# Entwicklung einer wissensbasierten Bewertungsfunktion zur Struktur- und Affinitätsvorhersage von Protein-Ligand-Komplexen

**Dissertation**

**zur**

**Erlangung des Doktorgrades**

**der Naturwissenschaften**

**(Dr. rer. nat.)**

dem

Fachbereich Pharmazie

der Philipps-Universität Marburg

vorgelegt von

**Holger Gohlke**

aus Langen / Hessen

Marburg / Lahn 2000

Vom Fachbereich Pharmazie der Philipps-Universität Marburg

| | |
|---|---|
| als Dissertation angenommen am: | 28. Juni 2000 |
| Erstgutachter: | Prof. Dr. G. KLEBE |
| Zweitgutachter: | Prof. Dr. G. FRENKING |
| Tag der mündlichen Prüfung: | 28. Juni 2000 |

Die vorliegende Arbeit wurde auf Anregung von Herrn Prof. Dr. G. KLEBE am Institut für Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von September 1997 bis Mai 2000 durchgeführt.

Herrn Prof. Dr. G. KLEBE gilt mein besonderer Dank für die sehr interessante Aufgabenstellung und die stetige Diskussionsbereitschaft sowie freundliche Unterstützung während dieser Arbeit. Es war die erlebte Mischung aus Anleitung und gewährter Eigenständigkeit, die stets neu motivierte.

Herrn Prof. Dr. F. DIEDERICH danke ich für die Ermöglichung eines Aufenthaltes an der Eidgenössischen Technischen Hochschule Zürich.

Herrn Dr. M. HENDLICH gilt mein Dank für die anfängliche Hilfestellung und die anregenden, kritischen Diskussionen. Herrn Dr. M. T. STUBBS danke ich für Hinweise zum kristallographischen Hintergrund der Arbeit.

Mein Dank gilt auch Herrn T. MIETZNER für seine beständige Unterstützung bei mathematischen Fragen.

Allen Mitgliedern der Arbeitsgruppe danke ich für das gute Arbeitsklima, die stetige Hilfsbereitschaft sowie die anregenden fachlichen Diskussionen. Besonderer Dank geht an Frau J. GÜNTHER, Herrn Dr. W. NISSINK, Frau A. SCHAFFERHANS und Frau K. SCHULZ für das Korrekturlesen.

Teile dieser Arbeit wurden bereits veröffentlicht:

**Aufsätze:**

- Klebe, G., Böhm, M., Dullweber, F., Grädler, U., Gohlke, H., Hendlich, M. (1999). Structural and Energetic Aspects of Protein-Ligand Binding in Drug Design. In *Molecular Modelling and Prediction of Bioactivity* (Gundertofte, K., Jorgensen, F., Ed.), S. 103-110, KLUWER Academic / Plenum Publ., New York.
- Gohlke H., Hendlich M., Klebe G. (2000). Knowledge-based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **295**, 337–356.
- Gohlke, H., Hendlich M., Klebe G. (2000). Predicting Binding Modes, Binding Affinities and "Hot Spots" for Protein-Ligand Complexes Using a Knowledge-based Scoring Function. *Persp. Drug Discov. Design*, im Druck.
- Klebe, G., Grädler, U., Grüneberg, S., Krämer, O., Gohlke, H. (2000). Understanding Receptor-Ligand Interactions as a Prerequisite for Virtual Screening. In *Virtual Screening for Bioactive Molecules* (Böhm, H. J., Schneider, S., Ed.), Wiley-VCH, Weinheim, im Druck.

**Tagungsbeiträge:**

- Gohlke, H., Hendlich, M., Klebe G., Knowledge-based Scoring Function to Predict Protein-Ligand Interactions, Poster präsentiert beim *12. Workshop on Molecular Modeling*, Darmstadt, 19. - 20. Mai 1998.
- Gohlke, H., Hendlich, M., Klebe, G., Development of a Knowledge-based Scoring Function to Predict Protein-Ligand Interactions, Vortrag präsentiert bei der *Doktorandentagung der Deutschen Pharmazeutischen Gesellschaft*, Freiburg, 11. – 13. März 1999.
- Gohlke, H., Hendlich, M., Klebe, G., Using Empirical Potentials to Predict Protein-Ligand Interactions, Vortrag präsentiert beim *Workshop on Virtual Screening*, Rauischholzhausen, 15. – 18. März 1999.

# Inhaltsverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| Abb. | Abbildung |
| Alg. | Algorithmus |
| ASP | Atomarer Solvatationsparameter |
| CSD | Datenbank mit Kristallstrukturen niedermolekularer Verbindungen (Cambridge Structural Database) (Allen *et al*., 1991) |
| FEP-MD | Freie Energie-Störungsrechnung / Molekulardynamik |
| GB/SA | Generalisierter Born-Ansatz |
| Kap. | Kapitel |
| LIE | Lineare Wechselwirkungsenergie |
| ME | Hauptgleichung (*Master Equation*) |
| NBTI | Nicht-Boltzmann Themodynamische Integration |
| PBE | Poisson-Boltzmann-Gleichung |
| PDB | Proteindatenbank (Bernstein *et al*., 1977) |
| PLS | *Partial Least Squares*-Verfahren |
| QSAR | Quantitative Struktur-Aktivitäts-Beziehungen |
| rmsd | mittlere quadratische Abweichung in den kartesischen Koordinaten einander entsprechender Atome in zwei Molekülen (*root mean-square deviation*) |
| SAMPLS | *Sample-distance Partial Least Squares*-Verfahren |
| SAS | Lösemittel-zugängliche Oberfläche (*Solvent Accessible Surface*) |
| Tab. | Tabelle |

# 1 Einleitung und Problemstellung

Das Paradigma der wechselseitigen molekularen Erkennung ist Grundlage für das Verständnis fast aller Prozesse in biologischen Systemen. Begründet vor etwa 100 Jahren durch Emil Fischers Erkenntnis, *„daß Enzym und Glycosid wie Schloß und Schlüssel zueinander passen müssen, um eine chemische Wirkung aufeinander ausüben zu können“* (Fischer, 1894) sowie Paul Ehrlichs Aussage *„Corpora non agunt nisi fixata“*[*] (Ehrlich, 1913), bildet es heute – in z. T. erweiterter Form (Koshland, 1994) – auch den Ausgangspunkt für die wissenschaftliche Erklärung der Wirkung von Arzneistoffen. Danach bewirkt die geometrische und chemisch komplementäre Anlagerung kleiner Moleküle (im folgenden: *Liganden*) an biologische makromolekulare Zielstrukturen (häufig Proteine, im folgenden allgemein *Rezeptoren* genannt) einen Eingriff in die Abfolge von Stoffwechsel- und Signaltransduktionsprozessen und löst somit einen physiologischen Effekt aus.

Dieses Wissen um die Zusammenhänge zwischen molekularer Struktur und biochemischer Wirkung hat in den letzten Jahren zu einem grundlegenden Wandel der Methoden der modernen Arzneistofforschung geführt. Molekularbiologische Verfahren ermöglichen es, die gestörte Aktivität eines Rezeptors als Ursache einer Krankheit zu identifizieren sowie Proteine in reiner Form zu gewinnen. Die Aufklärung ihrer dreidimensionalen Struktur ist dann vielfach mit Hilfe der Röntgenstrukturanalyse (Drenth, 1999; Glusker *et al*., 1994), NMR-Spektroskopie (Wüthrich, 1986) oder Kryo-Elektronenmikroskopie (Kühlbrandt & Williams, 1999) möglich. Zusätzlich wächst die Anzahl bekannter Proteinsequenzen als Folge der Genomsequenzierungs-Projekte (Collins *et al*., 1998) stark an. Davon ausgehend werden Technologien zur Vorhersage von Proteinfunktion (Lottspeich, 1999; Marcotte *et al*., 1999) und -struktur (Rost, 1998; Westhead & Thornton, 1998) bzw. zur Beschleunigung der Röntgenstrukturanalyse (Burley *et al*., 1999) entwickelt. Computergestützte Verfahren zur Identifikation und Selektion neuer biologischer Zielmoleküle (Jones & Fitzpatrick, 1999; Skolnick & Fetrow, 2000) werden bereits eingesetzt (Spaltmann *et al*., 1999). Als Folge dieser Anstrengungen ist für die Zukunft ein wesentlich breiteres Spektrum erkannter molekularer Zielsysteme für die Arzneimitteltherapie und ihrer biologischen Funktion zu erwarten.

