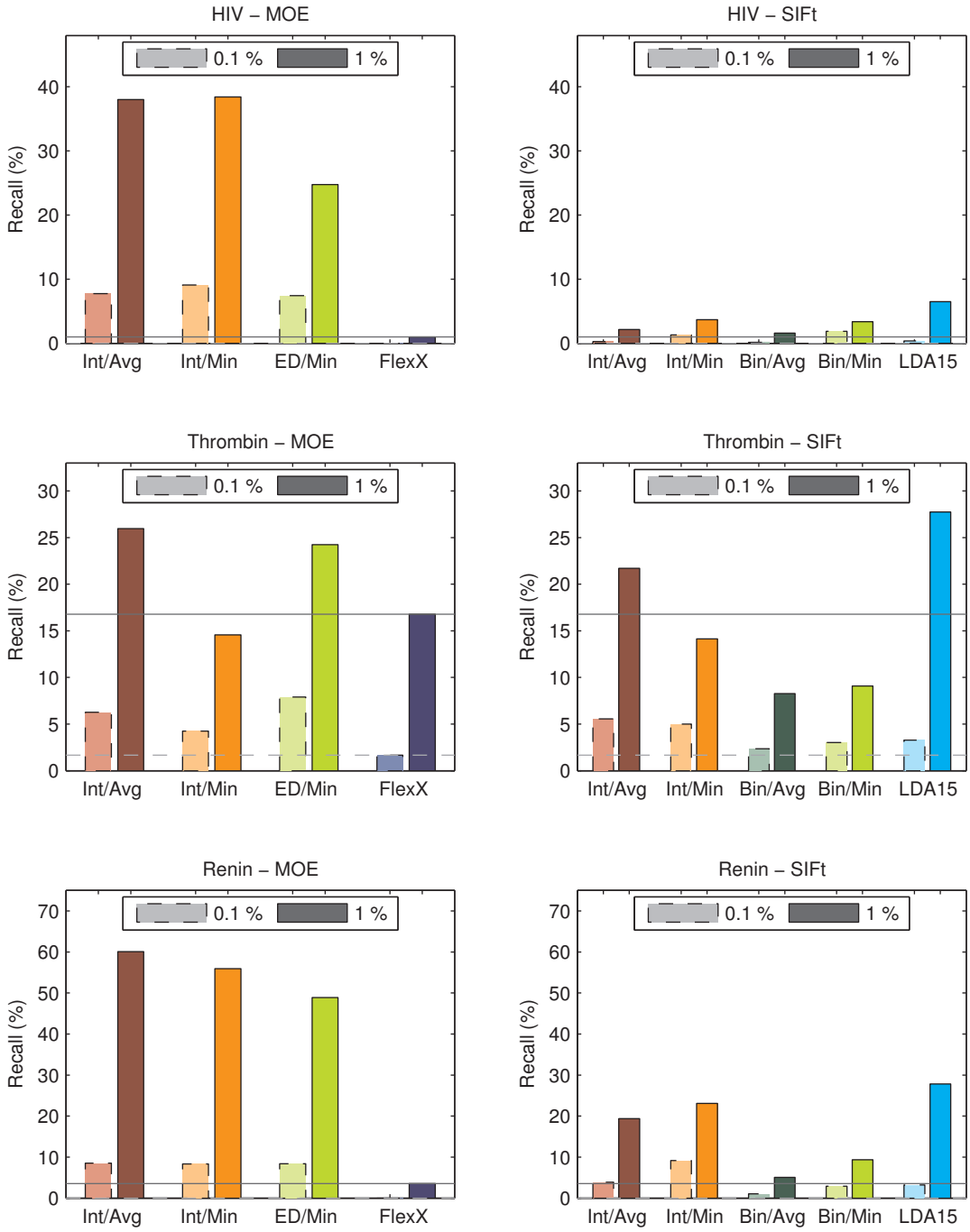


Abbildung 3.18 Recalls für die die Datensätze HIV, Thrombin und Renin bei 1 Prozent und 0.1 Prozent gescreenter Datenbank (Erläuterung siehe Abb. 3.17).

3.4.3.1 FlexX

Das virtuelle Screening auf Basis des FLEXX-Dockings liefert mit Ausnahme von Thrombin und ansatzweise ACE insgesamt enttäuschende Ergebnisse. Für AChE, COX, HIV und den obersten Bereich von Renin ist der Recall geringer als der, der bei einer Zufallsauswahl zu erwarten wäre; hier führt das Screening also sogar zu einer Abreicherung der Aktiven im oberen Teil der Rangliste. Auch bei ACE fällt die Anreicherung bei 0.1 Prozent nur gering aus und nimmt erst im weiteren Verlauf zu.

Einzig für den Datensatz Thrombin erzielt das FLEXX-Screening eine gute (bei 1 Prozent) bzw. zufriedenstellende (bei 0.1 Prozent) Datenbankanreicherung. Der hier gefundene Recall stimmt ungefähr mit der entsprechenden Untersuchung von CUMMINGS *et al.* (ebenfalls FLEXX und ein aus dem MDDR extrahierter Thrombin-Datensatz) überein.[136]

Die schlechte Anreicherung im COX-Datensatz steht in Übereinstimmung mit Erkenntnissen von STAHL *et al.*, denen zufolge FLEXX bei vielen im WDI (*World Drug Index*)[146] enthaltenen Verbindungen zusätzliche (falsche) Wasserstoffbrückenbindungen ermittelt, die noch dazu von der FLEXX-Scoringfunktion recht hoch bewertet werden.[96] Für den ähnlichen, in dieser Dissertation verwendeten MDDR[132] ist dasselbe Phänomen zu erwarten.

3.4.3.2 FlexX/SIFt

Die Verwendung der SIFts führt bei allen Datensätzen zu einer teilweise sehr deutlichen Erhöhung der Datenbankanreicherung im Vergleich zum alleinigen Einsatz von FLEXX; dies gilt zunächst weitgehend unabhängig von der Auswertungsmethode. Bei Thrombin wird allerdings bei 1 Prozent nur mit Int/LDA15 und Int/Avg eine Anreicherung erzielt, die über das gute Ergebnis von FLEXX hinausgeht. Bei 0.1 Prozent jedoch ist der Recall aller SIFt-basierten Methoden höher als der von FLEXX.

3.4.3.3 MOE

Das ligandbasierte Screening auf Grundlage der MOE-Deskriptoren erreicht bei allen untersuchten Datensätzen eine höhere Datenbankanreicherung als das entsprechende FLEXX-Screening. Bis auf den Thrombin-Datensatz liegt

der bei 0.1 Prozent erzielte Recall der MOE-Methoden sogar höher als der 1-Prozent-Recall von FLEXX. Auch diese Feststellung gilt unabhängig von der Auswertungsmethode, wobei erneut die Anreicherung im Datensatz Thrombin bei 1 Prozent die Ausnahme bildet: Hier schneiden nur Int/Avg und ED/Min besser ab als FLEXX. Bei 0.1 Prozent sind dagegen wiederum alle Methoden (außer Int/LDA80, in den Abbildungen nicht gezeigt) besser als FLEXX.

Die Methode Int/LDA80 ist in den Balkendiagrammen (Abb. 3.17 und 3.18) nicht dargestellt, da sowohl bei 0.1 Prozent als auch bei 1 Prozent stets ein anderes Auswertungsverfahren einen höheren Recall liefert. Darüber hinaus schneidet Int/LDA80 nur in einem einzigen Fall (Datensatz HIV bei 1 Prozent) signifikant besser ab als ED/Min, die Methode mit dem im Mittel höchsten Recall (siehe Tab. 3.6).

Für die Auswertung ED/Min wurde zusätzlich untersucht, wie sich die Anzahl von Referenzmolekülen auf die Datenbankanreicherung auswirkt. Erwartungsgemäß liegt der Recall niedriger, wenn für die Ähnlichkeitssuche statt zehn nur ein oder zwei Referenzen definiert werden. Wie Abb. 3.19 (Seite 110) am Beispiel der Datensätze Thrombin, Renin, COX und HIV zeigt, führt jedoch auch eine kleine Anzahl von Referenzen noch zu einer guten Anreicherung, die sich allerdings gegenüber der Standardmethode (zehn Vergleichsmoleküle) je nach Datensatz in unterschiedlichem Ausmaß verschlechtert.

3.4.4 Diskussion der Ergebnisse

3.4.4.1 FlexX

An den erzielten Datenbankanreicherungen fällt zunächst das allgemein schlechte Abschneiden von FLEXX auf. Eine Ausnahme bildet der Thrombin-Datensatz, bei dem der FLEXX-Recall auffällig gut ist. Dieser Datensatz ist der einzige, bei dem die mit FLEXX erzielte Anreicherung annähernd an die Ergebnisse der ligandbasierten Verfahren heranreicht. In den übrigen Fällen erreicht FLEXX lediglich für ACE bei 1 Prozent einen nennenswerten Recall.

Im Fall des Thrombin-Datensatzes ist sicherlich nicht zu unterschätzen, daß bei der Entwicklung von FLEXX der Algorithmus u. a. an genau diesem Tar-

Tabelle 3.5 Recall der SIFt-basierten Auswertungen bei 0.1 Prozent (oberer Teil der Tabelle) und 1 Prozent (unterer Teil). Der Wert der jeweils besten Methode ist fettgedruckt dargestellt. Zum Vergleich enthält die Spalte ganz rechts den mit FlexX erzielten Recall.

Recall bei 0.1 Prozent						
	Int/Avg	Int/Min	Bin/Avg	Bin/Min	Int/LDA15	FlexX
ACE	3.4	**5.1**	2.9	3.5	1.4	0.2
AChE	0.4	**3.3**	0.6	1.1	0.2	0.0
COX	0.3	**0.6**	0.6	0.0	0.0	0.0
HIV	0.3	1.3	0.1	**1.9**	0.3	0.0
Renin	3.9	**9.2**	1.1	2.9	3.2	0.0
Thrombin	**5.6**	5.0	2.4	3.0	3.3	1.7
Recall bei 1 Prozent						
	Int/Avg	Int/Min	Bin/Avg	Bin/Min	Int/LDA15	FlexX
ACE	**17.0**	16.4	10.9	12.4	15.1	7.2
AChE	2.1	**7.2**	2.8	3.3	2.0	0.0
COX	2.7	2.3	**3.3**	1.8	0.9	0.3
HIV	2.1	3.7	1.6	3.4	**6.5**	1.0
Renin	19.4	23.1	5.0	9.3	**27.8**	3.5
Thrombin	21.7	14.1	8.3	9.1	**27.7**	16.8

Tabelle 3.6 Recall der auf den MOE-Deskriptoren basierenden Screenings bei 0.1 Prozent (oberer Teil der Tabelle) und 1 Prozent (unterer Teil).

Recall bei 0.1 Prozent					
	Int/Avg	Int/Min	ED/min	Int/LDA80	FlexX
ACE	8.4	10.9	**14.3**	1.5	0.2
AChE	2.7	4.8	**6.7**	0.9	0.0
COX	1.2	3.4	**4.0**	0.2	0.0
HIV	7.7	**9.1**	7.4	3.3	0.0
Renin	**8.5**	8.3	8.4	3.2	0.0
Thrombin	6.3	4.3	**7.9**	0.3	1.7
Recall bei 1 Prozent					
	Int/Avg	Int/Min	ED/min	Int/LDA80	FlexX
ACE	26.5	22.2	**37.2**	29.5	7.2
AChE	10.6	14.1	**20.2**	15.4	0.0
COX	7.0	10.2	**11.0**	6.7	0.3
HIV	38.0	**38.4**	24.7	36.7	1.0
Renin	**60.1**	55.9	48.9	49.1	3.5
Thrombin	**26.0**	14.6	24.2	14.1	16.8

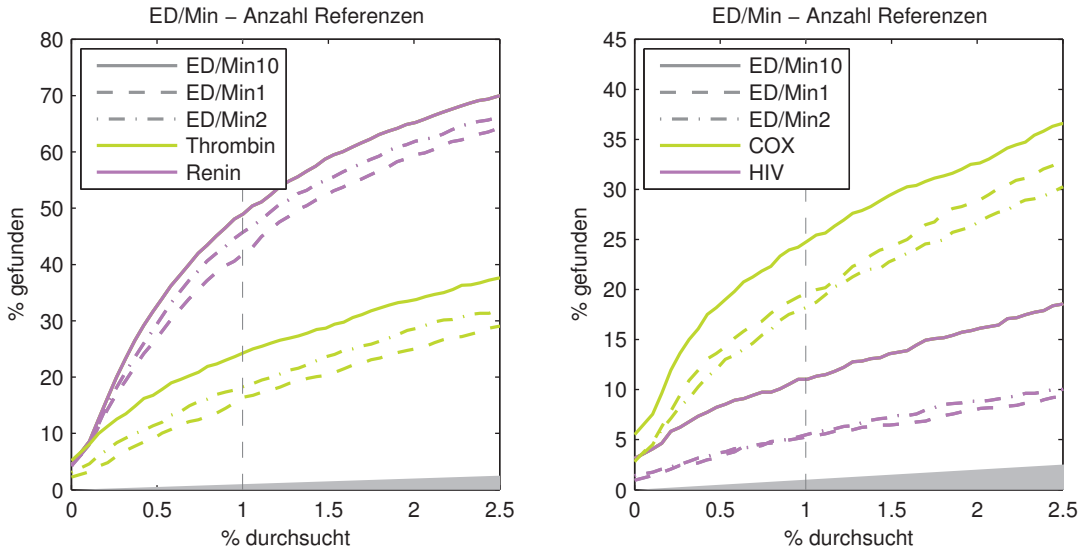


Abbildung 3.19 Einfluß der Anzahl von Referenzmolekülen, hier gezeigt an den Datensätzen Thrombin und Renin (links) bzw. COX und HIV (rechts). Die Kurven geben jeweils den erzielten Recall der Methode ED/Min (MOE-Deskriptoren) an, wenn 1, 2, oder 10 Referenzmoleküle verwendet werden (Standard: 10).

get optimiert wurde,[68] FLEXX also dafür in gewissem Sinne „maßgeschneidert“ ist. Ebenso ist bereits aus der ersten großen Validierung des Programms bekannt, daß es bei HIV-Proteasen keine guten Ergebnisse erzielt.[69]

Darüber hinaus stellen die Charakteristiken der Bindetaschen einen wichtigen Aspekt zur Erklärung der per Docking erzielten Ergebnisse dar. Eine Untersuchung von SCHULZ-GASCH *et al.* geht auf diese Thematik ein und präsentiert einige Erkenntnisse, die auch auf die hier verwendeten Targets zutreffen:[147] So wird etwa im Fall von COX-2 der Bindungsmodus typischer Liganden eher durch ihre allgemeine Gestalt bestimmt als durch Wasserstoffbrückenbindungen. Daher ist es für Dockingalgorithmen wie FLEXX, die einen inkrementellen Aufbau des Liganden verfolgen, schwierig, eine korrekte Basisfragmentplazierung zu finden. Dafür werden nämlich dezidierte, gerichtete Ligand-Rezeptor-Interaktionen benötigt. In solchen Situationen sind also Multikonformer-Algorithmen (z. B. FRED), die den Liganden nicht fragmentieren, möglicherweise besser geeignet.

Die Thrombin-Bindetasche dagegen ist zwar ebenfalls nicht sehr polar, bietet jedoch mit einem Aspartatrest (Asp189) an einer tiefen Stelle der Active

site einen hervorragenden Anker für die Plazierung des Basisfragments. Von dieser Position ausgehend ist der inkrementelle Aufbau des Liganden dann sehr erfolgreich. Umgekehrt kann aber eine größere Anzahl von polaren Gruppen auch dazu führen, daß der Dockingalgorithmus angesichts der Vielzahl guter Interaktionsmöglichkeiten nicht mehr zu unterscheiden vermag, welche Basisfragmentplazierung optimal ist.