Der Entwicklungsprozeß eines neuen Arzneistoffes, der bis zu 15 Jahre dauern und bis zu einer halben Milliarde Dollar verschlingen kann (Petsco, 1996), läßt sich in mehrere Phasen

[*] *“Die Körper wirken nicht, wenn sie nicht gebunden sind.”*

unterteilen (Böhm *et al*., 1996; Silvermann, 1994). Der am Anfang stehenden Suche nach einer Leitstruktur, d.h. einem Liganden mit zu erkennender Affinität zu einem gegebenen Rezeptor, folgen mehrere Optimierungsphasen. Neben der Erhöhung der Affinität und Selektivität des Liganden zu seinem biologischen Zielmolekül geht es zusätzlich um die Einstellung günstiger pharmakokinetischer Eigenschaften. Hierunter versteht man Aspekte der Absorption des Arzneistoffes, der Verteilung und Metabolisierung im Organismus sowie der Exkretion und Toxizität (Chan & Stewart, 1996). Abschließend erfolgt eine Phase der klinischen Testung.

Der anfänglichen, schnellen und verläßlichen Identifizierung von Liganden kommt eine besondere Bedeutung zu im Hinblick auf die zu erwartende Zunahme bekannter Rezeptoren und die sich daraus ergebende Notwendigkeit, schon in der Frühphase einer Arzneistoffentwicklung deren Aussicht auf Erfolg abschätzen zu müssen. Gegenwärtig existieren für diese Suche nach neuen Leitstrukturen zwei generelle, einander ergänzende Methoden: das *Screening*, d.h. das Testen großer Substanzbibliotheken, und das rationale, auf der Basis von vorhandenen Strukturinformationen beruhende *Design*.

Das *Random-Screening*-Verfahren orientiert sich an der Methode der traditionellen Arzneistoff-Suche, eine große Anzahl synthetischer und natürlicher Substanzen unabhängig von ihrer Struktur auf eine eventuelle Wirkung in einem Bioassay zu untersuchen (Carell *et al*., 1994; Gordon *et al*., 1996). Belebt wurde diese Methode in vergangener Zeit durch den Einsatz von Roboter-Verfahren zum Hochdurchsatztesten (Houston & Banks, 1997) sowie die Verfahren der kombinatorischen Chemie (Balkenhohl *et al*., 1996; Terrett *et al*., 1995) bzw. automatisierten Parallelsynthese, mit deren Hilfe in kurzer Zeit Bibliotheken mit mehreren zehntausend Substanzen ausgehend von einigen wenigen Reaktanden hergestellt werden können. Die Trefferraten dieser zeit- und kostenintensiven, wegen ihrer ungerichteten Vorgehensweise („so viel wie möglich und so schnell wie möglich“) auch als „irrational“ bezeichneten Verfahren betragen weniger als eine Promille der Anzahl der eingesetzten Substanzen (Lahana, 1999). In jüngster Zeit werden daher besonders *Nonrandom*- oder *Targeted-Screening*-Ansätze entwickelt, bei denen die Testsubstanzen hinsichtlich maximaler paarweiser Diversität z.B. in Bezug auf ihre chemischen Eigenschaften (Warr, 1997), zu erwartender günstiger pharmakokinetischer Eigenschaften (Clark & Pickett, 2000) oder anderer Kriterien computergestützt selektiert werden.

Einen ganz anderen Ansatz verfolgt das *rationale Wirkstoffdesign.* Auf der Basis der Analyse eines möglichen Wirkungsmechanismus werden gezielte Vorschläge für eine Leitstruktur entwickelt, die anschließend im Experiment getestet werden. Das Ergebnis fließt als neue Information wieder in den Designzyklus zurück (Abb. 1) (Boyd, 1998; Martin *et al*., 1999). Unter Einbeziehung des maximal zur Verfügung stehenden Wissens geht es hierbei in erster Hinsicht darum, aktiv verläßliche *Vorhersagen* zu treffen, und als Folge dessen weniger auf „glückliche Fügungen" vertrauen zu müssen.

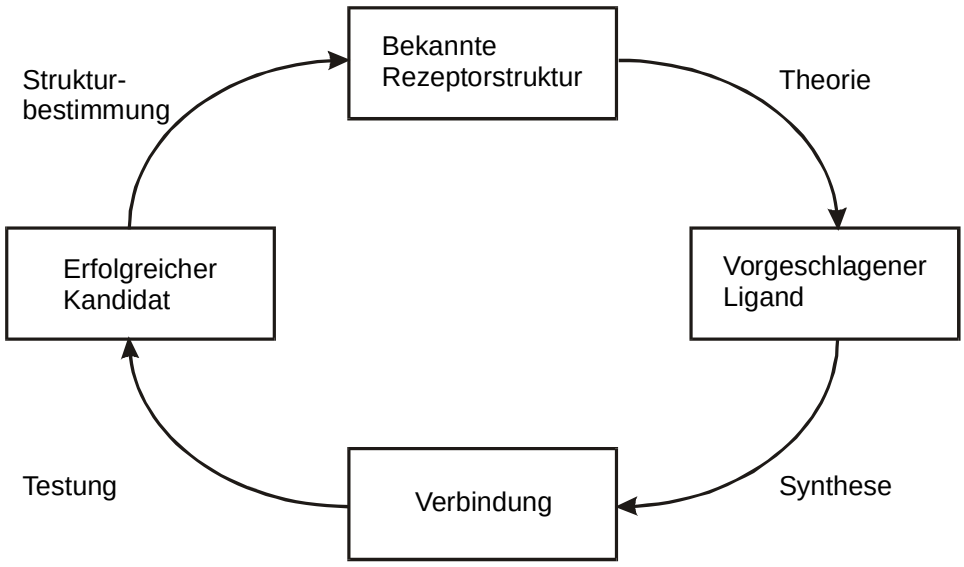



**Abb. 1:** Allgemeiner Ansatz zum strukturbasierten Design von Inhibitoren. Ausgehend von einer bekannten Rezeptorstruktur werden auf Grundlage theoretischer Verfahren mögliche Liganden vorgeschlagen. Nach Synthese und biologischer Testung wird von erfolgreichen Inhibitorkandidaten eine Rezeptor-Ligand-Struktur ermittelt, die als Start für einen neuen Designzyklus dient.

Obwohl dieser Ansatz noch relativ jung ist bzw. erst in neuerer Zeit von den Fortschritten in der Computertechnologie und –methodologie (Buzbee, 1993; Couzin, 1998) profitiert hat, gibt es schon eine ganze Reihe von Beispielen, bei denen er zu Neuentwicklungen bzw. Optimierungen von Wirkstoffen geführt hat (Babine & Bender, 1997; Boyd, 1990; Greer *et al*., 1994; Klebe, 1998b; Kubinyi, 1998). So wurden mit Kenntnis der Struktur des Rezeptors Liganden von Thymidylat-Synthase (Shoichet *et al*., 1993), Purin-Nucleosid-Phosphatase (Montgomery *et al*., 1993), Streptavidin (Katz *et al*., 1995; Weber *et al*., 1994), Sialidase (von Itzstein *et al*., 1996), Selectin (Kogan *et al*., 1995), FKBP12 (Andrus & Schreiber, 1993), Calmodulin (Hardcastle *et al*., 1995), Elastase (Veale *et al*., 1995), Thermolysin (Morgan *et al*., 1994), Thrombin (Hilpert *et al*., 1994), Papin (Cheng *et al*., 1994) und Renin (Plummer *et*

*al*., 1995) erfolgreich vorgeschlagen bzw. optimiert. Darüber hinaus existieren bereits in der Therapie eingesetzte Wirkstoffe, die auf diesem Weg gefunden wurden: z. B. der Angiotensin-*Converting*-Enzym(ACE)-Hemmer Captopril (Wermuth, 1996), der Carboanhydrase-Inhibitor Dorzolamid (Greer *et al.*, 1994) sowie die HIV-Protease-Inhibitoren Saquinavir, Indinavir, Ritonavir und Nelfinavir (Vacca & Condra, 1997).