Im Fall der Acetylcholinesterase stellen KELLENBERGER *et al.* fest, daß aufgrund der eher hydrophoben Natur der meisten bekannten Inhibitoren diejenigen Dockingprogramme, die ihr Hauptaugenmerk auf polare Interaktionen legen (wie FLEXX), ebenfalls häufig nur ungenügende Ergebnisse liefern.[148] So findet FLEXX bei dem in der vorliegenden Arbeit durchgeführten Docking in die Kristallstruktur 1EVE für den co-kristallisierten Liganden zwar zwölf hydrophobe, aber nur zwei der für die Basisfragmentplazierung wichtigen polaren Interaktionen (ausgehend von Tyr121 bzw. Arg289).

In derselben Publikation[148] wird außerdem auf die oft hohe Flexibilität von Renin- und HIV-Protease-Inhibitoren hingewiesen. In der Tat weisen im vorliegenden Fall die co-kristallisierten Liganden aus 1HRN (Renin) bzw. 4PHV (HIV) 20 bzw. 17 frei drehbare Bindungen auf. Wegen der Vielzahl sich daraus ergebender möglicher Fragmentierungen ist die Anwendung von FLEXX auf solche hochgradig flexiblen Moleküle problematisch.[137,138]

Für 1O86 (ACE) ist zu beachten, daß die Bindetasche ein Zinkatom enthält. Hier kann die Situation eintreten, daß das Metallion sowohl von Rezeptor- als auch von Ligandseite z. B. durch eine Carboxylatgruppe komplexiert wird. Dadurch kommen sich in der Koordinationssphäre des Zinkions die negativen Ladungen der Carboxylatgruppen zwangsläufig recht nahe: Bei 1O86 beträgt der Abstand zwischen den Carboxylat-Sauerstoffatomen von Glu411 und denen des Liganden nur 3.2 Å. Mit Scoringfunktionen, die für kleine Abstände gleichartig geladener Gruppen einen Strafterm vorsehen, wird diese eigentlich sehr günstige Ligand-Rezeptor-Interaktion dann abgewertet.

Im übrigen wurden die im virtuellen Screening erhaltenen Dockinglösungen nicht visuell inspiziert, etwa um nicht sinnvolle aber dennoch mit einem guten Score bewertete Plazierungen auszusortieren. Stattdessen erfolgte das Ranking ausschließlich „blind“ anhand der Dockingscores. Auch wenn dies in Einklang mit der üblichen (und leider oft einzig praktikablen) Vorgehensweise beim Screening von mehrere Millionen Moleküle enthaltenden virtu-

ellen Bibliotheken steht, so widerspricht es doch dem vorhandenen Wissen um die Ungenauigkeiten der Dockingscores. Wie KITCHEN *et al.* in diesem Zusammenhang treffend bemerken, ist es für die effiziente Auswahl von Verbindungen eben nicht ausreichend, sich ausschließlich auf die berechneten Scores zu verlassen: *„Experience and intuition are often still a key to success."*[149]

3.4.4.2 FlexX/SIFt

Die Berechnung der SIFts und ihre Verwendung zum ähnlichkeitsbezogenen Screening der Datenbank führt Information über bekannte aktive Inhibitoren in einen strukturbasierten VS-Ansatz ein. Diese Kombination der Information über die Targetstruktur einerseits und eine Serie aktiver Verbindungen andererseits ergab in den hier präsentierten Simulationen eine deutliche Verbesserung der Anreicherung gegenüber der alleinigen Verwendung von FLEXX. Gleichwohl ist bemerkenswert, daß das Screening auf Grundlage der SIFt-Ähnlichkeit nur in zwei Fällen (Thrombin, 1 Prozent, Int/LDA15 sowie Renin, 0.1 Prozent, Int/Min) zu einer geringfügig höheren Anreicherung führte als das rein ligandbasierte, MOE-Deskriptor-gestützte Screening.

Offenbar wird die in den zehn Referenzverbindungen enthaltene Information also nicht in vollem Umfang genutzt. Die Ursache dafür kann nicht abschließend geklärt werden. Einerseits kodieren die SIFts direkt den Bindungsmodus der Liganden, geben also ein recht realistisches Bild der für die Inhibition notwendigen Merkmale ab. Andererseits hängt der Wert der SIFt-codierten Information natürlich unmittelbar von der Güte der Dockinglösungen ab. Gelingt es dem Docking — etwa aufgrund der oben diskutierten Limitierungen — nicht, einen korrekten Bindungsmodus zu identifizieren, so besitzen auch die resultierenden SIFts kaum Aussagekraft. Somit steht und fällt der SIFt-Ansatz mit der Eignung des Dockingtools bzw. der Scoringfunktion für das betrachtete Target.

Darüber hinaus erfassen die MOE-Deskriptoren natürlich eine Vielzahl weiterer Merkmale der Liganden, die in den SIFts nicht enthalten sind. Die tatsächliche Aktivität einer Verbindung wird nicht nur durch den Bindungsmodus bestimmt, wie er in den SIFts gespeichert ist, sondern beispielsweise auch durch Eigenschaften, wie sie durch die einfachen Lipinski-Regeln[150] (*„Rule-of-five"*) beschrieben werden. Diese sind jedoch nur in den MOE-De-

skriptoren explizit kodiert, nicht aber in den SIFts. Auch wenn dies in der frühen Phase der Leitstrukturfindung vielleicht in realen Sreenings noch nicht in vollem Umfang berücksichtigt wird, stellen solche zusätzlichen Informationen in der Simulation einen wesentlichen Vorteil zum Wiederfinden von bekannten Aktiven dar. Im weiteren Verlauf der Entwicklung dieser Aktiven wurden nämlich genau diejenigen molekularen Eigenschaften optimiert, welche von den SIFts nicht erfaßt werden, wohl aber (zumindest teilweise) von den MOE-Deskriptoren.

Im allgemeinen liefern die Ergebnisse des SIFt-unterstützten Screenings jedoch einen deutlichen Hinweis darauf, daß diese Art der Kombination von ligand- und strukturbasierter Information ein hohes Potential zur Verbesserung der Datenbankanreicherung bietet. Ebenso trägt die Verwendung der SIFts dazu bei, das Problem der nicht vorgenommenen visuellen Beurteilung der Dockinglösungen zumindest abzumildern. Da nämlich in den vorgestellten Experimenten für jeden Liganden der Mittelwert über die SIFts seiner ersten 30 Plazierungen berechnet und für die Ähnlichkeitssuche herangezogen wurde, fallen einzelne nicht sinnvolle Posen weniger stark ins Gewicht.

Der Einsatz der SIFts weist eine gewisse Ähnlichkeit mit einem pharmakophorbeschränkten Docking auf, wie es etwa mit dem Modul FLEXX-PHARM möglich ist. Allerdings wird dabei bereits beim inkrementellen Aufbau des Liganden auf die Einhaltung der vom Benutzer vorgegebenen (aus einem Pharmakophormodell abgeleiteten) Interaktionen geachtet. Im Gegensatz dazu stellt der SIFt-Ansatz ein nachgelagertes Filtern der Dockinglösungen dar, jedoch ebenfalls anhand einer (automatisch generierten) „pharmakophorartigen“ Information.

3.4.4.3 MOE

Das Screening per Ähnlichkeitssuche auf Grundlage der MOE-Deskriptoren liefert durchweg überzeugende Ergebnisse. Im Mittel über alle sechs untersuchten Targets wird bei 0.1 Prozent eine beachtliche bis zu 81fache Anreicherung erzielt, bei 1 Prozent eine bis zu 28fache. Demgegenüber liegen die Anreicherungsraten von FLEXX/SIFt in beiden Fällen nur etwa halb so hoch.

Die Anzahl der verwendeten Referenzmoleküle besitzt zwar einen deutlichen, aber doch erstaunlich geringen Einfluß auf den Recall. Selbst beim Da-

tensatz Thrombin sinkt mit der Auswertungsmethode ED/Min der 0.1-Prozent-Recall nur von 8 % (10 Referenzen) auf 3 % (1 Referenz) und liegt damit noch immer doppelt so hoch wie der FLEXX-Recall; der 1-Prozent-Recall ist mit einer einzelnen Referenz genauso hoch wie der von FLEXX.

Eine Ursache für dieses günstige Verhalten ist möglicherweise in der Diversität der Referenzen zu suchen: Werden annähernd identische Vergleichsmoleküle verwendet, so ist der Informationsverlust beim Übergang von zehn Referenzen auf eine nur marginal und hat kaum Auswirkungen auf die Distanz ED/Min.

In der Praxis besitzt die Ähnlichkeitssuche gegenüber dem Docking einen erheblichen Geschwindigkeitsvorteil. Die Deskriptoren der Moleküle können zusammen mit den Strukturinformationen in der Screeningdatenbank gespeichert werden, müssen also nicht jedesmal neu berechnet werden. Während die benötigte Zeit für das Docking eines Moleküls in der Größenordnung von einigen Minuten liegt, dauert die Berechnung der euklidischen Distanz dann nur Sekundenbruchteile.

3.4.4.4 Einschränkungen

Die vorgestellten Untersuchungen zum strukturbasierten virtuellen Screening sind dadurch eingeschränkt, daß mit FLEXX nur ein einziges Dockingprogramm zur Verfügung stand. Gerade im Hinblick darauf, daß (wie oben erwähnt) die Qualität des Dockingergebnisses von der Wahl eines an die Charakteristik der Bindetasche angepaßten Dockingalgorithmus abhängt, wäre der parallele Einsatz mehrerer Dockingtools wünschenswert gewesen.[151] Gleiches gilt für die Scoringfunktion, die ebenfalls auf das in FLEXX implementierte Schema beschränkt war. Ein *Consensus scoring*, also die Bewertung einer Dockinglösung mit mehreren verschiedenen Scoringfunktionen, wäre sicherlich ein interessanter Aspekt gewesen.[152] Gleichwohl kommen CUMMINGS *et al.* in einer Studie zum dockinggestützten VS an mehreren Targets zu der Aussage, daß insgesamt die Anzahl identifizierter Hits bei Verwendung verschiedener Dockingprogramme recht ähnlich ist.[136] Dies steht jedoch nicht im Widerspruch zu der Erkenntnis, daß an einem einzelnen Target durch die Auswahl eines für die Charakteristik der Bindetasche adäquaten Dockingtools das Ergebnis optimiert werden kann.

Eine weitere Limitierung des simulierten VS stellt die Wahl des MDDR als Screeningdatenbank dar. Aufgrund seiner Entstehungsgeschichte entspricht der MDDR als Kollektion bereits auf dem Markt oder noch in Entwicklung befindlicher Wirkstoffe bzw. -kandidaten in seiner Zusammensetzung sicherlich nicht einer virtuellen Bibliothek, wie sie in industriellen Forschungsprojekten üblicherweise gescreent wird. Eine Untersuchung zum dockinggestützten virtuellen Screening von VERDONK *et al.* gibt jedoch Hinweise darauf, daß als simulierte Screeningdatenbank eine fokussierte Bibliothek verwendet werden sollte, deren Moleküle ähnliche Eigenschaften aufweisen wie die bekannten Aktiven.[153] Darüber hinaus schlagen GOOD *et al.* eine andere Art der Validierung vor: Nicht die bloße Datenbankanreicherung etwa in Form des Recalls sei zur Beurteilung der Leistungsfähigkeit des virtuellen Screenings zu betrachten, sondern auch die Eignung des Verfahrens, im Sinne eines Scaffold-Hoppings aktive Verbindungen mit neuen Strukturmotiven zu identifizieren.[154]

Schließlich ist zu beachten, daß die Ergebnisse einer großen, in der Praxis verwendeten Screeningdatenbank mit mehreren Millionen Verbindungen nicht beliebig auf den kleineren Maßstab einer Simulation skalierbar sind. So entsprechen im vorliegenden Fall die obersten 0.1 Prozent der simulierten Screeningdatenbank nur etwa 93 Molekülen, was zu einer gewissen Instabilität der Ergebnisse führt. Aus diesem Grund besitzt in der hier vorgestellten Simulation der Recall bei 1 Prozent eine höhere Relevanz als bei einer großen Bibliothek, wie sie in der Wirklichkeit eingesetzt wird.

3.4.5 Schlußfolgerungen

Die im Rahmen der hier durchgeführten Simulationen erzielten Datenbankanreicherungen sehen den ligandbasierten Ansatz deutlich im Vorteil gegenüber dem strukturbasierten virtuellen Screening. Auch unter Berücksichtigung der diskutierten Limitationen der vorliegenden Untersuchung ist die in der Praxis derzeit vorherrschende Fokussierung auf strukturbasierte Verfahren kritisch zu hinterfragen. Deren Herangehensweise ist meist mit einem höheren Aufwand verbunden, ohne bei rein numerischer Betrachtung der Anreicherung bessere Ergebnisse zu liefern.

Dessen ungeachtet bietet der strukturbasierte Ansatz einen klaren Vorteil hinsichtlich des Erkenntnisgewinns über das zugrundeliegende System: Das Docking erweist sich als äußerst nützlich für das Verständnis der möglichen Ligand-Rezeptor-Interaktion verschiedener chemischer Strukturmotive.[155] Im Gegensatz zu den ligandbasierten Techniken bietet es darüber hinaus die Möglichkeit zur Identifizierung wirklich neuartiger aktiver Verbindungen, ohne daß durch bereits bekannte Hits oder Leads eine Einschränkung oder Verzerrung im Strukturraum zu erwarten ist.[149] Das ligandbasierte Screening erweist sich zwar als sehr effizient bei der Suche nach Molekülen, die zu gegebenen Vergleichsverbindungen eine (durchaus hochkomplex definierte) Ähnlichkeit aufweisen; dennoch bleibt es — im Unterschied zum Docking — in seiner Aussagekraft beschränkt auf den durch die Referenzmoleküle abgesteckten Merkmalsraum.

Im Zusammenhang mit vergleichenden Studien zum ligand- und strukturbasierten Screening wird immer wieder die Frage nach „fairen" Bedingungen hinsichtlich der zur Verfügung stehenden Information aufgeworfen. Zunächst bleibt festzuhalten, daß beide Verfahren eine jeweils exklusive Information nutzen — die Struktur des Targets bzw. die Eigenschaften bereits bekannter Aktiver —, deren Wert sich kaum allgemeingültig beurteilen läßt. Weiterhin stellt sich der mit dieser Frage ausgedrückte Vorbehalt aus dem Blickwinkel der praktischen Wirkstoffentwicklung als irrelevant dar (was auch an entsprechenden vergleichenden Studien in der Literatur[155] deutlich wird): In der Praxis findet sich der Computerchemiker de facto von Zeit zu Zeit in der komfortablen Situation wieder, für ein virtuelles Screening sowohl auf struktur- als auch auf ligandbasierte Information zurückgreifen zu können. Spätestens an diesem Punkt weicht die Frage nach einem fairen Methodenvergleich der ergebnisorientierten Implementierung eines möglichst effizienten und erfolgreichen Screenings.