Die Verwendung der jeweiligen Verfahren des rationalen Designs hängt davon ab, ob die dreidimensionale Struktur des biologischen Zielmoleküls bekannt ist oder nicht. In letzterem Fall ermöglichen es „Quantitative-Struktur-Wirkungs-Beziehungen“ (QSAR-Verfahren) (Kubinyi, 1993; Kubinyi *et al*., 1997), Modelle für den Zusammenhang von Molekülstruktur und Wirkung auf einen Rezeptor aus einer Serie von Wirkstoffen und dazugehörigen experimentell bestimmten Affinitäten zu erstellen. Diese Modelle können dann nicht nur zur Erklärung beobachteter Affinitäten, sondern auch zur Vorhersage der Affinitäten unbekannter Verbindungen herangezogen werden. Ein alternatives Verfahren besteht darin, aus einer Reihe von wirksamen Verbindungen ein Pharmakophormodell zu erstellen (Martin, 1999), d.h. eine geometrische Abbildung der molekularen Eigenschaften der Wirkstoffe, die notwendig für eine biologische Aktivität sind. Mit diesem Pharmakophormodell kann in einem Folgeschritt in einer Datenbank von Molekülen nach neuen, potentiell aktiven Verbindungen gesucht werden (Marriot *et al*., 1999).

Die ständig wachsende Anzahl dreidimensionaler Strukturen der makromolekularen Rezeptoren (Bernstein *et al.*, 1977) bildet die Grundlage des *strukturbasierten Entwurfs* von Wirkstoffen. Hauptsächlich werden aus der Röntgenstrukturanalyse erhaltene Rezeptorgeometrien eingesetzt, wobei angenommen werden kann, daß diese Geometrien den aktiven Strukturen in Lösung gleichen (Blake *et al*., 1981; Doscher & Richards, 1963). Unter Ausnutzung der Informationen über die Eigenschaften der Ligandbindestelle und der auf dem Schlüssel-Schloß-Prinzip beruhenden zentralen Annahme, daß ein bindender Ligand deutliche strukturelle und chemische Komplementarität zum Wirkort aufweisen muß, ergeben sich zwei Ansätze zum computergestützten Wirkstoffdesign. Im Rahmen des *de novo-Designs* werden in der Bindetasche neue Leitstrukturen durch Verknüpfung vorher plazierter Atome oder Fragmente oder den sukkzessiven Aufbau eines Moleküls ausgehend von einer „Keimstelle“ erzeugt (Clark *et al*., 1997; Murcko, 1997). Alternativ dazu kann eine Datenbank aus Molekülen nach einem komplementären Liganden durchsucht werden, indem die einzelnen Moleküle (flexibel) in die Ligandbindestelle eingepaßt werden (*Docking-Verfahren*) (Lengauer & Rarey, 1996). In beiden Fällen erfolgt abschließend eine *Affinitätsvorhersage* des Liganden zu

seinem Rezeptor: nur wenn diese Bewertung mit akzeptabler Genauigkeit und Zuverlässigkeit erfolgt, können neue Leitstrukturen ohne aufwendige Experimente gefunden werden. In letzterem Fall spricht man daher auch von *virtuellem Screening* (Walters *et al*., 1998). Dies führt nicht nur zu einer deutlichen Reduzierung der Entwicklungszeiten und Kosten. Vielmehr können die gewonnen Struktur- und Affinitätsinformationen auch zur weitergehenden Leitstruktur*optimierung* genutzt werden.

Zwei Aspekte bestimmen maßgeblich den Erfolg des computergestützten, strukturbasierten Liganden-Designs unter Verwendung der Proteinstruktur: die Generierung relativer Anordnungen von Ligand und Rezeptor (*Konfigurations-Generierungsproblem*) sowie das Erkennen der korrekten Bindungsgeometrie und die Vorhersage der Affinität der so erhaltenen Rezeptor-Ligand-Komplexe (*Struktur- und Affinitäts-Vorhersageproblem*). Ein 1997 unabhängig durchgeführter Test von Docking-Verfahren (Dixon, 1997) bestätigte die Ansicht, daß Struktur- und Affinitätsvorhersagen nur unzureichend gelingen (Ajay & Murcko, 1995; Knegtel & Grootenhuis, 1998; Oprea & Marshall, 1998), wohingegen das Problem der Generierung der relativen Orientierung von Ligand und Rezeptor als weitgehend gelöst angesehen werden kann (Kramer *et al*., 1999; Kuntz *et al*., 1994).

Das Thema dieser Arbeit konzentriert sich daher auf das *Struktur- und Affinitätsvorhersageproblem*. Hierzu ergeben sich folgende Problemstellungen mit ansteigendem Komplexitätsgrad:

I. Für *einen* gegebenen Rezeptor, der in seiner dreidimensionalen Struktur bekannt ist, ist aus einem Satz von dazu relativen Orientierungen *eines* Liganden eine Ligandkonfiguration zu bestimmen, die der tatsächlichen (experimentellen) Orientierung entspricht.

II. Für *einen* gegebenen Rezeptor, der in seiner dreidimensionalen Struktur bekannt ist, ist die korrekte Reihenfolge *mehrerer* Liganden hinsichtlich ihrer Affinitäten zu diesem Rezeptor zu bestimmen.

III. Für *mehrere* gegebene Rezeptoren, die in ihrer dreidimensionalen Struktur bekannt sind, ist die jeweils korrekte Affinität *eines* oder *mehrerer* Liganden zu diesen Rezeptoren zu bestimmen.

Diese Formulierungen entsprechen Fragestellungen, die während eines strukturbasierten Liganden-Designs auftreten: der Vorhersage der korrekten Geometrie von Rezeptor und Ligand (I); der Notwendigkeit, im Rahmen des virtuellen Screenings Liganden für nachfolgende

experimentelle Evaluierungen zu priorisieren (II); der Vorhersage der Selektivität von Liganden in Bezug auf mehrere Rezeptoren (III).

Aus dem angestrebten praktischen Einsatz des zu entwickelnden Verfahrens ergeben sich zusätzliche Randbedingungen:

- *breite Anwendbarkeit* und *Robustheit* der Methode in Bezug auf verschiedene Kombinationen von Rezeptoren und Liganden
- *Genauigkeit* und *Fehlertoleranz* hinsichtlich der Vorhersage der Affinität nicht nur bei experimentell bestimmten Rezeptor-Ligand-Komplexen, sondern auch bei computer- bzw. handgenerierten Geometrien
- *Sensitivität* und *physikalische Interpretierbarkeit*, um aus den erhaltenen Ergebnissen Schlüsse für das weitere Vorgehen (etwa bei einer Ligandoptimierung) ziehen zu können
- ausreichende *Geschwindigkeit*, um auch virtuelle Screening-Verfahren mit einer großen Anzahl von zu testenden Verbindungen durchführen zu können
- *„Umgehung" fehlender Informationen* wie etwa solche über Protonierungszustände funktioneller Gruppen

In Kapitel 2 werden Faktoren beschrieben, die die Affinität eines Liganden zu einem Rezeptor bestimmen. Kapitel 3 gibt einen Überblick über existierende Verfahren zur Vorhersage von Rezeptor-Ligand-Affinitäten. Kapitel 4 beschreibt die dem entwickelten Ansatz zugrundeliegende Theorie sowie die angewendeten Methoden. Die erzielten Ergebnisse werden in Kapitel 5 ausgewertet und diskutiert. Kapitel 6 faßt die Ergebnisse der Arbeit zusammen und stellt Grenzen sowie mögliche Verbesserungen des Ansatzes dar.

# 2 Bindungsaffinität und inter- / intramolekulare Beiträge - makroskopische Größe und mikroskopische Ursachen

## *2.1 3D-Rezeptor-Ligand-Strukturen – Sichtfenster in die Welt der Wechselwirkungen*

3D-Strukturen von Rezeptoren (und Liganden) bilden nicht nur die Basis für das strukturbasierte Wirkstoffdesign. Viele der im folgenden Abschnitt erläuterten Beiträge zur Bindungsaffinität von Rezeptor-Ligand-Komplexen wären ohne die Möglichkeit ihrer Untersuchung auf molekularer Ebene bis heute nur unvollständig verstanden. Dazu notwendige Informationen über die Anordnung der Atome im Raum liefert – neben den Methoden der hochaufgelösten NMR-Spektroskopie (Clore & Gronenborn, 1991; Wüthrich, 1986) – v.a. die Proteinkristallographie (Chayen *et al*., 1996; Drenth, 1999; Glusker *et al.*, 1994).

Die Grundlage der (Protein-)Kristallographie bildet die dreidimensional-periodische Anordnung der Bausteine im (Protein-)Kristall: durch die Anwendung von Symmetrieoperationen erzeugt man aus der asymmetrischen Einheit, dem „Grundmotiv“, die Elementarzelle. Nachfolgende Anwendung von Gittertranslationen erzeugt daraus die Kristallstruktur (Borchardt-Ott, 1993). Trifft elektromagnetische Strahlung mit gegebener Wellenlänge auf ein Objekt gegebener Ausdehnung, so kommt es zu Beugungseffekten, wenn Wellenlänge und Ausdehnung von gleicher Größenordnung sind. Im Falle von Kristallen gilt dies für Strahlung im Nanometerbereich. Röntgenstrahlen werden durch das Phänomen der kohärenten Streuung an den Elektronen der Atome gebeugt; Neutronenstrahlung hingegen an den Atomkernen.

Dies hat Auswirkungen auf die erhaltenen Ergebnisse: während die Amplitude gestreuter Röntgenstrahlung proportional zur Anzahl der Elektronen um dieses Atom sowie eine Funktion des Streuwinkels ist (Hartree, 1925; James, 1954), bestimmt die Masse des Atomkerns sowie die Energie der Wechselwirkung zwischen diesem und dem Neutron die Streuung der Neutronenstrahlung. Röntgenstrahlung kann daher nicht zwischen Isotopen bzw. nur schlecht zwischen Elementen ähnlicher Ordnungszahl unterscheiden; leichte Elemente wie Wasserstoff haben ein sehr geringes Streuvermögen. Im Falle von Proteinkristallen bedeutet das, daß die Positionen der terminalen N- und O-Atome von Asparagin und Glutamin nur aufgrund eines in sich konsistenten Wasserstoffbückennetzwerkes zugeordnet werden können. Analog sind häufig zwei mögliche Orientierungen des Imidazol-Rings von Histidin denkbar. Die Positionen von Wasserstoffen bleiben in Proteinkristallen meist unbestimmt. Dies ist v.a. bedeutsam für konformativ flexible H-Atome (z.B. an terminalen Hydroxyl- oder Aminogrup-

pen) sowie für (de-)protonierbare Gruppen. Allerdings lassen sich durch die Analyse der umgebenden Atome in einem Kristall meist weitere Hinweise auf den Protonierungszustand finden.

Durch positive und negative Interferenz der an allen Atomen des Kristalls gemäß den Laue- bzw. Bragg-Gleichungen gebeugten Röntgenstrahlen entsteht ein Beugungsbild. Aus den Intensitäten der Reflexe dieses Beugungsbildes, sowie den im Falle der Proteinkristallographie meist indirekt bestimmten relativen Phasenlagen (Beauchamp & Isaacs, 1999) läßt sich unter Anwendung einer Fouriertransformation eine Elektronendichtekarte als Darstellung der Atomanordnung im Kristall berechnen. Diese experimentell bestimmte Repräsentation dient nun als Grundlage für die Erstellung eines Molekülmodells. In folgenden Schritten der Verfeinerung werden die Unterschiede zwischen der beobachteten Elektronendichte sowie einer aus dem Modell berechneten durch Variation des Modells mittels Auswertung von Differenzdichtekarten, Verfahren des kleinsten Fehlerquadrates sowie globalen Optimierungsverfahren (*simulated annealing*) minimiert.

Als Maß für die Übereinstimmung zwischen Modell und Experiment dient während der Verfeinerung der R-Faktor. Während er für zufällig in die Einheitszelle plazierte Atome einer nicht-zentrosymmetrischen Struktur zu 0.59 berechnet werden kann, zeigen optimal verfeinerte Proteinstrukturen Werte unter 0.2. Methoden zur Kreuzvalidierung (Brünger, 1992) zur Verhinderung einer Überanpassung werden heute ebenfalls verwendet. Allerdings wird die Qualität des abschließenden Modells maßgeblich schon durch das Streuvermögen des untersuchten Kristalls limitiert. Baufehler und Unordnungsphänomene im Kristall begrenzen die Auflösung, d.h. das Vermögen, zwei nahe Objekte noch als getrennte Entitäten wahrnehmen zu können. Für Proteinkristalle werden selten Auflösungen kleiner als 1.5 Å gemessen (also in der Größenordnungen von Bindungslängen), meist erhält man Werte zwischen 2 und 3 Å. Von atomarer Auflösung spricht man bei Werten unter 1.2 Å. Inhärent verknüpft mit dem Maß der Auflösung ist die Größe des Fehlers der Koordinaten der Atompositionen: für eine Auflösung von 2.5 Å ergibt er sich zu etwa 0.4 Å (Dauber-Osguthorpe *et al*., 1988; Kossiakoff *et al*., 1992; Wlodawer *et al*., 1987). Dies ist auch bei der Betrachtung intermolekularer Wechselwirkungen zu berücksichtigen.

Für kristallographisch bestimmte Proteinstrukturen ist außerdem zu berücksichtigen, daß sie als zeitliche und räumliche Mittelung über eine Vielzahl einzelner Moleküle entstanden sind (Brünger, 1997). Kristallkontakte zwischen im Gitter benachbarten Molekülen können zu

intermolekularen Wechselwirkungen führen, die einzelne Teile einer Struktur beeinflussen können. Zusätzlich kommt es neben dem Auftreten von statischer Unordnung durch unterschiedliche Besetzung analoger Atompositionen zu dynamischer Unordnung durch die thermische Bewegung der Atome um ihre Gleichgewichtslagen. Selbst für Atome mit einem durchschnittlichen Temperaturfaktor von B = 20 $Å^2$ beträgt die mittlere Verschiebung von der Gleichgewichtsposition 0.5 Å. Durch den Mittelungsprozeß während der Dauer der Sammlung der Beugungsdaten können außerdem nur solche Atome im Kristall erkannt werden, die einen hohen Ordnungsgrad aufweisen. Dies gilt insbesondere für Wassermoleküle, die bis zu 70 % der Atomanzahl eines Proteinkristalls ausmachen können (Carugo & Bordo, 1999; Levitt & Park, 1993). Während die Wassermoleküle der ersten Hydratationssphäre um ein Protein bzw. einen Liganden i.a. wohl geordnet sind, nimmt die Unordnung mit wachsendem Abstand davon zu (Karplus & Faerman, 1994).

Im Fall der Bindung eines Liganden an einen Rezeptor ist als Folge des Mittelungsprozesses außerdem das Auftreten multipler Bindungsmoden zu erwähnen, bei dem ein und derselbe Ligand mehrere verschiedene, energetisch gleichwertige Lagen in der Bindetasche eines Proteins einnehmen kann. Zusätzliche strukturelle Variationen im kristallinen Zustand können darüber hinaus durch die Erscheinung der Polymorphie auftreten. Hierunter wird die Eigenschaft einer Verbindung verstanden, bei kinetisch kontrollierten Bedingungen in unterschiedlichen Kristallstrukturen – mit durchaus verschiedenen physikochemischen Eigenschaften – auftreten zu können (Bernstein, 1989; Verwer & Leusen, 1998). In Anbetracht dieser letzten beiden Punkte wird deutlich, daß nur eingeschränkt von „der“ Kristallstruktur einer Verbindung bzw. eines Protein-Ligand-Komplexes gesprochen werden kann.