Für die konkrete Vorgehensweise erscheint eine Abwägung anhand der gegebenen Zielsetzung sinnvoll: Für das Auffinden neuer, patentierbarer Leitstrukturen besitzt das docking-/strukturbasierte Screening prinzipiell weitergehende Möglichkeiten. Insbesondere für spätere, eher auf die Leitstrukturoptimierung ausgerichtete Entwicklungsphasen hingegen steht mit dem ligandbasierte Ansatz ein einfach zu handhabendes und effizientes Werkzeug mit hohen Anreicherungsraten zur Verfügung. Ein großes Potential

verspricht die Kombination der beiden Herangehensweisen, beispielsweise durch Anwendung der SIFts: Die Fähigkeit des Dockings zum Auffinden neuer Chemotypen wird hier komplementär ergänzt durch den bekanntermaßen erfolgversprechenden Bindungsmodus, der aus den aktiven Referenzstrukturen extrahiert wurde.

Dazu wurde in dieser Arbeit ein vereinfachter SIFt-Ansatz so implementiert, daß die Interaktions-Fingerprints unter ausschließlicher Verwendung der von FLEXX generierten Daten automatisiert berechnet werden konnten. Die hier gezeigten Simulationen stellen die erstmalige Anwendung im großen Maßstab — an einer fast 100 000 Moleküle umfassenden virtuellen Screeningdatenbank und sechs verschiedenen Targets — dar.

3.4.5.1 Ausblick

Es ist anzunehmen, daß die Ergebnisse des ligandbasierten virtuellen Screenings stark von der Wahl der Deskriptoren abhängig sind. In der vorliegenden Studie wurde aufgrund bestehender Erfahrungen der MOE-Deskriptorsatz verwendet. Es bleibt jedoch offen, wie sich die gezielte Auswahl einer Submenge von MOE-Deskriptoren oder auch der Einsatz anderer Deskriptoren auf die Datenbankanreicherung auswirkt. Auch die Eignung klassischer Fingerprints zur Beschreibung der Moleküle wurde hier nicht untersucht.

Beim strukturbasierten Screening ist für den Einsatz eines Consensus-Dockings und -Scorings, also die parallele Anwendung mehrerer Dockingprogramme und Scoringfunktionen, eine Verbesserung der Anreicherung zu erwarten. Eine genauere Untersuchung dieses Effekts, wiederum im Vergleich mit den ligandbasierten Techniken, wäre wünschenswert.

Die Ähnlichkeitssuche anhand der SIFts stellt nur eine Möglichkeit zur Verknüpfung von ligand- und strukturbasierter Information dar. Hier sind, etwa durch die Einbindung bestehender Pharmakophormodelle[156], auch andere Herangehensweisen denkbar. Unabhängig davon sollte auch an der Beantwortung der Frage gearbeitet werden, ob und wie durch Analyse des Targets, der Referenzverbindungen und/oder der Screeningdatenbank eine Aussage darüber möglich ist, welcher Ansatz im konkreten Fall den größeren Erfolg verspricht — das ligand- oder das strukturbasierte Screening.

3.5 Kovalentes Docking von Aziridinen als Cathepsin-Inhibitoren

Aus einer vorangehenden Arbeit[157] lag eine Serie N-acylierter Aziridin-2,3-dicarboxylate bzw. -dicarbonsäureester als Cathepsin-Inihibitoren vor. Die im folgenden vorgestellten Dockingexperimente gehen der Frage nach der Ursache ihrer Aktivität sowie der Selektivität bezüglich Cathepsin B und L nach und liefern Hinweise zum gezielten Design entsprechender Liganden.

Die kovalente Ligand-Rezeptor-Bindung stellte dabei eine besondere Herausforderung dar, weil die verfügbaren Dockingprogramme diesen Fall kaum adäquat behandeln können. Aus diesem Grund wurde ein neues FLEXX-Dockingprotokoll entwickelt, das an den bearbeiteten Komplexen zu guten Resultaten führte und auch darüber hinaus für ein breites Anwendungsspektrum geeignet ist.

3.5.1 Entwicklung eines FlexX-Dockingpotokolls

Ausgehend von den Betrachtungen zum zweistufigen Inhibitionsmechanismus in 2.4.5 liegt es nahe, die Aziridin-Liganden in ihrer ursprünglichen ringgeschlossenen Form zu docken. Sofern der Dockingalgorithmus dabei eine Plazierung findet, bei der der Aziridinring ausreichend nahe am aktiven Cystein liegt, besteht die Möglichkeit zur Ausbildung der kovalenten Bindung. Die Scoringfunktion berücksichtigt bei diesem Ansatz natürlich nur die nichtkovalenten Interaktionsenergien; dies gilt aber für alle Liganden gleichermaßen, so daß sich keine Einschränkung hinsichtlich des Vergleichs der Liganden untereinander ergibt.

Eine erfolgreiche Plazierung der Liganden in der Bindetasche wird bei dieser Herangehensweise jedoch paradoxerweise genau durch das nukleophile Schwefelatom des aktiven Cysteins vereitelt. Bei üblichen Dockingvorgängen sieht der Algorithmus nämlich einen Strafterm vor, der Plazierungen verhindern soll, in denen sich Ligand- und Rezeptoratome zu nahe kommen. Erlaubt sind dann nur solche nicht-kovalente Interaktionen, die einen Abstand zwischen Rezeptor und Ligand wahren, der größer als die entsprechende kovalente Bindungslänge ist. Wird die Distanz dagegen geringer, spricht man

von einem *Clash*; eine solche Lösung wird vom Dockingalgorithmus sofort als ungültig erkannt und verworfen.

Damit der Aziridin-Ligand also überhaupt im Abstand von etwa einer S–C-Bindungslänge (ca. 1.8 Å) zum Cystein-Schwefelatom plaziert werden kann, muß eine Anpassung vorgenommen werden, die den Clash verhindert bzw. dazu führt, daß er vom Dockingalgorithmus ignoriert wird. Da eine entsprechende Modifikation des Algorithmus und der Scoringfunktion kaum praktikabel ist, wurde stattdessen das Schwefelatom aus der Bindetasche entfernt. Somit steht für den Liganden genug Raum zur Verfügung, um ohne Clash plaziert werden zu können.

3.5.2 Validierung des Dockingprotokolls

Das vorgestellte Dockingprotokoll erscheint zunächst paradox, da genau das Atom, das auf der Rezeptorseite an der charakteristischen kovalenten Bindung beteiligt ist, entfernt wird. Die Erklärung liegt jedoch in der o. g. zweistufigen Betrachtung, daß zunächst der ring-geschlossene Aziridinring an der richtigen Stelle vorfixiert werden muß (nicht-kovalenter Schritt), bevor die S–C-Bindung geknüpft werden kann (kovalenter Schritt). Durch die Entfernung des Schwefelatoms wird also die Simulation des ersten Schritts ermöglicht, während der zweite Schritt (für den das Schwefelatom essentiell wäre) ohnehin nicht direkt vom Dockingprogramm dargestellt werden kann.

Zur Validierung dieses Ansatzes wurde in gleicher Weise eine Reihe von Dockingexperimenten durchgeführt. Dabei wurden u. a. Epoxid-Verbindungen als Liganden verwendet, die in Analogie zu den eigentlich untersuchten Aziridin-Strukturen stehen, für die aber auch die Röntgenkristallstrukturen von Co-Kristallisaten zur Verfügung stehen. Dadurch ist eine Überprüfung des Dockingansatzes anhand der experimentellen Kristallstruktur möglich: Gelingt es, mit der vorgestellten Vorgehensweise die Position des einkristallisierten Epoxid-Liganden im Dockingexperiment zu reproduzieren, so ist davon auszugehen, daß derselbe Ansatz auch für das Docking der analogen Aziridine geeignet ist. Neben Epoxiden als Liganden existiert auch eine Anzahl von Cysteinprotease-Kristallstrukturen mit co-kristallisierten kovalent gebundenen Fluoromethylketonen bzw. Vinylsulfonen; auch für diese Ligandklassen wurden entsprechende Dockingexperimente durchgeführt.

Tabelle 3.7 Validierung des Dockingprotokolls: RMSD-Werte der jeweiligen Dockinglösung bezüglich der Position des co-kristallisierten Liganden („Top20“: Die 20 Posen mit dem besten FlexX-Score, „Alle“: Alle Posen).

PDB-Code	Enzym	Ligand	Bester RMSD [Å] Top20	Alle
1AIM[159]	Cruzain	Fluoromethylketon	1.9	2.2
1ATK[160]	Cathepsin K	Epoxid (E-64)	1.4	1.5
1CSB[161]	Cathepsin B	Epoxid (Ca030)	1.7	5.5
1ITO[162]	Cathepsin B	Epoxid (E-64c)	1.7	4.6
1MEM[163]	Cathepsin K	Vinylsulfon (APC3328)	1.7	1.7
1QDQ[164]	Cathepsin B	Epoxid (Ca074)	1.8	1.9

Im allgemeinen wird ein Dockingexperiment als erfolgreich angesehen, wenn die mittlere Abweichung (RMSD) zwischen der im Docking gefundenen Plazierung des Liganden und seiner Position in der Kristallstruktur nicht mehr als etwa 2 oder 2.5 Ångström beträgt.[68,158] Tab. 3.7 zeigt sechs Komplexe, für die dieses Kriterium mit dem beschriebenen Dockingprotokoll erfüllt werden konnte. Für jeden Komplex sind zwei RMSD-Werte angegeben: Derjenige der Dockinglösung mit dem besten RMSD-Wert unter den ersten 20 nach Dockingscore sortierten Lösungen sowie derjenige der Lösung mit dem insgesamt besten Dockingscore. Der erstgenannte RMSD-Wert entspricht also dem in der Praxis häufig angewandten Verfahren, die ersten (nach Score sortierten) Plazierungen manuell in Augenschein zu nehmen und nicht nur nach dem Score zu beurteilen. Zur Veranschaulichung ist zusätzlich die Überlagerung von gedockter und co-kristallisierter Struktur dreier Komplexe in Abb. 3.20 dargestellt.

Bei zwei weiteren Komplexen (hier nicht gezeigt) lag der RMSD noch im akzeptablen Bereich zwischen 2.5 und 3.0 Å, weitere zwei wiesen einen hohen RMSD von 4.8 bzw. 6.8 Å auf (1PE6 bzw. 1PPP). Die Differenzen resultierten dabei meist aus der abweichenden Plazierung von Seitengruppen. Das für die kovalente Bindung entscheidende elektrophile Zentrum (also z. B. C1/C2 des Epoxid-Rings) wurde jedoch bei fast allen Experimenten korrekt in unmittelbarer Nähe des aktiven Cysteins plaziert. Der Median der RMSD-Werte aller zehn untersuchten Komplexe liegt bei 1.85 Å. Somit ist das Dockingprotokoll mit der Entfernung des aktiven Cystein-Schwefelatoms aus der Bindetasche und Verwendung der ring-geschlossenen Liganden valide, sei-

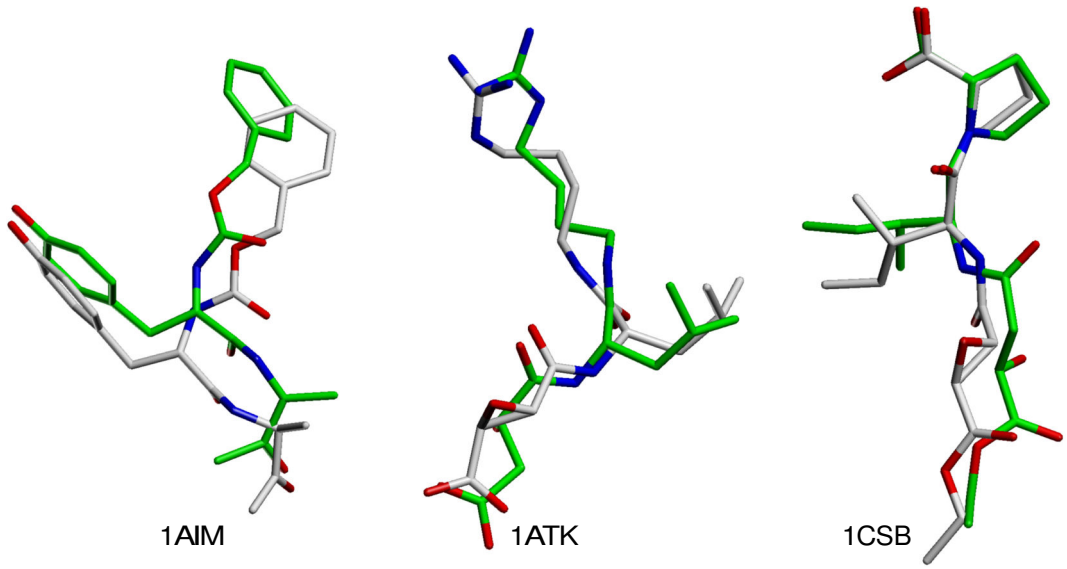


Abbildung 3.20 Überlagerung von Kristallstruktur (grün) und Dockinglösung (grau) für die drei Komplexe 1AIM, 1ATK und 1CSB. Jeweils dargestellt ist die Lösung mit dem bestem RMSD unter den Top20 (FlexX-Score).

ne Anwendbarkeit auf das Docking der Aziridine kann aufgrund der hohen strukturellen Analogie als gesichert gelten. Der Vorteil dieses neuen Ansatzes liegt in seiner universellen Anwendbarkeit — es werden keine vom Benutzer vorgegebenen und damit verzerrenden Nebenbedingungen benötigt. Außerdem können die Standardeinstellungen des Dockingprogramms verwendet werden, d. h. ein Eingriff in den meist sehr fein abgestimmten und sensiblen Dockingalgorithmus ist nicht notwendig.

3.5.3 Anwendung des Dockingprotokolls

3.5.3.1 Aufbereitung der Strukturen

Die Eingangsstrukturen der Liganden wurden mit SYBYL[165] energieminimiert. Dazu wurden das Tripos-Kraftfeld, 2000 Iterationen und der Powell-Minimierer verwendet.

Als Targets für Cathepsin B bzw. L wurden die Röntgenkristallstrukturen mit den PDB-Codes 1GMY[166] bzw. 1MHW[167] verwendet. 1GMY entspricht dem Enzym, das auch im fluorimetrischen Assay zur Bestimmung der Inhi-

bitionskonstanten zum Einsatz kam. Für Cathepsin L wurde im Assay das Enzym aus *Paramecium tetraurelia*[168] verwendet, für das allerdings keine Kristallstruktur vorliegt; daher mußte im Docking auf das entsprechende humane Enzym (1MHW) zurückgegriffen werden. Vorangehende Studien zeigen jedoch, daß die Inhibitoren an beiden Enzymen gleiche Aktivitäten aufweisen.[78]

Die Bindetasche wurde mit einem Radius von 12 Å um das aktive Cystein (Cys29 bei 1GMY bzw. Cys25 bei 1MHW) definiert. Das Schwefelatom dieses Cysteinrests wurde gelöscht, ebenso alle Wassermoleküle. Den positiv geladenen Histidingruppen der Bindetasche wurde in FLEXX das entsprechende Templat `his+` zugewiesen.

Zur Vergrößerung der Menge intern evaluierter Plazierungen wurden alle Dockingexperimente mit den Optionen `set solutions_per_it` 800 und `set solutions_per_frag` 400 durchgeführt.