## *2.2 Faktoren, die die Bindungsaffinität von Rezeptor-Ligand-Komplexen bestimmen*

Die selektive Bindung niedermolekularer Liganden an ein spezifisches Protein wird durch die strukturelle und energetische Erkennung zwischen beiden bestimmt (Böhm & Klebe, 1996; Klebe & Böhm, 1998; Lehn, 1988). Liganden können sowohl kovalent als auch nichtkovalent mit der biologischen Zielstruktur wechselwirken. Komplexe, bei denen es zur Ausbildung kovalenter Bindungen kommt, wie z.B. zwischen D-Phe-Pro-Arg-chlormethylketon (PPACK) und Thrombin (Bode *et al*., 1989), Trifluormethylketon-Inhibitoren und Elastase (Bernstein *et al*., 1994; Damewood *et al*., 1994) sowie Aspirin und Prostaglandin-Synthase (Roth *et al*., 1975) sollen hier nicht weiter betrachtet werden.

Die nicht-kovalente, reversible Assoziation von Rezeptor (*R*) und Ligand (*L*) zum Rezeptor-Ligand-Komplex (*R'L'*) tritt i.a. in wäßriger, elektrolythaltiger Lösung auf.

$$R_{(aq.)} + L_{(aq.)} \rightleftharpoons R'L'_{(aq.')}$$

Unter der Bedingung der Einstellung des thermodynamischen Gleichgewichts dieser Reaktion läßt sich die Freie Standardenthalpie der Bindung $\Delta G^0$ (im folgenden auch Bindungsaffinität genannt) aus der experimentell ermittelten Assoziationskonstanten $K_A$ (bzw. der dazu reziproken Dissoziations- bzw. Inhibtionskonstanten ($K_D$ bzw. $K_i$)) bestimmen. $\Delta G^0$ setzt sich aus einem enthalpischen ($\Delta H^0$) und einem entropischen bedingten ($T\Delta S^0$) Anteil zusammen. *T* steht für die absolute Temperatur (di Cera, 1995).

$$K_A = K_D^{-1} = K_i^{-1} = \frac{[R'L']}{[R][L]} \qquad \Delta G^0 = -RT \ln K_A = \Delta H^0 - T\Delta S^0 \qquad \textbf{Gl. 1}$$

Gemäß

$$\Delta G^0 = \mu^0_{R'L'(aq'.)} - (\mu^0_{R(aq.)} + \mu^0_{L(aq.)}) \qquad \textbf{Gl. 2}$$

mit $\mu^0_i$ als chemischem Standardpotential der Spezies *i* läßt sich $\Delta G^0$ auch als Funktion der Stabilität des Gesamtkomplexes gegenüber dem freien Liganden und dem freien Rezeptor auffassen (Davis & Teague, 1999; Gilson *et al*., 1997). Experimentell bestimmte Inhibitionskonstanten liegen i.a. in einem Bereich zwischen $10^{-2}$ und $10^{-12}$ M, was einer Freien Standardbindungsenthalpie von –10 bis –70 kJ / mol bei $T$ = 298 K entspricht (Böhm & Klebe, 1996). Eine Änderung der Freien Standardenthalpie um 5.7 kJ / mol bewirkt bei dieser Temperatur eine Änderung der Inhibitionskonstanten um eine Größenordnung.

Allgemein anerkannt ist, daß bei der nicht-kovalenten Bindung eines Liganden an einen Rezeptor elektrostatische Wechselwirkungen – hierunter faßt man Salzbrücken, Wasserstoffbrückenbindungen, Dipol/Dipol-Wechselwirkungen und Wechselwirkungen zu Metallionen zusammen - Solvatations- und Desolvatationsbeiträge sowie die Komplementarität der Raumstruktur eine Rolle spielen (Abb. 2) (Andrews *et al*., 1984; Davis & Teague, 1999; Dean, 1987; Fersht, 1985). Zusätzlichen Einfluß nehmen noch Faktoren, die durch intramolekulare Veränderungen bei Rezeptor (R → R') und Ligand (L → L') im Verlauf der Komplexbildung bedingt sind (Böhm & Klebe, 1996; Klebe & Böhm, 1998).

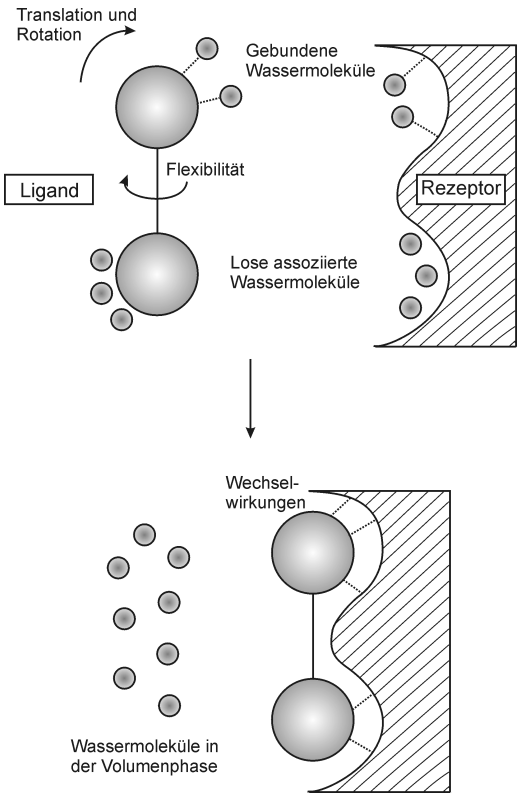



**Abb. 2:** Übersicht über thermodynamische Beiträge zur Bindungsaffinität, die bei der Bildung eines Rezeptor-Ligand-Komplexes aus den solvatisierten Edukten auftreten. Neben der Ausbildung von (elektrostatischen) Wechselwirkungen zwischen Protein und Ligand treten Desolvatationseffekte, Reorganisationen in der Solvathülle, Einschränkungen von Flexibilität und Mobilität bzw. neue niederfrequente Schwingungsmoden auf.

### 2.2.1 Elektrostatische Wechselwirkungen

Die Erkennung der Bedeutung der Wasserstoffbrückenbindung (Jeffrey, 1997; Jeffrey & Saenger, 1991) für die Strukturen von Proteinen und ihren Komplexen mit Liganden geht bis auf Arbeiten von Pauling zurück (Pauling & Corey, 1951). Nichtsdestotrotz herrscht auch heute noch kein einheitliches Bild über ihren relativen Energiebeitrag zur Thermodynamik der Proteinfaltung und Ligandbindung (Fersht, 1987; Murphy & Gill, 1991; Yang *et al*., 1992).

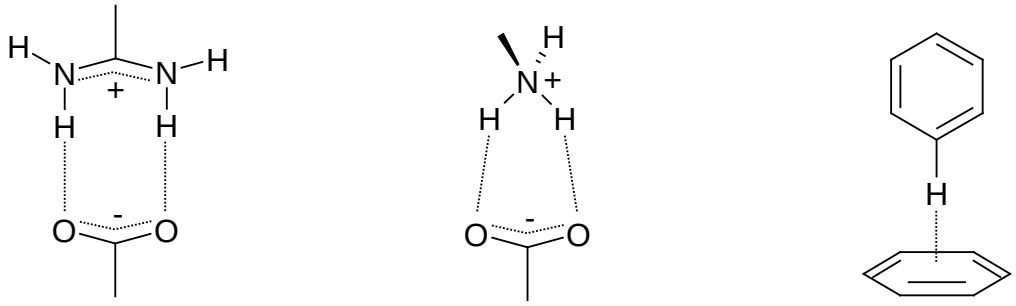
Wasserstoffbrücken beruhen hauptsächlich auf der elektrostatischen Anziehung zwischen einem an ein elektronegatives Atom X (meist N oder O) gebundenes Wasserstoffatom und einem weiteren elektronegativen Atom Y oder einem π-Elektronensystem. Charakteristisch sind Abstände zwischen 2.5 und 3.2 Å (Jeffrey, 1997) zwischen Wasserstoffbrückendonor X und –akzeptor Y sowie ein X-H ··· Y Winkel von 130-180° (Klebe, 1994). Kürzere Abstände bis zu 2.2 Å gehen einher mit einem eher kovalenten Bindungscharakter und größerer Bindungsenergie (Jeffrey, 1997).