3.5.3.2 Anpassungen von FlexX an die spezielle Aziridin-Geometrie

Die hier gedockten Aziridine weisen eine strukturelle Besonderheit auf, die eine von den Standardeinstellung abweichende Anpassung von FLEXX erforderlich machte. Die Amidbindung, an der das Stickstoffatom des Aziridinrings beteiligt ist, weist nämlich nicht wie üblich eine planare Konformation auf.[169] Vielmehr zeigen quantenchemische Berechnungen, daß der Winkel zwischen der Aziridin-Ringebene (CCN) und der Ebene der Carbonylgruppe (C=O) etwa 115° bis 130° beträgt (statt 180° wie bei planaren Amidbindungen). Die niedrige Energiebarriere von nur $4\,\mathrm{kcal \cdot mol^{-1}}$ zwischen dem gewinkelten und dem planaren Zustand spiegelt dabei den Hybridisierungszustand des Aziridin-Stickstoffatoms wider, der zwischen sp^2- und sp^3-Hybridisierung liegt.

Da dieser geringe Energiebetrag bei der Bindung des Liganden an den Rezeptor leicht kompensiert werden kann, müssen beim Docking beide Zustände berücksichtigt werden. Deshalb wurden alle Liganden nacheinander sowohl in der sp^2- als auch in der sp^3-hybridisierten Form gedockt. Dies wurde realisiert, indem in den Eingangsstrukturen einmal der Atomtyp `N.3` (sp^3) und einmal `N.am` (sp^2) gewählt wurde.

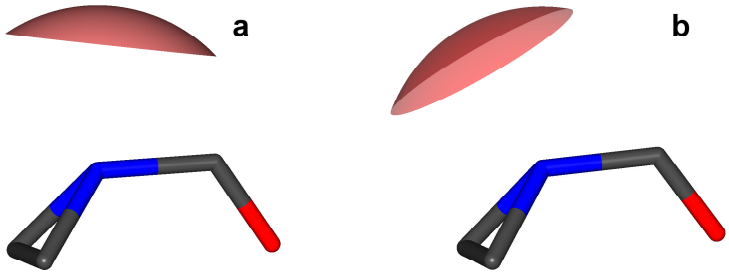


Abbildung 3.21 Interaktionsgeometrie des sp^3-hybridisierten (`N.3`) Aziridin-Stickstoffatoms (a) vor und (b) nach der Anpassung von FlexX. Durch die Modifikation wird eine realistischere Ausrichtung der H-Akzeptor-Funktion erreicht.

FLEXX benutzt jedoch zur Bestimmung der Interaktionsmöglichkeiten des Liganden eine interne Bibliothek von Molekülfragmenten, mit denen der Ligand verglichen wird. Aufgrund der genannten strukturellen Besonderheiten war es notwendig, die Wechselwirkungsgeometrie der H-Akzeptor-Funktion des sp^3-hybridisierten Aziridin-Stickstoffatoms anzupassen. Dazu wurde in der Datei `geometry.dat` die entsprechende Definition `cone e -1 -0.55 -0.55 0 40` eingefügt und in `contact.dat` ein `@subgraph` definiert, der auf das sp^3-hybridisierte Strukturelement zutrifft (als Vorlage wurde der entsprechende Amino-`@subgraph` benutzt). Abb. 3.21 zeigt diese angepaßte Interaktionsgeometrie.

Aus der speziellen Geometrie ergab sich ein weiteres Problem: Zur Erzeugung flexibler Ringkonformationen verwendet FLEXX das Programm CORINA, das für alle aliphatischen Ringsysteme die möglichen Konformationen berechnet. Dabei würde jedoch die gewinkelte Geometrie des Aziridin-Stickstoffatoms verlorengehen. In FLEXX kann die Erzeugung von Ringkonformationen allerdings nur global ein- und ausgeschaltet werden, nicht aber gezielt für einen einzelnen Ring. Um die quantenmechanisch berechnete Konformation des Aziridinrings wie in der Eingangsstruktur vorgegeben beizubehalten und dennoch die konformelle Flexibilität der übrigen Ringsysteme zu gewährleisten, wurde der Bindungstyp der Aziridin-C–C-Bindung auf un (undefined; statt `1`, Einfachbindung) gesetzt. Dies führt für den so modifizierten Aziridinring zum Programmabsturz von CORINA, beeinträchtigt jedoch weder die Konformationsberechnung der übrigen Ringsysteme, noch den Dockingalgorithmus oder das Scoring.

3.5.3.3 Gedockte Liganden

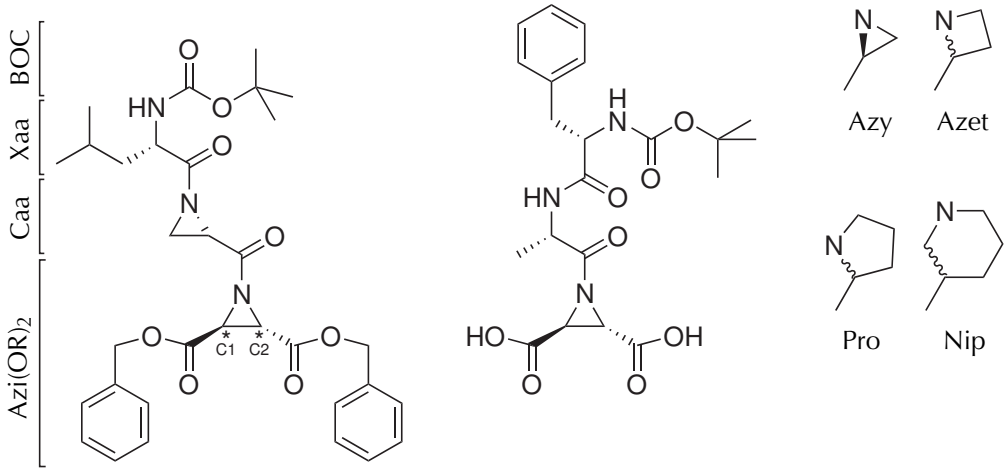
Die folgenden Inhibitoren wurden in CB und CL gedockt (siehe Tab. 3.8): **18a** mit nanomolarer Aktivität an CL und mikromolarer Aktivität an CB, sein Diastereomer **18e**, das nur CL inhibiert, die Phe-Ala-enthaltende Disäure **20a**, die an beiden Enzymen die gleiche Aktivität zeigt, und die Leu-Pro-enthaltenden Dibenzylester **13b**, **13c**, **13e** (CL-inhibierend). Die Disäure **20a** wurde ebenso als freie Säure, als Dicarboxylat-Dianion wie auch in Form der beiden Monoanionen gedockt. Zusätzlich wurden die Azet-enthaltenden* Verbindungen **11c**, **11d**, **12a** und **12b** sowie die Nip-enthaltenden† Moleküle **16a**, **16b**, **16e** und **16f** gedockt, die ausschließlich CL inhibieren. Da die letztgenannten Substanzen im Assay nicht diastereomerenrein getestet wurden, wurde das Docking für jeweils beide Diastereomere durchgeführt.

3.5.4 Ergebnisse

Für **18a** zeigt sich unabhängig von der Konfiguration (*R*/*S*) und vom Hybridisierungszustand (sp^2/sp^3) des Aziridin-Stickstoffatoms ein gemeinsamer Bindungsmodus an CL. Der Inhibitor bindet sowohl in die *Primed* als auch in die *Non-primed site* (siehe Abb. 3.22, Seite 126). Entweder eine Benzylester- oder die Isopropylgruppe des N-terminalen BOC-Leu-Rests liegt in der hydrophoben S2-Tasche, die aus den Aminosäuren Met70, Ala135, Ala214 und Met161 besteht (siehe Abb. 3.23, Seite 127). Der zweite Benzylester (die BOC-Gruppe im Fall der planaren Struktur) bindet in die S1'-Tasche (Ala138, Asp162, His163, Trp189). Die BOC-Leu-Azy-Sequenz interagiert mit der S2'-Bindetasche (Gln19, Gly20, Gln21, Cys22, Gly23) durch Wasserstoffbrückenbindungen zu Gln19 (NH_2 der Seitengruppe), Gly20 (CO des Backbones) oder Cys22 (NH des Backbones). Der für die *S*-Konfiguration (sp^3-hybridisiertes Aziridin-Stickstoffatom) gefundene Bindungsmodus erscheint am plausibelsten, da er durch eine große Anzahl von Wasserstoffbrückenbindungen (sieben H-Brücken, hauptsächlich im Bereich der S2'-Tasche) sowie hydrophobe Interaktionen der OBn-Gruppen (zu Met70, Trp189 und Ala138) stabilisiert wird.

*Azet = Azetidin-2-carbonsäure
†Nip = Nipecotinsäure

Tabelle 3.8 Aziridinyltripeptide mit der allgemeinen Struktur BOC-Xaa-Caa-Azi$(OR)_2$; exemplarisch gezeigt ist Verbindung **18a**. Davon abweichend folgt Verbindung **20a** der allgemeinen Sequenz BOC-Phe-Ala-(*S,S*)-Azi$(OR)_2$. Die beiden rechten Spalten der Tabelle geben jeweils die in einem fluorimetrischen Assay bestimmten Inhibitionskonstanten an CL bzw. CB an („k.I.": keine Inhibition, $K_i > 140\,\mu M$).



18a **20a**

#	Xaa	Caa	C1/C2	R	K_i CL (μM)	K_i CB (μM)
11c	(*S*)-Leu	(*R*)-Azet	(*R,R*)	Et	4.8 ± 0.3	k.I.
11d	(*S*)-Leu	(*S*)-Azet	(*R,R*)	Et	4.8 ± 0.3	k.I.
12a	(*S*)-Leu	(*R*)-Azet	(*S,S*)	Bn	3.8 ± 0.2	k.I.
12b	(*S*)-Leu	(*S*)-Azet	(*S,S*)	Bn	3.8 ± 0.2	k.I.
13b	(*S*)-Leu	(*R*)-Pro	(*S,S*)	Bn	6.0 ± 0.8	k.I.
13c	(*S*)-Leu	(*S*)-Pro	(*R,R*)	Bn	0.4 ± 0.2	115 ± 18
13e	(*R*)-Leu	(*S*)-Pro	(*S,S*)	Bn	4.0 ± 0.2	k.I.
16a	(*S*)-Leu	(*R*)-Nip	(*S,S*)	Bn	4.4 ± 0.4	k.I.
16b	(*S*)-Leu	(*S*)-Nip	(*S,S*)	Bn	4.4 ± 0.4	k.I.
16e	(*R*)-Leu	(*R*)-Nip	(*S,S*)	Bn	4.2 ± 0.4	k.I.
16f	(*R*)-Leu	(*S*)-Nip	(*S,S*)	Bn	4.2 ± 0.4	k.I.
18a	(*S*)-Leu	(*S*)-Azy	(*S,S*)	Bn	0.013 ± 0.001	9.4 ± 1.1
18e	(*R*)-Leu	(*S*)-Azy	(*S,S*)	Bn	6.4 ± 0.4	k.I.
20a					15.3 ± 1.0	18.3 ± 0.7

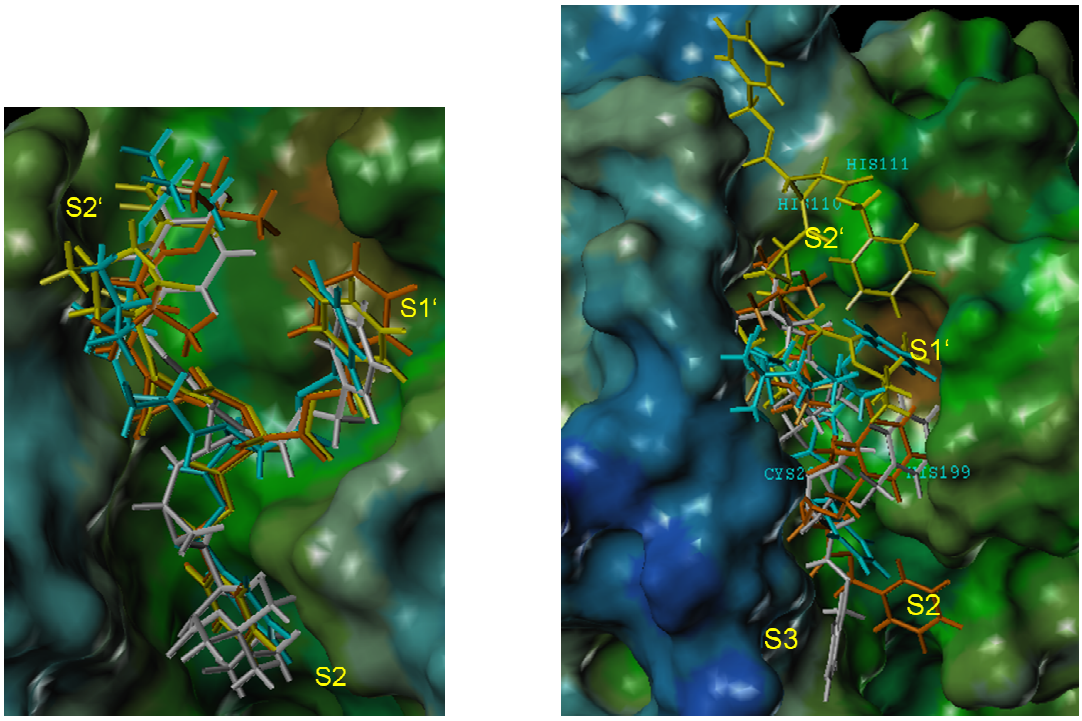


Abbildung 3.22 Überlagerung der Dockingergebnisse von **18a** an CL (links) und CB (rechts). Gelb: $\mathbf{18a_R}$ (N: sp^3), orange: $\mathbf{18a_S}$ (N: sp^3), cyan: $\mathbf{18a_R}$ (N: sp^2), weiß: $\mathbf{18a_S}$ (N: sp^2).

Allen im Docking an CB erhaltenen Plazierungen von **18a** fehlt dagegen mindestens eine der Interaktionen, die an CL gefunden wurden. Es werden also nur zwei der drei Bindetaschen (S2, S1', S2') adressiert. Dies ist möglicherweise durch die Tatsache begründet, daß die S'-Taschen von CB enger sind und somit nicht genügend Raum für zwei große hydrophobe Gruppen bieten. Weiterhin wurde an CB im Gegensatz zu CL kein gemeinsamer Bindungsmodus für die verschiedenen Konfigurationen von **18a** gefunden.

Gemäß den Dockingergebnissen für **18e** bindet dieser Ligand ebenfalls in die Primed und in die Non-primed site von CL. Aufgrund der gegenüber **18a** invertierten Konfiguration der Leu-Einheit paßt jedoch die Isopropylgruppe der Seitenkette nicht optimal in die S2'-Tasche von CL, was wahrscheinlich die Ursache für die geringere Affinität dieses Diastereomers darstellt.