Bedingt durch ihren Charakter hängt die Stärke einer Wasserstoffbrücke von ihrer jeweiligen mikroskopischen Umgebung ab: die Abschirmung elektrostatischer Wechselwirkungen ist abhängig von der Dielektrizitätskonstanten $\varepsilon$ des umgebenden Mediums. Während im Inneren von Proteinen $\varepsilon$-Werte von 2 - 4 angenommen werden, beträgt der Wert an der Peripherie des Proteins in unmittelbarer Nähe zum umgebenden Wasser etwa 80 (Honig & Nicholls, 1995). Zusätzlich wird in der Nähe von polaren Gruppen eine höhere Dielektrizitätskonstante erwartet als in unpolarer Umgebung (Nakamura, 1996). Vergrabene Wasserstoffbrücken werden daher als bedeutsamer für die Protein-Ligand-Wechselwirkung angesehen als solche, die zum Lösemittel exponiert sind (Beeson *et al*., 1993; Stahl & Böhm, 1998).

Bei der Bildung einer Wasserstoffbrücke in einem Rezeptor-Ligand-Komplex darf nicht übersehen werden, daß zuvor bestehende, in ihrer Stärke vergleichbare Wasserstoffbücken der funktionellen Gruppen des freien Rezeptors bzw. Liganden mit dem umgebenden Wasser dazu aufgebrochen werden müssen. Die Differenz der Freien Enthalpien aus diesen Beiträgen des H-Brücken-Austauschprozesses bestimmt letztlich, ob die H-Brücken-Bildung zwischen Rezeptor und Ligand zur Bindungsaffinität beiträgt. Ein Beispiel aus der Literatur macht dies deutlich: in einer Arbeit von Connelly *et al.* (Connelly *et al*., 1994) wurde im FK506 bindenden Protein (FKBP-12) Tyrosin-82 zu Phenylalanin mutiert. Damit entfällt bei der Bindung von Tacrolismus bzw. Rapamycin die Möglichkeit zur Ausbildung einer Wasserstoffbrücke zu dem Liganden. Dennoch wurde eine starke enthalpische *Stabilisierung* gegenüber der Bindung an das Wildtyp-Protein gefunden. Hochaufgelöste Röntgenkristallstrukturen zeigten, daß während der Ligandbindung zwei vorher an die Hydroxylgruppe des Tyrosin-82 gebundene Wassermoleküle verdrängt werden mußten. Dies erwies sich in einer nachfolgenden thermodynamischen Analyse als enthalpisch stark ungünstiger Beitrag zur Ligandbindung. Einen weiteren Hinweis auf die Bedeutung der Beteiligung von Wasser in der Gesamtbilanz der Wechselwirkungen gibt die Tatsache, daß nur 1 – 2 % aller vergrabenen N-H bzw. C=O-Gruppen der Amidbindungen eines Proteins keine Wasserstoffbrücke ausbilden (McDonald & Thornton, 1994). Nicht abgesättigte, vergrabene polare Gruppen von Liganden oder Proteinen werden daher als der Komplexbildung stark abträglich angesehen.

Beim physiologischem pH-Wert (ca. 7.4) ist anzunehmen, daß in Proteinen i.a. die Guanidinseitenkette von Arginin ($pK_a$ = 12.5) bzw. die Aminseitenkette von Lysin ($pK_a$ = 10.8) protoniert, die Carboxylgruppen von Asparagin- ($pK_a$ = 3.9) und Glutaminsäure ($pK_a$ = 4.1) dagegen deprotoniert vorliegen ($pK_a$-Werte nach (Dawson *et al*., 1969)). Allerdings hängt der genaue Protonierungszustand auch hier von den lokalen elektrostatischen Verhältnissen in der

Umgebung der jeweiligen funktionellen Gruppe ab und kann sich sogar während der Ligandbindung ändern (Antosiewicz *et al*., 1996). Besitzt der gebundene Ligand in sterisch passender Anordnung entgegengesetzt geladene Gruppen, führt das zu anziehenden elektrostatischen Wechselwirkungen, die auch als „Salzbrücken“ bezeichnet werden (Abb. 3) (Barril *et al*., 1998).



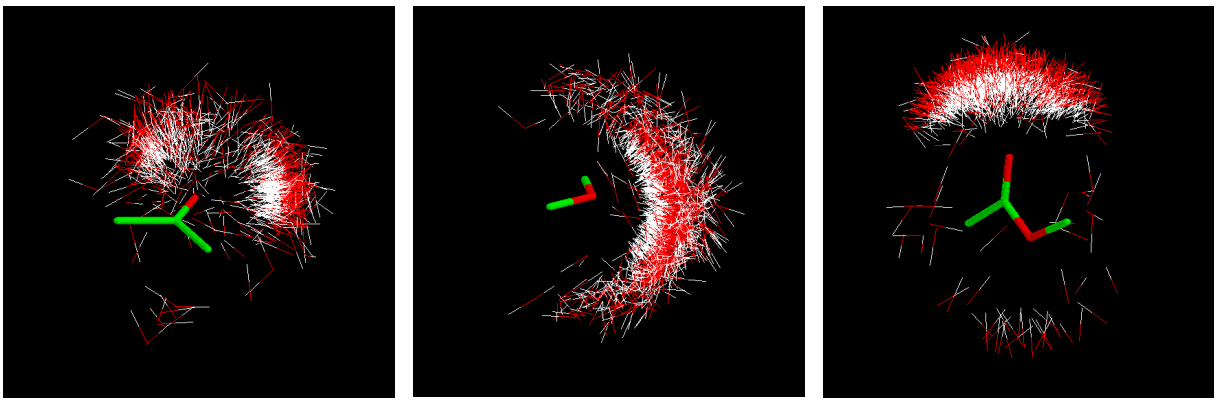

**a)** **b)** **c)**

**Abb. 3:** Beispiele spezieller Wasserstoffbrücken: a), b) bidentate ionische („Salzbrücken“), c) CH¨π-Wechselwirkung

Abschätzungen über die Beiträge der Wasserstoff- und Salzbrücken zur Bindungsaffinität eines Rezeptor-Ligand-Komplexes resultieren aus verschiedenen Ursprüngen. Daten aus Proteinmutationsstudien ergeben Werte für den Beitrag einer Wechselwirkung zwischen ungeladenen Partnern von $\Delta G^0 = -5 \pm 2.5$ kJ / mol (Connelly *et al.*, 1994; Fersht, 1987; Fersht *et al*., 1985). Ähnliche Zahlen resultieren aus Untersuchungen von Strukturen und Lösungsenergien kristalliner zyklischer Dipeptide (Habermann & Murphy, 1996) sowie Studien zum Stabilisierungsbeitrag von intramolekularen Wasserstoffbrückenbindungen in Proteinen (Dill, 1990; Fernandez-Recio *et al*., 1999; Thorson *et al*., 1995). Für ladungsunterstütze Wasserstoffbrükken bzw. Salzbrücken werden dagegen Werte von -10 - -20 kJ / mol angegeben (Fersht *et al.*, 1985; Hossain & Schneider, 1999). Ein Problem bei diesen Untersuchungen ergibt sich allerdings bei der Interpretation der experimentell bestimmten *„augenscheinlichen*“ Bindungsbeiträge: nur wenn zusätzliche Effekte ausgeschlossen werden können, dürfen sie mit dem „*intrinsischen*“ Beitrag einer Wechselwirkung gleichgesetzt werden (Fersht *et al.*, 1985). So wurde von Williams *et al.* der Beitrag einer Wasserstoffbrücke zunächst mit etwa -25 kJ / mol angegeben (Doig & Williams, 1992; Searle *et al*., 1992), nach Berücksichtigung vernachlässigter Einflüsse in späteren Arbeiten aber auf -1 bis -7 kJ / mol reduziert (Williams *et al*., 1993; Williams & Westwell, 1998). Analog werden in einer Studie von Andrews *et al.* zur Ermittlung von Bindungsbeiträgen funktioneller Gruppen in Proteinliganden Wasserstoff-

brücken hinsichtlich ihrer Stabilisierungseigenschaft überschätzt, weil der Bindung abträgliche entropische Beiträge bei der abschließenden Analyse zu groß angesetzt wurden (Andrews *et al.*, 1984).