Im Gegensatz zu **18a** und **18e** sagen die in Abb. 3.24 (Seite 128) gezeigten Dockingergebnisse für **20a** die Bindung in die S2-S3-Tasche von CL vorher

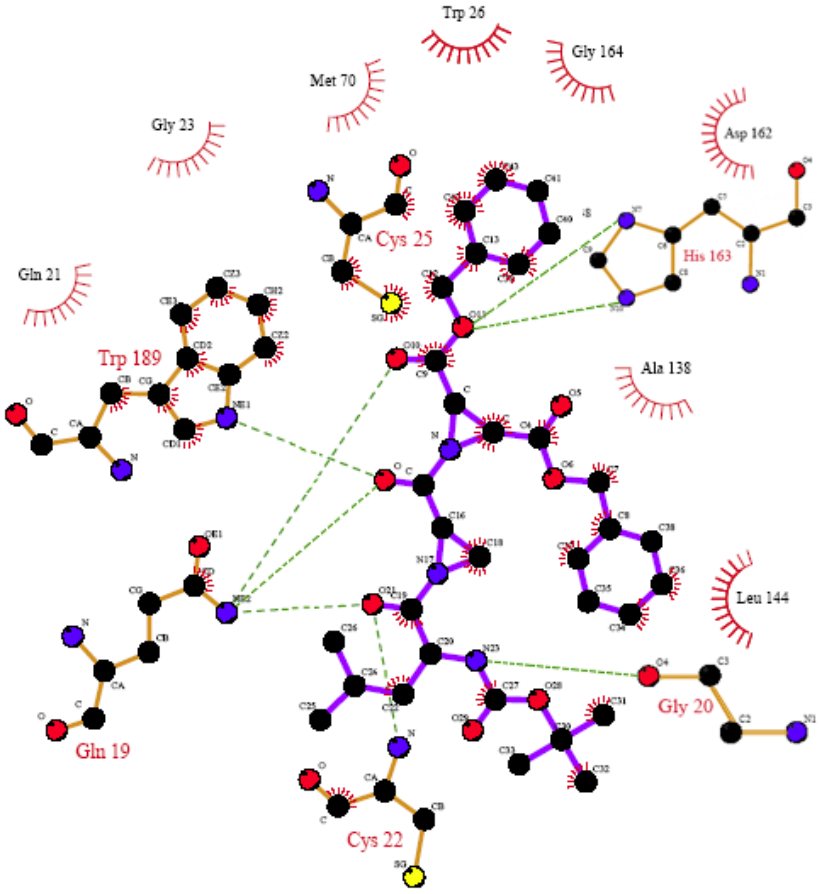


Abbildung 3.23 Im Docking gefundener Bindungsmodus von **18a$_S$** (N: sp^3) an CL. Die Bindung ist hauptsächlich stabilisiert durch die Wasserstoffbrückenbindungen der Aminosäuren der S2′-Tasche (Cys22, Gly20), dem sogenannten *Oxyanion hole* (Gln19) und der S1′-Tasche (Trp189, His163) sowie durch hydrophobe Interaktionen mit Met70 (S2), Trp189 und Ala138 (S1′).

(S2: Met70, Ala135, Ala214, Met161; S3: Leu69, Tyr72, Gly68, Gly61, Glu63). Die Benzylgruppe der Phe-Einheit bindet in die hydrophobe S2-Tasche; es werden Interaktionen zwischen den Carbonsäuregruppen des Inhibitors und dem Imidazoliumrest von His163 gefunden. Diese Bindung wird im Docking unabhängig vom Protonierungszustand der Säuregruppen identifiziert (die unter den Bedingungen des Assays deprotoniert sind).

Ähnliche Ergebnisse werden für die Bindung von **20a** an CB gefunden. Sowohl die Disäure als auch das Dianion binden in die S2-S3-Tasche (S2: Glu245,

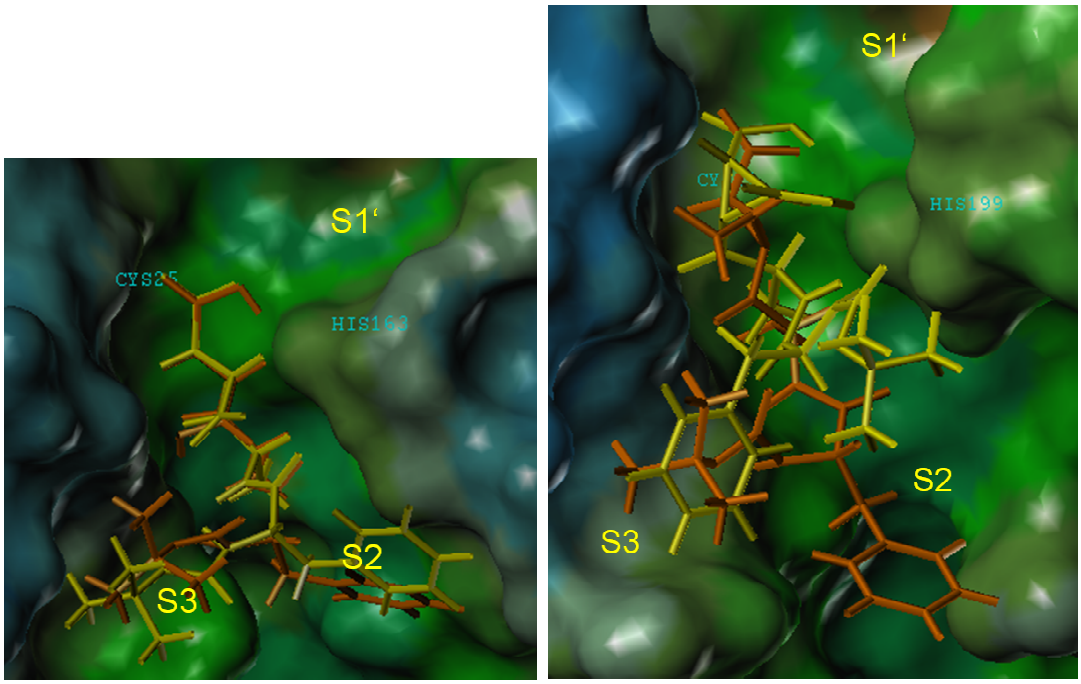


Abbildung 3.24 Überlagerung der Dockingergebnisse von **20a** an CL (links) und CB (rechts). Gelb: Dicarboxylat, orange: Disäure. Es sind nur die Bindungsmodi der Liganden mit sp^3-hybridisiertem Aziridinstickstoffatom gezeigt.

Gly198, Ala173, Ala200; S3: Tyr75, Asn72, Asp69, Pro76). Demnach werden für beide Enzyme ähnliche Bindungsmodi vorhergesagt, die jeweils nur die Non-primed subsites adressieren. Dies stellt eine gute Erklärung für die mangelnde Selektivität dieses Inhibitors dar.

Die Ergebnisse für die Leu-Pro-enthaltenden Verbindungen (**13b**, **13c**, **13e**) an CL (siehe Abb. 3.25) unterscheiden sich je nach verwendetem Hybridisierungszustand des Aziridin-Stickstoffatoms. In allen Fällen werden jedoch Bindungsmodi gefunden, in denen sowohl die Primed als auch die Non-primed sites vom Liganden adressiert werden; der Aziridinring liegt jeweils in der Nähe des aktiven Cys25-Schwefelatoms. Im Gegensatz zu **18a** und in Übereinstimmung mit **18e** passen diese Inhibitoren nicht optimal in die Substratbindetasche von CL, was gut mit der 30- bis 50fach schlechteren Inhibition verglichen mit **18a** korreliert.

Für die Azet- und Nip-enthaltenden Inhibitoren **11c**, **11d**, **12a**, **12b**, **16a**, **16b**, **16e**, **16f**, **17a** und **17b** sind die Resultate vergleichbar mit denen, die für

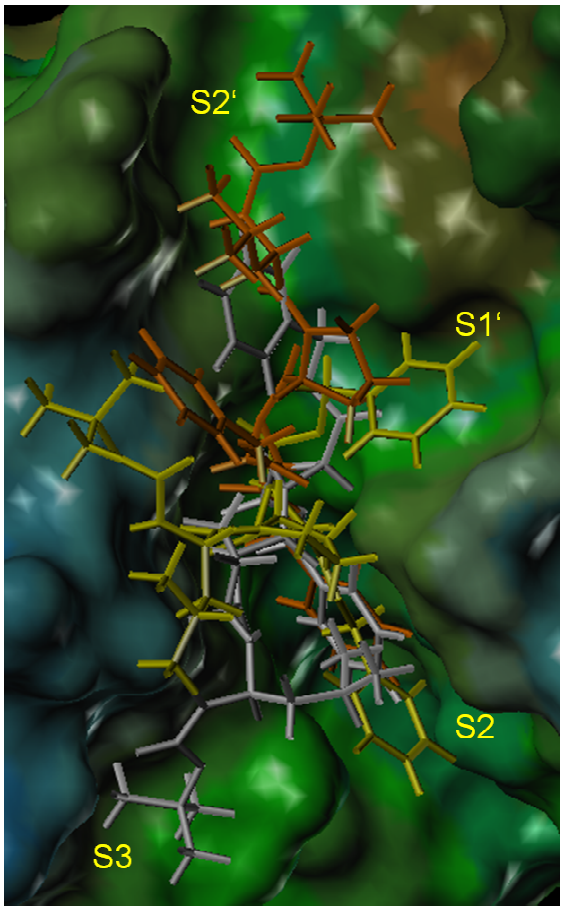


Abbildung 3.25 Überlagerung der Dockingergebnisse der Inhibitoren **13b** (orange), **13c** (gelb) und **13e** (weiß) an CL. Es sind nur die Bindungsmodi mit sp^3-hybridisiertem Aziridinstickstoffatom gezeigt.

13b, **13c** und **13e** erhalten wurden. In keinem der Fälle wird eine optimale Einpassung des Liganden vorhergesagt, was in guter Übereinstimmung mit der 300fach schlechteren Inhibition verglichen mit **18a** steht.

Zusammenfassend bleibt festzuhalten, daß sich Verbindungen mit der Sequenz Leu(Gly)-Caa über die gesamte Active site von CL erstrecken. Die Interaktionen, die zwischen diesen Inhibitoren und der S1'- und S2'-Tasche von CL gefunden werden, sind für CB aufgrund der positiv geladenen His110/111-Reste und der kleineren S1'/S2'-Taschen nicht möglich. Letzteres ist offensichtlich der Hauptgrund für die CL-Selektivität der meisten dieser

Verbindungen. Für den im Assay aktivsten Inhibitor **18a** ergibt sich auch im Docking eine perfekte Einpassung in die S2-, S1'- und S2'-Tasche von CL.

Im Gegensatz dazu zeigen die Dockingvorhersagen für **20a** an beiden Enzymen CB und CL sehr ähnliche Bindungsmodi, in denen die Primed subsites nicht adressiert werden. Die Hauptinteraktion besteht zwischen den Carbonsäuregruppen des Inhibitors und den positv geladenen Histidiniumgruppen der Active site (His163 in CL bzw. His199 in CB). Dies stimmt mit dem experimentellen Befund überein, daß **20a** nicht selektiv, sondern an beiden Enzymen in gleichem Maße aktiv ist.

Kapitel 4

Zusammenfassung

Nichts setzt dem Fortgang der Wissenschaft mehr Hindernis entgegen, als wenn man zu wissen glaubt, was man noch nicht weiß.
—Georg Christoph Lichtenberg, Aphoristiker und erster deutscher Professor für Experimentalphysik, † 1799

Diese Dissertation beschreibt Methoden zur Lösung wichtiger anwendungsorientierter Aspekte des struktur- und ligandbasierten *in-silico*-Wirkstoffdesigns. Dabei liegt der Fokus auf der Entwicklung chemometrischer Verfahren und der Überprüfung ihrer Leistungsfähigkeit. Die vorgeschlagenen Algorithmen werden mit entsprechenden etablierten Techniken verglichen. Die folgenden Abschnitte fassen die Vorgehensweisen und Resultate in den einzelnen Projektbereichen zusammen.

Identifizierung von Outliern Die Untersuchung eines QSAR-Datensatzes mit dem Ziel der Outlier-Identifizierung wird in der Praxis häufig vernachlässigt. Dabei ist es offensichtlich, daß kein QSAR-Modell auf jede nur denkbare chemische Verbindung anwendbar sein kann. Vielmehr handelt es sich um empirische mathematische Modelle, die nur innerhalb jenes Datenraums Gültigkeit besitzen, der von den Trainingsobjekten aufgespannt wird. Daher ist jedes Modell auf gewisse Grenzen beschränkt, außerhalb derer eine verläßliche Vorhersage unmöglich ist.

Die in dieser Arbeit entwickelte Methode ODD dient der Ermittlung dieser Grenzen und damit der Identifizierung von Outliern, also Objekten außerhalb des Anwendungsbereichs des Modells. Ziel der Entwicklung war ein

nur auf den unabhängigen Variablen (X-Daten) basierendes Verfahren, das auch auf hochdimensionale Datensätze anwendbar ist und weitestgehend auf den Eingriff des Benutzers (etwa die Definition von Grenzwerten) verzichtet. Ebenfalls wünschenswert war die Fähigkeit zur Identifikation von Inliern. Eine ausreichend hohe Geschwindigkeit sollte die Einsetzbarkeit im virtuellen Screening gewährleisten. Die Methode mußte der Überprüfung standhalten, den Vorhersagefehler eines Modells bei Vorhandensein extremer Outlier zu reduzieren, gleichzeitig aber unkritische Datensätze unbeeinflußt zu lassen.

ODD basiert auf der Beurteilung der euklidischen Distanz eines Testobjekts zu seinem am nächsten benachbarten Trainingsobjekt. Der Schwellenwert für die Betrachtung eines Objekts als Outlier wird dabei aus der Verteilung der Nächster-Nachbar-Distanzen der Trainingsobjekte berechnet. Durch dieses intrinsische Maß ergibt sich die gewünschte Dimensionsunabhängigkeit und vor allem die automatische Anpassung des Grenzwerts an die Charakteristik des Kalibrierdatensatzes ohne Eingriff des Benutzers. Die Validierung zeigt, daß ODD extreme Outlier zuverlässig erkennt und sich gleichzeitig durch eine im Vergleich zu anderen gebräuchlichen Verfahren geringere Anzahl falsch positiver Identifizierungen auszeichnet.

Ensemble-Techniken In einer vergleichenden Studie wurde die Leistungsfähigkeit verschiedener Ensemble-Techniken hinsichtlich ihres Einflusses auf den Vorhersagefehler untersucht. Dazu wurden umfangreiche Simulationen anhand mehrerer realer QSAR-Datensätze durchgeführt. Die Verwendung von Ensembles (d. h. einer Sammlung vieler Modelle, die mit geringfügig manipulierten Varianten des Trainingsdatensatzes kalibriert wurden) wirkt sich im allgemeinen positiv auf den Vorhersagefehler (RMSEP) aus. Diese Reduzierung des RMSEP wurde hier ermittelt und für verschiedenen Ansätze zur Ensemble-Generierung verglichen.

Insgesamt betrachtet erwiesen sich die Methoden der konvexen Pseudodaten und des Baggings als die effektivsten Verfahren zur Ensemble-Generierung, da sie den Vorhersagefehler am deutlichsten verbesserten. Die konvexen Pseudodaten wurden erstmalig zur Erzeugung von Ensembles in der QSAR-Analyse eingesetzt; sie werden als neuer Standard zur Reduzierung des RMSEP bei QSAR-Problemen vorgeschlagen, die Regressionsmodelle auf

Basis von latenten Variablen verwenden. Darüber hinaus bieten die Studien eine Abschätzung der mit Hilfe von Ensembles zu erzielenden Reduktion des Vorhersagefehlers bei typischen QSAR-Datensätzen.

Virtuelles Screening Beim virtuellen Screening handelt es sich um eine Technik zum Durchsuchen großer (virtueller) Molekülbibliotheken — oft mehrere Millionen Verbindungen — nach den aussichtsreichsten Wirkstoffkandidaten. Dies kann sowohl durch strukturbasierte als auch mit Hilfe ligandbasierter Verfahren geschehen.

Es wurden umfangreiche Simulationen anhand sechs verschiedener Targets und einer Bibliothek von mehr als 90 000 Molekülen durchgeführt, um das Potential strukturbasierter (Docking mit FLEXX) und ligandbasierter (Ähnlichkeitssuche mit mehreren Referenzen) Verfahren zu vergleichen. Darüber hinaus wurde durch Berechnung von Interaktionsfingerprints eine Möglichkeit geschaffen, die Information der beiden sonst getrennten Herangehensweisen zu kombinieren. Um den Einfluß des Klassifizierungsalgorithmus zu untersuchen, wurden verschiedene statistische Methoden zur Datenauswertung herangezogen. Als Bewertungskriterium für die Leistungsfähigkeit eines Verfahrens diente jeweils die Anzahl der wiedergefundenen aktiven Moleküle in der simulierten Screeningdatenbank.

Die Resultate führen zu dem Schluß, daß ligandbasierte Verfahren, die einfacher einzusetzen sind aber mehr *a-priori*-Information benötigen, dem strukturbasierten virtuellen Screening hinsichtlich der Datenbankanreicherung überlegen sind. Weiterhin konnte gezeigt werden, wie nutzbringend die Zusammenführung von strukturbasierter Information und solcher über das Interaktionsmuster bekanntermaßen aktiver Verbindungen für die Erhöhung der Wiederfindungsrate ist. Bei der Datenanalyse stellte sich heraus, daß im Mittel bestimmte statistische Methoden (minimale euklidische Distanz ED/Min bzw. Tanimoto-Ähnlichkeit der Integer-Fingerprints Int/Min) zu bevorzugen sind.

Kovalentes Docking von Cathepsin-Inhibitoren Die Cysteinproteasen Cathepsin B und L sind interessante pharmakologische Targets. Geeignete Inhibitoren stammen u. a. aus der Strukturklasse der Aziridine. Ein nukleophiler

Angriff des Cysteinrests des Enzyms auf den elektrophilen Aziridinring führt hier zur Ausbildung einer kovalenten Ligand-Rezeptor-Bindung.

Praktisch alle erhältlichen Dockingprogramme konzentrieren sich jedoch auf nicht-kovalente Ligand-Rezeptor-Interaktionen und lassen kein uneingeschränktes kovalentes Docking zu. Daher wurde für FLEXX ein Dockingprotokoll entworfen, das den entscheidenden nicht-kovalenten Zustand vor Ausbildung der kovalenten Bindung simulieren kann. Auf diese Weise konnte untersucht werden, ob sich die Reaktionszentren von Ligand und Enzym ausreichend nahe für die Ausbildung einer kovalenten Bindung kommen. Der vorgestellte Ansatz läßt sich leicht auf andere kovalente Ligand-Rezeptor-Systeme übertragen und bietet somit eine breite Anwendbarkeit.

Weiterhin wurde die Parametrisierung der in FLEXX vorgesehenen Interaktionsgeometrien an die strukturellen Eigenheiten der zu dockenden Aziridide angepaßt. Diese weisen nämlich formal eine Amidbindung auf, deren geometrische und elektronische Eigenschaften jedoch deutlich von den Werten eines typischen Amids abweichen.

Die Ergebnisse der Dockingstudien liefern wertvolle Einblicke für das Verständnis der Selektivität der untersuchten Liganden bezüglich Cathepsin B beziehungsweise L. Umgekehrt erbringt die gute Übereinstimmung der FLEXX-Resultate mit den experimentell bestimmten Inhibitionskonstanten den Nachweis für die Validität des verwendeten Dockingprotokolls.

Summary

This thesis describes methods for solving important application-oriented aspects of structure-based and ligand-based *in silico* drug design. The proposed algorithms are compared to well established techniques. The focus is particularly on the development and benchmarking of different chemometric techniques. In the following, the approaches and results within the different project areas are summarised.

Outlier Identification The inspection of QSAR datasets in order to identify prediction outliers is often omitted in practice. However, it is clear that no QSAR model is applicable to every conceivable chemical compound. Since QSAR models represent empirical mathematical models, these are only valid within the data space spanned by the training data. Hence, every model is restricted to certain borders beyond which a reliable prediction is impossible.

The method ODD developed in this work can be used to determine these borders and thus to identify outliers. Those are objects outside the data space spanned by the training data (i.e. the applicability domain of the model). The aim of the method is to detect outliers solely based on the predictor variables (X data). Moreover, the method must be capable to handle high-dimensional datasets with minimal user interference (e.g. setting of cut-offs). Furthermore, the ability to identify inliers would be preferable. The computational speed should be high enough to apply the method to virtual screening. The developed technique had to prove that it provides a reduction of the model's error of prediction if extreme outliers are present. At the same time, it should leave non-critical datasets unaffected.

ODD is based on the evaluation of the Euclidean distance of a test object towards its nearest neighbouring training object. The cut-off for deeming an object as outlier is calculated from the distribution of the nearest neighbour distances of the training set. This intrinsic value leads to the desired indepen-

dence from data dimensionality and, above all, to an automatic adjustment of the cut-off to the characteristics of the calibration dataset without any user intervention. The validation shows that ODD reliably identifies extreme outliers. On the other hand, it offers a low rate of false positives compared to other common techniques for outlier identification.

Ensemble Techniques In a benchmark study, the impact of different ensemble techniques on the prediction error was investigated. For this purpose, comprehensive simulations on several real QSAR datasets were carried out. The application of ensembles (i.e. a collection of many models trained with sligthly perturbed versions of the training set) usually lowers the error of prediction (RMSEP). The RMSEP reduction was determined and compared for different approaches of ensemble generation.

Overall, the methods of convex pseudo data and bagging proved to be the most efficient ways for ensemble generation (i.e. they resulted in the largest reduction of the prediction error). Convex pseudo data, which were applied to QSAR data sets for the first time as ensemble technique, are proposed as the new standard for lowering RMSEP in QSAR problems using latent variable regression models. Furthermore, the effect size of ensemble averaging was quantified for typical QSAR data sets.

Virtual Screening Virtual screening is a technique to screen large (virtual) molecular databases — often several million compounds — for the most promising drug candidates. This can be done by structure-based as well as by ligand-based approaches. Comprehensive computations on six different targets and a library of more than 90 000 compounds were carried out to compare the potential of structure-based techniques (docking with FLEXX) and ligand-based techniques (similarity searching with multiple queries). In addition to that, interaction fingerprints were computed in order to combine the information of the otherwise distinct approaches. Several statistical methods were applied for data analysis to investigate the impact of the machine learning algorithm. Figure of merit for each approach was the number of active compounds retrieved from the assembled screening database with known actives.

The results lead to the following conclusions: Ligand-based approaches, which are simpler to use but require more *a priori* information, turned out to be superior to structure-based virtual screening techniques in terms of database enrichment. In addition, it could be shown that combination of structure-based information with information of the interaction pattern of known actives is beneficial for increasing retrieval rates. Data analysis revealed that certain statistical methods (minimum Euclidean distance ED/Min, and Tanimoto similarity of integer fingerprints Int/Min, respectively) are on average to be preferred.

Covalent Docking of Cathepsin Inhibitors Cysteine proteases Cathepsin B and L are interesting pharmacological targets. Suitable inhibitors, amongst others, come from the structural class of aziridines. A nucleophilic attack of the enzyme's active site cysteine moiety on the electrophilic aziridine ring leads to formation of a covalent bond between ligand and receptor.

However, virtually all available docking programs concentrate on non-covalent ligand-receptor interactions and do not provide sophisticated, unrestricted covalent docking. Thus, a docking protocol for FLEXX was designed which is able to represent the essential non-covalent state before formation of the covalent bond. That way, it could be studied whether or not the reaction centres of both ligand and receptor adopt a position close enough to each other to actually form the covalent bond. The approach presented here can easily be transferred to other covalent ligand-receptor systems and therefore provides a broad applicability.

Furthermore, the parametrisation of the FLEXX interaction geometries was adapted to account for the special structural features of aziridides. Those show a formal amide bond, but its geometric and electronic properties differ noticeably from a typical amide.

The results of the docking studies provide valuable insights for understanding the Cathepsin B/L selectivity of the ligands under scrutiny. Vice versa, the good correspondence of the FLEXX results and the inhibition constants obtained experimentally provide evidence for the validity of the applied docking protocol.

Anhang

A.1 Datensätze

Die folgenden Datensätze wurden im Rahmen der Outlier-Identifizierung in Abschnitt 3.1 verwendet (Anzahl Moleküle $\times$ Anzahl Deskriptoren):

logP (346$\times$101) Oktanol-Wasser-Verteilungskoeffizient, Deskriptor: SE-Vektoren.[170,171]

Sol (257$\times$44) Wasserlöslichkeit unterschiedlicher Verbindungen, ausgewählte MOE-Deskriptoren (4SC AG, interne Daten, nicht veröffentlicht).

DHODH (206$\times$183) IC_{50}-Werte, ausgewählte MOE-Deskriptoren (DHODH-Projekt der 4SC AG, interne Daten, nicht veröffentlicht).

Bei der Evaluierung der Ensemble-Techniken in Abschnitt 3.2 fanden zusätzlich folgende Datensätze Verwendung:

HEPT (79$\times$181) Inhibitoren der HIV1 Reversen Transkriptase, Derivate des nichtnukleosidischen Inhibitors 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymin (HEPT), MOE-Deskriptoren.[172]

ACE (114$\times$56) ACE-Inhibitoren, Datensatz von J. SUTHERLAND[173,174]

DHFR (361$\times$70) DHFR-Inhibitoren, Datensatz von J. SUTHERLAND[173,174]

Die simulierten virtuellen Screenings in Abschnitt 3.4 basieren auf folgenden Datensätzen:

ACE (355×181) Eigene Zusammenstellung bekannter ACE-Inhibitoren aus dem MDDR; MOE-Deskriptoren.

AChE (701×181) Eigene Zusammenstellung bekannter AChE-Inhibitoren aus dem MDDR; MOE-Deskriptoren.

COX (636×181) COX-Inhibitoren aus dem MDDR; MOE-Deskriptoren.[107,133]

HIV (747×181) Inhibitoren der HIV1 Reversen Transkriptase aus dem MDDR; MOE-Deskriptoren.[107,133]

Renin (1121×181) Renin-Inhibitoren, MDDR; MOE-Deskriptoren.[107,133]

Thrombin (803×181) Thrombin-Inhibitoren aus dem MDDR; MOE-Deskriptoren.[107,133]

InA (93451×181) Verbindungen aus dem MDDR ohne bekannte Aktivität an COX, HIV, Renin oder Thrombin; MOE-Deskriptoren.[107,133]

InA-ACE (93154×181) Verbindungen aus dem MDDR ohne bekannte Aktivität an ACE, COX, HIV, Renin, Thrombin; MOE-Deskriptoren; eigene Zusammenstellung basierend auf InA.

InA-AChE (92843×181) Verbindungen aus dem MDDR ohne bekannte Aktivität an AChE, COX, HIV, Renin, Thrombin; MOE-Deskriptoren; eigene Zusammenstellung basierend auf InA.

A.2 Deskriptoren

Die folgende Liste enthält die im Programm MOE[175] verfügbaren Deskriptoren. Die zehn mit „#“ auskommentierten Zeilen (x3D-Klasse, rotations- und translationsvariant) wurden nicht verwendet, so daß sich ein Standardsatz von 181 Deskriptoren ergibt.

```
apol            2D   Sum of atomic polarizabilities
ASA             i3D  Water accessible surface area
ASA+            i3D  Positive accessible surface area
ASA-            i3D  Negative accessible surface area
ASA_H           i3D  Total hydrophobic surface area
ASA_P           i3D  Total polar surface area
a_acc           2D   Number of H-bond acceptor atoms
a_acid          2D   Number of acidic atoms
a_aro           2D   Number of aromatic atoms
a_base          2D   Number of basic atoms
a_count         2D   Number of atoms
a_don           2D   Number of H-bond donor atoms
a_heavy         2D   Number of heavy atoms
a_hyd           2D   Number of hydrophobic atoms
a_IC            2D   Atom information content (total)
a_ICM           2D   Atom information content (mean)
a_nB            2D   Number of boron atoms
a_nBr           2D   Number of bromine atoms
a_nC            2D   Number of carbon atoms
a_nCl           2D   Number of chlorine atoms
a_nF            2D   Number of fluorine atoms
a_nH            2D   Numberof hydrogen atoms
a_nI            2D   Number of iodine atoms
a_nN            2D   Number of nitrogen atoms
a_nO            2D   Number of oxygen atoms
a_nP            2D   Number of phosphorus atoms
a_nS            2D   Number of sulfur atoms
balabanJ        2D   Balaban averaged distance sum connectivity
bpol            2D   Difference of bonded atom polarizabilities
b_1rotN         2D   Number of rotatable single bonds
b_1rotR         2D   Fraction of rotatable single bonds
b_ar            2D   Number of aromatic bonds
b_count         2D   Number of bonds
b_double        2D   Number of double bonds
b_heavy         2D   Number of heavy-heavy bonds
b_rotN          2D   Number of rotatable bonds
b_rotR          2D   Fraction of rotatable bonds
b_single        2D   Number of single bonds
b_triple        2D   Number of triple bonds
CASA+           i3D  Charge-weighted positive surface area
CASA-           i3D  Charge-weighted negative surface area
chi0            2D   Atomic connectivity index (order 0)
chi0v           2D   Atomic valence connectivity index (order 0)
chi0v_C         2D   Carbon valence connectivity index (order 0)
chi0_C          2D   Carbonconnectivity index (order 0)
chi1            2D   Atomic connectivity index (order 1)
chi1v           2D   Atomic valence connectivity index (order 1)
chi1v_C         2D   Carbon valence connectivity index (order 1)
chi1_C          2D   Carbon connectivity index (order 1)
DASA            i3D  Absolute difference in surface area
DCASA           i3D  Absolutedifference in charge-weighted areas
dens            i3D  Mass density (AMU/A^3)
density         2D   Mass density (AMU/A**3)
diameter        2D   Largest vertex eccentricity in graph
dipole          i3D  Dipole moment
# dipoleX      x3D  Dipole moment (X)
# dipoleY      x3D  Dipole moment (Y)
```