Wasserstoffbrücken beeinflussen den Ligandbindungsprozeß außerdem durch die starke geometrische Ausrichtung ihrer Wechselwirkungen. Neben theoretischen (Gordon & Jensen, 1996) und spektroskopischen Untersuchungen liefern hierzu v.a. die Analyse kristallographischer Datenbanken wertvolle Informationen (Allen, 1998; Klebe, 1994). So bilden Carbonyl- und Carboxylat-Sauerstoffatome Wechselwirkungen hauptsächlich in Richtung ihrer freien Elektronenpaare aus (Kroon *et al.*, 1975; Murray-Rust & Glusker, 1984); bei letzteren sind die Elektronenpaare in *syn*-Stellung gegenüber denen in *anti*-Stellung bevorzugt (Klebe, 1994; Roe & Teeter, 1993; Tintelnot & Andrews, 1989). Eine Zusammenstellung in Kristallpackungen beobachteter Wechselwirkungsgeometrien ist in der Datenbank IsoStar enthalten (Bruno *et al.*, 1997) (Abb. 4).



**a)** **b)** **c)**

**Abb. 4:** In Kristallpackungen von niedermolekularen Verbindungen beobachtete intermolekulare Wechselwirkungen, wie sie in der Datenbank IsoStar (Bruno *et al.*, 1997) enthalten sind. Gezeigt sind Anordnungen von Hydroxylgruppen um aliphatische Ketone (a), aliphatische Ether (b) und aliphatische Ester (c) als Zentralgruppen.

Ebenfalls bekannt als gerichtete (Cole *et al.*, 1998), wenn auch „schwache“, Wasserstoffbrückenbindungen sind Wechselwirkungen zwischen C-H ··· O, C-H ··· N, C-H ··· π-Systemen und C-H ··· Cl (Desiraju, 1996; Harder, 1999; Taylor & Kennard, 1982), die in den hydrophoberen Bereichen der Proteine auftreten können. Wesentliche Beiträge zur Affinität eines Liganden zu seinem Rezeptor liefern auch sog. π-π-Wechselwirkungen (Hunter & Sanders, 1990) zwischen aromatischen Systemen im Liganden und den Seitenketten von Phenylalanin, Tyrosin und Tryptophan (Hunter *et al.*, 1991; Samanta *et al.*, 1999) sowie Wechselwirkungen

zwischen Kationen – etwa tetra-alkylierten Aminen – und aromatischen Systemen (Dougherty, 1996). Koordinative Bindungen funktioneller Gruppen von Liganden (z. B. Hydroxamate, Carboxylate, Phosphate, Thiole) an Metallionen im Protein stabilisieren ebenfalls Rezeptor-Ligand-Komplexe (Harding, 1999).

### 2.2.2 Beiträge durch Solvatation und Desolvatation

Die molekulare Erkennung zwischen zwei Molekülen findet in biologischen Systemen in wäßriger Umgebung statt. Wasser kommt daher - zusätzlich zu der im vorherigen Kapitel beschriebenen Rolle bei der Energetik von Wasserstoffbrückenbindungen - ein besonderer Einfluß bei der Bildung von Protein-Ligand-Komplexen zu (Covell & Wallquist, 1997; Israelachvili & Wennerstrom, 1996; Lemieux, 1996).

Wassermoleküle können nicht nur in der reinen, kondensierten Phase als H-Brückendonatoren und –akzeptoren wirken und so ein Netz von 3 – 4 H-Brücken pro Molekül ausbilden (Jeffrey & Saenger, 1991). Diese Eigenschaft ist auch bei einer Analyse von 19 hochaufgelösten Kristallstrukturen von Protein-Ligand-Komplexen bei etwa 80% der Wassermoleküle gefunden wurden, die verbrückend zwischen Protein und Ligand wirken (Poornima & Dean, 1995a). Für diese indirekte, Solvens-vermittelte Wechselwirkung wurde unter der Voraussetzung optimaler Geometrie ein Beitrag zur Bindungsaffinität von –10.5 bis -12.5 kJ / mol (Ben-Naim, 1992; Wang & Ben-Naim, 1996) bzw. -7 kJ / mol (Ladbury, 1996) abgeschätzt. Der letztere Wert ergibt sich aus der Bilanz zwischen dem entropischen (-30 J / (mol K) entspr. 9 kJ / mol für $-T\Delta S$ bei 298 K) (Dunitz, 1994) sowie dem enthalpischen (-16 kJ / mol) (Connelly, 1997) Beitrag, die bei der Überführung eines Wassermoleküls aus der Lösungsmittelphase in das Bindungsepitop auftreten.

Bei der Untersuchung der Topographie der angrenzenden Molekülоberflächen zeigt sich, daß verbrückende Wassermoleküle im Bereich der Protein-Ligand-Wechselwirkungen bevorzugt in Einbuchtungen auf der Proteinseite und weniger in solchen auf der Ligandseite sitzen (Kuhn *et al.*, 1992; Ladbury, 1996; Poornima & Dean, 1995b). Eine Serie von hochaufgelösten Kristallstrukturen von an das Oligopeptid-bindende Protein (OppA) gebundenen Lys-xxx-Lys-Liganden (xxx: natürliche und nicht-natürliche Aminosäuren) (Tame *et al.*, 1996; Tame & Wilkinson, 1994) zeigt, daß Wasser quasi als Teil der Proteinstruktur die Spezifität des Proteins beeinflusst. Dabei bewegt es sich jedoch nicht frei in der Ligandbindetasche, sondern besetzt jeweils energetisch günstige, partiell konservierte Positionen (Ladbury, 1996; Poornima & Dean, 1995c).

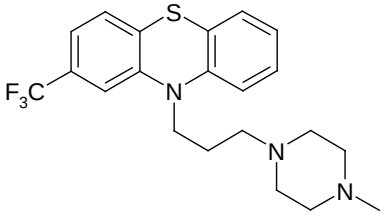
Die besondere Stellung von Wasser unter den Lösungsmitteln – bei außergewöhnlich kleinem Molekülvolumen ein Tetraederkoordination aufweisendes molekulares Netzwerk zu bilden (Eisenberg & Kauzmann, 1969; Franks, 1972-1982) – tritt auch bei der Desolvatation von Protein und Ligand im Verlauf der Komplexbildung hervor. Damit verbunden ist nicht nur das Aufbrechen von H-Brücken zu funktionellen Gruppen, sondern auch die Reorganisation der Wasserstruktur an der Grenzfläche; sowohl enthalpische als auch entropische Einflüsse auf die Bindungsaffinität resultieren daraus (Jencks, 1981; Page, 1977; Searle & Williams, 1992; Searle *et al.*, 1992).

Mit dem Begriff des „hydrophoben Effektes“ beschreibt man die Tatsache, daß die Überführung einer unpolaren Substanz / eines unpolaren Oberflächenbereiches in Wasser a) stark ungünstig ist und b) mit einer Abnahme der Entropie bei Raumtemperatur sowie c) mit einer Zunahme der Wärmekapazität einhergeht (Ben-Naim, 1980; Ben-Naim, 1987; Blokzijl & Engberts, 1993; Dill, 1990; Silverstein *et al.*, 1998; Tanford, 1980). Ein erster Ansatz zu seiner Interpretation geht auf das „Eisbergmodell“ von Frank und Evans (Frank & Evans, 1945; Nemethy & Scheraga, 1962) zurück: bei der Hydratisierung einer unpolaren Substanz kommt es zwar zur *Verringerung* der Anzahl von H-Brücken zwischen Wassermolekülen, allerdings bilden die unmittelbar an der Grenzfläche liegenden Wassermoleküle *stärkere* H-Brücken aus als die in der reinen Phase. Hieraus resultiert eine Clathrat-artige Strukturierung der angrenzenden Wasserhülle sowie ein partielles Einfrieren der Beweglichkeit der Wassermoleküle (Laidig & Daggett, 1996; Muller, 1990; Pertsemlidis *et al.*, 1996). Während der enthalpische Beitrag bei Raumtemperatur dafür gegen Null geht („wenigere, aber stärkere H-Brücken“), nimmt die Entropie aufgrund der erhöhten Ordnung der Moleküle ab. Umgekehrt ergibt sich daraus für das Vergraben einer hydrophoben Oberfläche während der Komplexbildung ein günstiger, entropiegetriebener ($\Delta H \approx 0$, $\Delta S > 0$) Vorgang. Allerdings ist diese klassische Sichtweise nicht allgemein akzeptiert (Blokzijl & Engberts, 1993; Muller, 1990). Ein alternativer Ansatz sieht nicht in der Strukturierung der Wassermoleküle die Ursache für hydrophobe Wechselwirkungen, sondern in einer positiven Enthalpie bedingt durch das Aufbrechen von H-Brücken zur Ausbildung einer Kavität im Wasser, die anschließend die unpolare Substanz aufnimmt (Mancera, 1996; Murphy *et al.*, 1990). Dementsprechend wurden Beiträge von 25 – 100 % der Bindungsenthalpie von Liganden an Proteine bedingt durch Lösemittelreorganisation in kalorimetrischen Studien gefunden (Chervenak & Toone, 1994).