# dipoleZ	x3D	Dipole moment (Z)
E	i3D	Potential Energy
E_ang	i3D	Angle Bend Energy
E_ele	i3D	Electrostatic energy
E_nb	i3D	Non-bonded energy
E_oop	i3D	Out-of-plane Energy
# E_rele	x3D	Electrostatic Interaction Energy
# E_rnb	x3D	Non-bonded Interaction Energy
# E_rsol	x3D	Solvation Correction Difference
# E_rvdw	x3D	Van der Waals Interaction Energy
E_sol	i3D	Solvation energy
E_stb	i3D	Stretch-bend energy
E_str	i3D	Bond stretch energy
E_tor	i3D	Torsion energy
E_vdw	i3D	Van der Waals energy
FASA+	i3D	Fractional positive accessible surface area
FASA-	i3D	Fractional negative accessible surface area
FASA_H	i3D	Fractional hydrophobic surface area
FASA_P	i3D	Fractional polar surface area
FCASA+	i3D	Fractional charge-weighted positive surface area
FCASA-	i3D	Fractional charge-weighted negative surface area
FCharge	2D	Sum of formal charges
glob	i3D	Molecular globularity
Kier1	2D	First kappa shape index
Kier2	2D	Second kappa shape index
Kier3	2D	Third kappa shape index
KierA1	2D	First alphamodified shape index
KierA2	2D	Second alpha modified shape index
KierA3	2D	Third alpha modified shape index
KierFlex	2D	Molecular flexibility
logP(o/w)	2D	Log octanol/water partition coefficient
mr	2D	Molar refractivity
PC+	2D	Total positive partial charge
PC-	2D	Total negative partial charge
PEOE_PC+	2D	Total positive partial charge
PEOE_PC-	2D	Total negative partial charge
PEOE_RPC+	2D	Relative positive partial charge
PEOE_RPC-	2D	Relative negative partial charge
PEOE_VSA+0	2D	Total positive 0 vdw surface area
PEOE_VSA+1	2D	Total positive 1 vdw surface area
PEOE_VSA+2	2D	Total positive 2 vdw surface area
PEOE_VSA+3	2D	Total positive 3 vdw surface area
PEOE_VSA+4	2D	Total positive 4 vdw surface area
PEOE_VSA+5	2D	Total positive 5 vdw surface area
PEOE_VSA+6	2D	Total positive 6 vdw surface area
PEOE_VSA-0	2D	Total negative 0 vdw surface area
PEOE_VSA-1	2D	Total negative 1 vdw surface area
PEOE_VSA-2	2D	Total negative 2 vdw surface area
PEOE_VSA-3	2D	Total negative 3 vdw surface area
PEOE_VSA-4	2D	Total negative 4 vdw surface area
PEOE_VSA-5	2D	Total negative 5 vdw surface area
PEOE_VSA-6	2D	Total negative 6 vdw surface area
PEOE_VSA_FHYD	2D	Fractional hydrophobic vdw surface area
PEOE_VSA_FNEG	2D	Fractional negative vdw surface area
PEOE_VSA_FPNEG	2D	Fractional polar negative vdw surface area
PEOE_VSA_FPOL	2D	Fractional polar vdw surface area
PEOE_VSA_FPOS	2D	Fractional positive vdw surface area
PEOE_VSA_FPPOS	2D	Fractional polar positive vdw surface area
PEOE_VSA_HYD	2D	Total hydrophobic vdw surface area
PEOE_VSA_NEG	2D	Total negative vdw surface area
PEOE_VSA_PNEG	2D	Total polar negative vdw surface area
PEOE_VSA_POL	2D	Total polar vdw surface area
PEOE_VSA_POS	2D	Total positive vdw surface area
PEOE_VSA_PPOS	2D	Total polar positive vdw surface area
petitjean	2D	(diameter - radius) / diameter
petitjeanSC	2D	(diameter - radius) / radius
pmi	i3D	Principal moment of inertia
# pmiX	x3D	Principal moment of inertia (X)
# pmiY	x3D	Principal moment of inertia (Y)

# pmiZ	x3D	Principal moment of inertia (Z)
Q_PC+	2D	Total positive partial charge
Q_PC-	2D	Total negative partial charge
Q_RPC+	2D	Relative positive partial charge
Q_RPC-	2D	Relative negative partial charge
Q_VSA_FHYD	2D	Fractional hydrophobic vdw surface area
Q_VSA_FNEG	2D	Fractional negative vdw surface area
Q_VSA_FPNEG	2D	Fractional polar negative vdw surface area
Q_VSA_FPOL	2D	Fractional polar vdw surface area
Q_VSA_FPOS	2D	Fractional positive vdw surface area
Q_VSA_FPPOS	2D	Fractional polar positive vdw surface area
Q_VSA_HYD	2D	Total hydrophobic vdw surface area
Q_VSA_NEG	2D	Total negative vdw surface area
Q_VSA_PNEG	2D	Total polar negative vdw surface area
Q_VSA_POL	2D	Total polar vdw surface area
Q_VSA_POS	2D	Total positive vdw surface area
Q_VSA_PPOS	2D	Total polar positive vdw surface area
radius	2D	Smallest vertex eccentricity in graph
reactive	2D	Molecule contains reactive groups
rgyr	i3D	Radius of gyration
RPC+	2D	Relative positive partial charge
RPC-	2D	Relative negative partial charge
SlogP	2D	Log Octanol/Water Partition Coefficient
SlogP_VSA0	2D	Bin 0 SlogP (-10 ,-0.40]
SlogP_VSA1	2D	Bin 1 SlogP (-0.40,-0.20]
SlogP_VSA2	2D	Bin 2 SlogP (-0.20, 0.00]
SlogP_VSA3	2D	Bin 3 SlogP (0.00, 0.10]
SlogP_VSA4	2D	Bin 4 SlogP (0.10, 0.15]
SlogP_VSA5	2D	Bin 5 SlogP (0.15, 0.20]
SlogP_VSA6	2D	Bin 6 SlogP (0.20, 0.25]
SlogP_VSA7	2D	Bin 7 SlogP (0.25, 0.30]
SlogP_VSA8	2D	Bin 8 SlogP (0.30, 0.40]
SlogP_VSA9	2D	Bin 9 SlogP (0.40,10]
SMR	2D	Molar Refractivity
SMR_VSA0	2D	Bin 0 SMR (0.000,0.110]
SMR_VSA1	2D	Bin 1 SMR (0.110,0.260]
SMR_VSA2	2D	Bin 2 SMR(0.260,0.350]
SMR_VSA3	2D	Bin 3 SMR (0.350,0.390]
SMR_VSA4	2D	Bin 4 SMR (0.390,0.440]
SMR_VSA5	2D	Bin 5 SMR (0.440,0.485]
SMR_VSA6	2D	Bin 6 SMR (0.485,0.560]
SMR_VSA7	2D	Bin 7 SMR (0.560,10]
std_dim1	i3D	Standard dimension 1
std_dim2	i3D	Standard dimension 2
std_dim3	i3D	Standard dimension 3
TPSA	2D	Topological Polar Surface Area (A**2)
VAdjEq	2D	Vertex adjacency information (equal)
VAdjMa	2D	Vertex adjacency information (mag)
VDistEq	2D	Vertex distance equality index
VDistMa	2D	Vertex distance magnitude index
vdw_area	2D	Van der Waals surface area (A**2)
vdw_vol	2D	Van der Waals volume (A**3)
vol	i3D	Van der Waals volume
VSA	i3D	Van derWaals surface area
vsa_acc	2D	VDW acceptor surface area (A**2)
vsa_acid	2D	VDW acidic surface area (A**2)
vsa_base	2D	VDW basic surface area (A**2)
vsa_don	2D	VDW donor surface area (A**2)
vsa_hyd	2D	VDW hydrophobe surface area (A**2)
vsa_other	2D	VDW other surface area (A**2)
vsa_pol	2D	VDW polar surface area (A**2)
Weight	2D	Molecular weight (CRC)
weinerPath	2D	Weiner path number
weinerPol	2D	Weiner polarity number
zagreb	2D	Zagreb index

Literaturverzeichnis

[1] S. Kraljevic, P. J. Stambrook, K. Pavelic, *EMBO Reports* **2004**, *5*, 837–842.

[2] C. M. Dobson, *Nature* **2004**, *432*, 824–828.

[3] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50.

[4] J. Bajorath, *Nat. Rev. Drug Discovery* **2002**, *2*, 882–894.

[5] P. Kuhn, K. Wilson, M. G. Patch, R. C. Stevens, *Curr. Opin. Chem. Biol.* **2002**, *6*, 704–710.

[6] M. Itzstein, W.-Y. Wu, G. B. Kok, M. S. Pegg, J. C. Dyason, B. Jin, T. V. Phan, M. L. Smythe, H. F. White, S. W. Oliver, P. M. Colman, J. N. Varghese, D. M. Ryan, J. M. Woods, R. C. Bethell, V. J. Hotham, J. M. Cameron, C. R. Penn, *Nature* **1993**, *363*, 418–423.

[7] C. U. Kim, W. Lew, M. A. Williams, H. Liu, L. Zhang, S. Swaminathan, N. Bischofberger, M. S. Chen, D. B. Mendel, C. Y. Tai, W. G. Laver, R. C. Stevens, *J. Am. Chem. Soc.* **1997**, *119*, 681–690.

[8] F. L. Stahura, J. Bajorath, *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.

[9] H. Kubinyi, *Nat. Rev. Drug Discovery* **2003**, *2*, 665–668.

[10] D. A. Smith, *Drug Discov. Today* **2002**, *7*, 1080–1081.

[11] H. van de Waterbeemd, E. Gifford, *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.

[12] A. Crum-Brown, T. R. Fraser, in *Proceedings of the Royal Society of Edinburgh* **1868**, S. 151.

[13] C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

[14] S. M. J. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395–399.

[15] H. van de Waterbeemd, *Quant. Struct. Act. Relat.* **1992**, *11*, 200–204.

[16] J. J. Sutherland, L. A. O'Brien, D. F. Weaver, *J. Med. Chem.* **2004**, *47*, 5541–5554.

[17] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Bd. 11 von *Methods and Principles in Medicinal Chemistry*, Wiley-VCH, Weinheim **2000**.

[18] R. D. Cramer, M. Milne, in *Abstracts of the ACS Meeting*, Honolulu **1979**, S. 44.

[19] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

[20] G. Klebe, *Persp. Drug Discov. Design* **1998**, *12*, 87–104.

[21] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.

[22] R. C. Wade, in *3D QSAR in drug design: Theory, methods and applications* (Hrsg. H. Kubinyi), ESCOM, Leiden **1993**, S. 486–505.

[23] *http://www.daylight.com/dayhtml/doc/theory/theory.finger.html*, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, USA.

[24] H. Kubinyi, *Quant. Struct. Act. Relat.* **2002**, *21*, 348–356.

[25] R. Henrion, G. Henrion, *Multivariate Datenanalyse*, Springer-Verlag, Heidelberg **1994**.

[26]R. Kramer, *Chemometric techniques for quantitative analysis*, Marcel Dekker AG, Basel **1998**.

[27]J. Mandel, *Amer. Statist.* **1982**, *36*, 15–24.

[28]I. T. Jolliffe, *Principal Components Analysis*, Springer, New York **1986**.

[29]S. Wold, H. Martens, H. Wold, in *Matrix Pencils* (Hrsg. A. Ruhe, B. Kagström), Springer, Heidelberg **1983**, S. 286–293.

[30]K. Baumann, *Trends in Analytical Chemistry* **2003**, *22*, 395–406.

[31]S. Wold, *Technometrics* **1978**, *20*, 397–405.

[32]S. Geisser, *J. Amer. Statist. Assoc.* **1975**, *70*, 320–328.

[33]J. Shao, *J. Amer. Statist. Assoc.* **1993**, *88*, 486–494.

[34]D. Jouan-Rimbaud, E. Bouveresse, D. L. Massart, O. E. d. Noord, *Anal. Chim. Acta* **1999**, *388*, 283–301.

[35]D. A. Belsley, E. Kuh, R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York **1980**.

[36]S. Chatterjee, A. S. Hadi, *Statistical Science* **1986**, *1*, 379–416.

[37]W. J. Egan, S. L. Morgan, *Anal. Chem.* **1998**, *70*, 2372–2379.

[38]B. Walczak, D. L. Massart, *Chemom. Intell. Lab. Syst.* **1998**, *41*, 1–15.

[39]M. Meloun, J. Militky, *Anal. Chim. Acta* **2001**, *439*, 169–191.

[40]J. A. F. Pierna, F. Wahl, O. E. d. Noord, D. L. Massart, *Chemom. Intell. Lab. Syst.* **2002**, *63*, 27–39.

[41]A. Tropsha, P. Gramatica, K. G. Vijay, *QSAR Comb. Sci.* **2003**, *22*, 69–77.

[42]S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, O. Mekenyan, *J. Chem. Inf. Model.* **2005**, *45*, 839–849.

[43]J. A. F. Pierna, L. Jin, M. Daszykowski, F. Wahl, D. L. Massart, *Chemom. Intell. Lab. Syst.* **2003**, *68*, 17–28.

[44]J. Mandel, *J. Res. Nat. Bur. Stand.* **1985**, *90*, 465–476.

[45]E. Fix, J. L. Hodges, *Discriminatory analysis: Nonparametric discrimination: Consistency properties*, Techn. Bericht, USAF School of Aviation Medicine, Randolph Fields TX, USA **1951**.

[46]M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombar, A. Tropsha, *J. Med. Chem.* **2003**, *46*, 3013–3020.

[47]T. G. Dietterich, *Ensemble Methods in Machine Learning*, Techn. Bericht, Oregon State University, Corvallis, Corvallis, Oregon **2000**.

[48]L. Breiman, *Machine Learning* **2000**, *40*, 229–242.

[49]L. Breiman, *Using Convex Pseudo-Data to Increase Prediction Accuracy*, Techn. Bericht, Statistics Department, University of California Berkeley, Berkeley, CA **1998**.

[50]L. Breiman, *Machine Learning* **1996**, *24*, 123–140.

[51]J. G. Topliss, R. J. Costello, *J. Med. Chem.* **1972**, *15*, 1066–1068.

[52]K. Baumann, *QSAR Comb. Sci.* **2005**, *24*, 1033–1046.

[53]R. A. Fisher, *Annals of Eugenics* **1936**, *7*, 179–188.

[54]M. Otto, *Chemometrie: Statistik und Computereinsatz in der Analytik*, VCH, Weinheim **1997**.

[55]P. Willett, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[56]R. De Maesschalck, D. Jouan-Rimbaud, D. L. Massart, *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18.

[57]M. Köppen, in *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, via Internet **2000**.

[58]M. Verleysen, D. Francois, in *8th International Workshop on Artificial Neural Networks (IWANN)*, Barcelona **2005**, S. 758–770.

[59]P. N. Yianilos, in *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, San Francisco **2000**, S. 361–370.

[60]E. Fischer, *Ber. Dtsch. Chem. Ges.* **1894**, *2*, 2985–2993.

[61]F. W. Lichtenthaler, *Angew. Chem.* **1994**, *106*, 2456–2467.

[62]H. Kubinyi, *Pharmazie in unserer Zeit* **1994**, *23*, 281–290.

[63]H.-J. Böhm, G. Klebe, H. Kubinyi, *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg **1996**.

[64]I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269–288.

[65]M. McGann, H. Almond, A. Nicholls, J. A. Grant, F. Brown, *Biopolymers* **2003**, *68*, 76–90.

[66]D. S. Goodsell, G. M. Morris, A. J. Olson, *J. Mol. Recognit.* **1996**, *9*, 1–5.

[67]G. Jones, P. Willet, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.

[68]B. Kramer, G. Metz, M. Rarey, T. Lengauer, *Med. Chem. Res.* **1999**, *9*, 463–478.

[69]B. Kramer, M. Rarey, T. Lengauer, *Proteins* **1999**, *37*, 228–241.

[70]D. Joseph-McCarthy, B. E. Thomas IV, M. Belmarsh, D. Moustakas, J. C. Alvarez, *Proteins* **2003**, *51*, 172–188.

[71]A. N. Jain, *J. Med. Chem.* **2003**, *46*, 499–511.

[72]H.-J. Böhm, *J. Comp.-Aided Mol. Des.* **1994**, *8*, 243–256.

[73]R. D. Taylor, P. J. Jewsbury, J. W. Essex, *J. Comp.-Aided Mol. Des.* **2002**, *16*, 151–116.

[74]Z. Deng, C. Chuaqui, J. Singh, *J. Med. Chem.* **2004**, *47*, 337–344.

[75]C. Chuaqui, Z. Deng, J. Singh, *J. Med. Chem.* **2005**, *48*, 121–133.

[76]T. Schirmeister, U. Käppler, *Mini Rev. Med. Chem.* **2003**, *3*, 361–373.

[77]V. Martichonok, C. Plouffe, A. C. Storer, R. Menard, J. B. Jones, *J. Med. Chem.* **1995**, *38*, 3078–3085.

[78]T. Schirmeister, M. Peric, *Bioorg. Med. Chem.* **2000**, *8*, 1281–1291.

[79]R. Vicik, M. Busemann, K. Baumann, T. Schirmeister, *Curr. Top. Med. Chem.* **2006**, *6*, 331–353.

[80]H. Helten, T. Schirmeister, B. Engels, *J. Phys. Chem. A* **2004**, *108*, 7691–7701.

[81]H. Helten, T. Schirmeister, B. Engels, *J. Org. Chem.* **2005**, *70*, 233–237.

[82]*http://merops.sanger.ac.uk*.

[83]Z. Werb, in *Textbook of Rheumatology* (Hrsg. W. N. Keller, E. D. Harris, S. Ruddy, C. S. Sledge), W. B. Saunder Co., Philadelphia PA, USA **1989**, S. 300–321.

[84]N. Katunuma, E. Kominami, *Rev. Physiol. Biochem. Pharmacol.* **1987**, *108*, 1–20.

[85]B. F. Sloane, K. Moin, E. Krepela, J. Rhozhin, *Cancer Metastasis Rev.* **1990**, *9*, 333–352.

[86]H.-J. Böhm, G. Schneider, *Virtual Screening for Bioactive Molecules*, Bd. 10 von *Methods and Principles in Medicinal Chemistry*, Wiley-VCH, Weinheim **2000**.

[87]K. H. Bleicher, H.-J. Böhm, K. Müller, A. I. Alanine, *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.

[88]G. Klebe, in *Perspectives in Drug Discovery and Design* (Hrsg. G. Klebe), Springer Netherlands, Bd. 20 **2000**, S. 7–11.

[89]J. Altshuler, A. Flanagan, P. Guy, M. Steiner, P. Tollman, *A Revolution in R&D: How Genomics and Genetics are Transforming the Biopharmaceutical Industry*, Techn. Bericht, The Boston Consulting Group **2001**.

[90]M. H. J. Seifert, K. Wolf, D. Vitt, *Biosilico* **2003**, *1*, 143–149.

[91]B. K. Shoichet, *Nature* **2004**, *432*, 862–865.

[92]P. D. Lyne, *Drug Discov. Today* **2002**, *7*, 1047–1055.

[93]J. Hert, P. Willet, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

[94]A. Bender, R. C. Glen, *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

[95]J. C. Alvarez, *Curr. Opin. Chem. Biol.* **2004**, *8*, 365–370.

[96]M. Stahl, M. Rarey, *J. Med. Chem.* **2001**, *44*, 1035–1042.

[97]I. Halperin, B. Ma, H. Wolfons, R. Nussinov, *Proteins* **2002**, *47*, 409–443.

[98]P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, C. L. Brooks, *J. Med. Chem.* **2004**, *47*, 3032–3047.

[99]R. Wang, Y. Lu, S. Wang, *J. Med. Chem.* **2003**, *46*, 2287–2303.

[100]P. S. Charifson, J. J. Corkery, M. A. Murcko, W. P. Walters, *J. Med. Chem.* **1999**, *42*, 5100–5109.

[101]C. Bissantz, G. Folkers, D. Rognan, *J. Med. Chem.* **2000**, *43*, 4759–4767.

[102]T. Klabunde, G. Hessler, *ChemBioChem* **2002**, *3*, 928–944.

[103]G. Rum, W. C. Herndon, *J. Am. Chem. Soc.* **1991**, *113*, 9055–9060.

[104]Y. C. Martin, J. L. Kofron, L. M. Traphagen, *J. Med. Chem.* **2002**, *45*, 4350–4358.

[105]C. Lemmen, T. Lengauer, *J. Comp.-Aided Mol. Des.* **2000**, *14*, 215–232.

[106]J. Mestres, R. M. A. Knegtel, *Persp. Drug Discov. Design* **2000**, *20*, 191–207.

[107]J. Hert, P. Willett, D. J. Wilton, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

[108]*http://www.daylight.com/smiles/*, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, USA.

[109]*http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html*, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, USA.

[110]P. Ehrlich, *Chem. Ber.* **1909**, *42*, 17–47.

[111]P. Gund, *Prog. Mol. Subcell. Biol.* **1977**, *11*.

[112]T. Langer, E. M. Krovat, *Curr. Opin. Drug Discov. Devel.* **2003**, *6*, 370–376.

[113]R. P. Sheridan, B. P. Feuston, V. N. Maiorov, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

[114]Y. j. Xu, H. Gao, *QSAR Comb. Sci.* **2003**, *22*, 422–429.

[115]L. He, P. C. Jurs, *J. Mol. Graph. Model.* **2005**, *23*, 503–523.

[116]R. R. Picard, K. N. Berk, *Amer. Statist.* **1990**, *44*, 140–147.

[117]MATLAB, *Release 13, Version 6.5.0.196271*, The MathWorks Inc., Natick, MA, USA **2003**.

[118]J. S. Rao, R. Tibshirani, *The out-of-bootstrap method for model averaging and selection*, Techn. Bericht, University of Toronto, Statistics Department, Toronto **1996**.

[119]M. LeBlanc, R. Tibshirani, *J. Am. Stat. Assoc.* **1996**, *91*, 1641–1662.

[120]J. K. Lanctot, S. Putta, C. Lemmen, J. Greene, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2163–2169.

[121]B. E. Mattioni, G. W. Kauffman, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 949–963.

[122]R. Guha, P. Jurs, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179–2189.

[123]C. Merkwirth, H. Mauser, T. Schulz-Gasch, O. Roche, M. Stahl, T. Lengauer, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971–1978.

[124]V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, Q. Song, *J. Chem. Inf. Model.* **2005**, *45*, 786–799.

[125]U. Fechner, G. Schneider, *ChemBioChem* **2004**, *5*, 538–540.

[126]A. Bender, R. C. Glen, *Org. Biomol. Chem.* **2004**, 3204–3218.

[127]D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, L. E. Weinberger, *J. Med. Chem.* **1996**, *39*, 3049–3059.

[128]G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed. Engl.* **1999**, *38*, 2894–2896.

[129]K. M. Andrews, R. D. Cramer, *J. Med. Chem.* **2000**, *43*, 1723–1740.

[130]T. Lengauer, C. Lemmen, M. Rarey, M. Zimmermann, *Drug Discov. Today* **2004**, *9*, 27–34.

[131]J. W. Eaton, *GNU Octave Manual, Octave 2.1.40*, Network Theory Limited, Bristol, UK **2002**.

[132]*MDL Drug Data Report (MDDR)*, http://www.mdl.com/products/knowledge/drug_data_report.

[133]*http://www.cheminformatics.org/datasets/hert/index.shtml*.

[134]J. Gasteiger, C. Rudolph, J. Sadowski, *Tetrahedron Comp. Method.* **1990**, *3*, 537–547.

[135]F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. Mol. Biol.* **1977**, *112*, 535–542.

[136]M. D. Cummings, R. L. DesJarlais, A. C. Gibbs, V. Mohan, E. P. Jaeger, *J. Med. Chem.* **2005**, *48*, 962–976.

[137]M. Kontoyianni, L. M. McClellan, G. S. Sokol, *J. Med. Chem.* **2004**, *47*, 558–565.

[138]H. Chen, P. D. Lyne, F. Giordanetto, T. Lovell, J. Li, *J. Chem. Inf. Model.* **2006**, *46*, 401–415.

[139]R. Natesh, S. Schwager, E. Sturrock, K. Acharya, *Nature* **2003**, *421*, 551.

[140]G. Kryger, I. Silman, J. L. Sussman, *Structure Fold Des.* **1999**, *7*, 297.

[141]R. Bone, J. P. Vacca, P. S. Anderson, M. K. Holloway, *J. Am. Chem. Soc.* **1991**, *113*, 9382.

[142]R. G. Kurumbail, A. M. Stevens, J. K. Gierse, J. J. McDonald, R. A. Stegeman, J. Y. Pak, D. Gildehaus, J. M. Miyashiro, T. D. Penning, K. Seibert, P. C. Isakson, W. C. Stallings, *Nature* **1996**, *384*, 644.

[143]D. W. Banner, P. Hadvary, *J. Biol. Chem.* **1991**, *266*, 20085–20093.

[144]L. Tong, S. Pav, D. Lamarre, L. Pilote, S. LaPlante, P. C. Anderson, G. Jung, *J. Mol. Biol.* **1995**, *250*, 211.

[145]MATLAB, *Release 14, Version 7.0.4.365 SP2*, The MathWorks Inc., Natick, MA, USA **2005**.

[146]*World Drug Index (WDI)*, Derwent Information.

[147]T. Schulz-Gasch, M. Stahl, *J. Mol. Mod.* **2003**, *9*, 47–57.

[148]E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, *Proteins* **2004**, *57*, 225–242.

[149]D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

[150]C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.

[151]V. N. Maiorov, R. P. Sheridan, *J. Chem. Inf. Model.* **2005**, *45*, 1017–1023.

[152]J.-M. Yang, Y.-F. Chen, T.-W. Shen, B. S. Kristal, D. F. Hsu, *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.

[153]M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor, P. Watson, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.

[154]A. C. Good, M. A. Hermsmeier, S. A. Hindle, *J. Comp.-Aided Mol. Des.* **2004**, *18*, 529–536.

[155]A. Evers, G. Hessler, H. Matter, T. Klabunde, *J. Med. Chem.* **2005**, *48*, 5448–5465.

[156]S. A. Hindle, M. Rarey, C. Buning, T. Lengauer, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 129–149.

[157]R. Vicik, *Synthese und Eigenschaften N-Acylierter Aziridin-2,3-dicarboxylate als selektive, peptidomimetische Inhibitoren von Cystein-Proteasen der Cathepsin-L-Subfamilie*, Dissertation, Universität Würzburg **2004**.

[158]H. Gohlke, M. Hendlich, G. Klebe, *J. Mol. Biol.* **2000**, *295*, 337–356.

[159]S. A. Gillmor, C. S. Craik, R. J. Fletterick, *Protein Sci.* **1997**, *6*, 1603–1611.

[160]B. Zhao, C. A. Janson, B. Y. Amegadzie, K. D'Alessio, C. Griffin, C. R. Hanning, C. Jones, J. Kurdyla, M. McQueney, X. Qiu, W. W. Smith, S. S. Abdel-Meguid, *Nat. Struct. Biol.* **1997**, *4*, 109–111.

[161]D. Turk, M. Podobnik, T. Popovic, N. Katunuma, W. Bode, R. Huber, V. Turk, *Biochemistry* **1995**, *34*, 4791–4797.

[162]A. Yamamoto, T. Tomoo, K. Matsugi, T. Hara, Y. In, M. Murata, K. Kitamura, T. Ishida, *Biochim. Biophys. Acta* **2002**, *1597*, 244–251.

[163]M. E. McGrath, J. L. Klaus, M. G. Barnes, D. Bromme, *Nat. Struct. Biol.* **1997**, *4*, 105–109.

[164]A. Yamamoto, K. Tomoo, T. Hara, M. Murata, K. Kitamura, T. Ishida, *J. Biochem.* **2000**, *127*, 635–643.

[165]SYBYL *7.0*, Tripos Inc., St. Louis MO, USA **2004**.

[166]P. D. Greenspan, K. L. Clark, R. A. Tommasi, S. D. Cowen, L. W. Mcquire, D. L. Farley, J. H. Van Duzer, R. L. Goldberg, H. Zhou, Z. Du, J. J. Fitt, D. E. Coppa, Z. Fang, W. Macchia, L. Zhu, M. P. Capparelli, R. Goldstein, A. M. Wigg, J. R. Doughty, R. S Bohacek, A. K. Knap, *J. Med. Chem.* **2001**, *44*, 4524–4534.

[167]S. Chowdhury, J. Sivaraman, J. Wang, G. Devanathan, P. Lachance, H. Qi, R. Menard, J. Lefebvre, Y. Konishi, M. Cygler, T. Sulea, E. O. Purisima, *J. Med. Chem.* **2002**, *45*, 5321–5329.

[168]H. Volkel, U. Kurz, J. Linder, S. Klumpp, V. Gnau, G. Jung, J. E. Schultz, *Eur. J. Biochem.* **1996**, *238*, 198–206.

[169]H. Shao, X. Jiang, P. Gantzel, M. Goodman, *Chem. Biol.* **1994**, *1*, 231–234.

[170]D. B. Turner, P. Willet, A. M. Ferguson, T. Heritage, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409–422.

[171]N. Bodor, M. J. Huang, *Pharm. Sci.* **1992**, *81*, 272–281.

[172]J. M. Luco, F. H. Ferretti, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392–401.

[173]J. J. Sutherland, L. A. O'Brien, D. F. Weaver, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906–1915.

[174]*http://pubs3.acs.org/acs/journals/supporting_information.page?in_manuscript=ci034143r*.

[175]MOE *2002.03*, Chemical Computing Group Inc., Montreal, Canada **2002**.

Danke!

Nullum enim officium referenda gratia magis necessarium est.
—Marcus Tullius Cicero, römischer Konsul, Redner und Philosoph, † 43 v. Chr.

- Mein Dank gilt an erster Stelle meinem Mentor Knut Baumann. Die Anschaulichkeit seiner Erklärungen waren der initiale Funke für mein wachsendes Interesse an der Chemometrie — einer Disziplin, die mir während des Studiums unbekannt geblieben war. Bei der Entwicklung eigener Ideen hat mir seine kritisch-hinterfragende Blickweise ebenso geholfen wie seine umgängliche und freundschaftliche Art. Letztere war auch für die Überbrückung der räumlichen Distanz zwischen Würzburg und Martinsried ein entscheidender Faktor.

- Bernd Kramer danke ich für viele hilfreiche Ratschläge und interessante Diskussionen sowie die Entwicklung so mancher verrückter Idee. Meinen übrigen Kollegen bei der 4SC AG — insbesondere Jürgen, Kristina, Markus und Thomas — bin ich für ihre Hilfsbereitschaft und die gute Arbeitsatmosphäre dankbar.

- Eine ebenso angenehme wie erfolgreiche Kooperation mit guten Ergebnissen bestand im Cysteinprotease-Projekt. Dafür danke ich Radim Vicik und Tanja Schirmeister.

- Bei meinen Mit-Doktoranden an der Uni Würzburg bedanke ich mich für die gute Zusammenarbeit. Von Nik konnte ich besonders in der Anfangszeit viel lernen, Sepp hat mir vor allem beim Korrekturlesen dieser Arbeit sehr geholfen, und Sebastian und Ulrike verdanke ich nicht zuletzt die Einsicht, daß meine Besuche in Würzburg gerade zur Weinfest-Zeit viel zu selten waren.

- Sehr dankbar bin ich schließlich meinen Eltern. Sie haben meine Interessen von Kindheit an gefördert und mich während Studium und Promotion stets großzügig unterstützt. Ohne ihre Hilfe hätte ich niemals den Punkt erreicht, an dem ich heute stehe.

Curriculum vitae

Matthias Busemann, geboren am 27. September 1976 in Würzburg.

Seit 01.2006	Computational Chemist, Jenapharm GmbH & Co. KG, Jena
Seit 10.2003	Doktorarbeit unter Anleitung von PD Dr. Knut Baumann
03.2003	Abschluß Dipl.-Chem., Universität Würzburg
11.2002–12.2005	Diplomand/Doktorand, 4SC AG, Martinsried
08.2002–03.2003	Diplomarbeit unter Anleitung von Dr. Knut Baumann
10.1999–07.2002	Hauptstudium Chemie, Universität Würzburg
10.1999	Vordiplom Chemie, Universität Würzburg
10.1997–10.1999	Grundstudium Chemie, Universität Würzburg
07.1996–08.1997	Zivildienst im Zentrum für Labordiagnostik, St.-Bernward-Krankenhaus, Hildesheim
1996	Abitur
1989–1996	Bischöfliches Gymnasium Josephinum, Hildesheim
1987–1989	Röntgen-Gymnasium, Würzburg
1983–1987	Heuchelhof-Grundschule, Würzburg