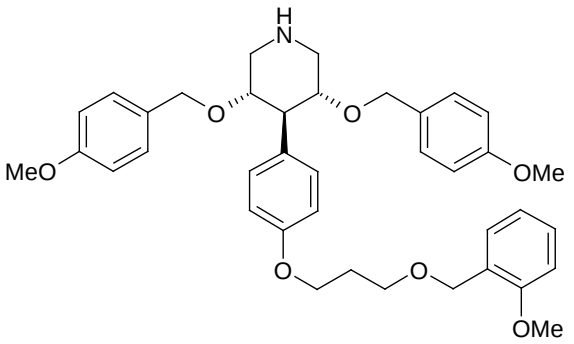
Der Beitrag hydrophober Wechselwirkungen zur Freien Enthalpie bei der Proteinfaltung bzw. der Protein-Ligand-Komplexbildung kann als proportional zur Größe der während dieser

Vorgänge vergrabenen hydrophoben Oberfläche angesehen werden (Chothia, 1974; Chothia & Janin, 1975; Eisenberg & McLachlan, 1986; Ooi *et al*., 1987; Pace, 1992). Damit ergibt sich ein Zugang zur quantitativen Charakterisierung dieser Effekte (Carrupt *et al*., 1997; Williams & Bardsley, 1999). Bei Löslichkeitsstudien von Kohlenwasserstoffen in Wasser erhält man so als Beitrag der hydrophoben Wechselwirkungen -0.10 bis -0.14 kJ / (mol $Å^2$) (Chothia, 1974; Chothia & Janin, 1975; Hermann, 1972; Reynolds *et al*., 1974). Die Analyse der vergrabenen Oberfläche bei Rezeptor-Ligand-Bindungsstellen im Zusammenhang mit experimentell bestimmten Bindungsaffinitäten ergibt Werte von -0.11 bis -0.24 kJ / (mol $Å^2$) als Beiträge zur Freien Enthalpie (Nicholls *et al*., 1991; Ross & Rekharsky, 1996; Searle *et al*., 1992). Die Vergrabung einer Methylgruppe (≈ 25 $Å^2$) liefert daher -2.75 bis –6 kJ / mol, was bei 298 K einer Erhöhung der Assoziationskonstanten um das 3- bis 11-fache entspricht. Aus Mutationsstudien zur Bestimmung des Einflusses der hydrophoben Wechselwirkungen auf die Stabilität von Proteinen resultieren dagegen Werte von -0.08 bis -0.64 kJ / (mol $Å^2$) (Kellis *et al*., 1989; Matsumura *et al*., 1988; Serrano *et al*., 1992; Shortle *et al*., 1990), wobei die Mehrzahl der Werte auch hier größer als die aus Löslichkeitsstudien sind.

Hydrophobe Wechselwirkungen werden auch als Haupttriebkraft für konformative Änderungen des Rezeptors bei der Ligandbindung angesehen. Diese induzierte Anpassung kann man sich als „Zusammenklappen“ des Rezeptors um den Liganden vorstellen (Davis & Teague, 1999). So bedingt die Bindung von Trifluoperazin (**1**) an $Ca^{2+}$-Calmodulin eine Konformationsänderung des Proteins von einer verlängerten Hantel zu einer kompakten Form (Vandonselaar *et al*., 1994) und Kristallstrukturanalysen von 3,4,5-substituierten Piperidinderivaten (wie etwa **2**) gebunden an Renin ergaben eine induzierte Anpassung der Rezeptortasche zur Unterbringung der Substituenten (Märki *et al*., 1997).



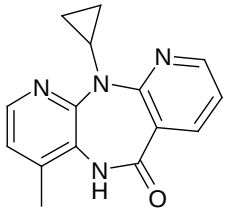

**1**



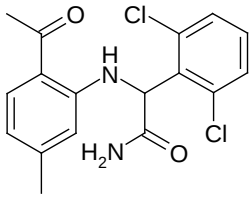

**2**

Diese Anpassung kann auch das Auftreten günstiger hydrophober Wechselwirkungen bei Wirkstoffen unterschiedlicher Gestalt in derselben Bindetasche eines Rezeptors ermöglichen. Ein eindrucksvolles Beispiel stellt die Bindung zwischen HIV-1-Reverse-Transkriptase und

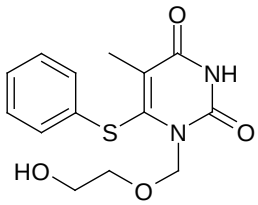
den Inhibitoren Nevirapin (**3**), α-APA (**4**) und HEPT (**5**) (Ren *et al*., 1995) dar: während beim Vergleich der Strukturen keine der C-α-Atompositionen der Proteine um mehr als 2.7 Å verschoben werden, lassen sich die beobachteten Änderungen als Folge einer Anpassung an die strukturellen Variationen der Inhibitorsubstituenten erklären (Davis & Teague, 1999). Da im Fall der HIV-1-Reverse-Transkriptase jedoch auch bei der RNA-Bindung eine deutliche Konformationsänderung zu beobachten ist, läßt sich hier vermuten, daß durch die Anwesenheit der Inhibitoren jeweils unterschiedliche „Stufen" dieser Umlagerung ausgebildet werden.




**3**




**4**




**5**

### 2.2.3 Beiträge durch *intra*molekulare Veränderungen bei Ligand und Rezeptor

Bei der Bildung eines Protein-Ligand-Komplexes kommt es zu Veränderungen der Freiheitsgrade der daran beteiligten Komponenten, die zu einer mit diesem Vorgang verbundenen Entropieänderung führen (Brady & Sharp, 1997b). Sieht man die Komplexbildung – ohne Berücksichtigung des Wassers – als bimolekulare Assoziationsreaktion an, so gehen je drei Freiheitsgrade der Translation und Rotation verloren, während sechs neue Schwingungszustände entstehen (Page, 1977; Page, 1973; Page & Jencks, 1971; Sturtevant, 1977). Obwohl die Separierbarkeit der Standardentropie in einzelne Beiträge für Vorgänge in Lösung formal nicht möglich ist (Dunitz, 1995), bildet dieses Vorgehen doch den Einstieg zum Verständnis des Einflusses von Flexibilität und Mobilität von Protein und Ligand auf die Komplexbildung (Brady & Sharp, 1997b).

Unter Verwendung des Ansatzes von Sackur-Tetrode bzw. der Trouton'schen Regel (Atkins, 1990) und der nicht unkritischen (Murphy *et al*., 1994) Annahme der Übertragbarkeit der Ergebnisse auf Vorgänge in Lösung (Gilson *et al.*, 1997; Janin, 1996) ergibt sich für den Entropieverlust durch vollständige Einschränkung von Translation und Rotation eines Moleküles ein Wert von etwa –420 J / (mol K) (Janin, 1995). Unter Berücksichtigung der Restmobilität der Moleküle im Komplex - abgeschätzt aus den experimentell beobachteten Bewegungen von Lysozym- (Doucet & Benoit, 1987) bzw. Insulinmolekülen (Caspar *et al*., 1988) in

ihren Kristallen – bzw. aus Entropieänderungen bei inter- und intramolekularen Reaktionen wurden allerdings nur etwa halb so große Entropieverluste (≈ -200 J / (mol K) (Finkelstein & Janin, 1989; Page, 1977; Page, 1973; Page & Jencks, 1971) entspr. 60 kJ / mol bei 298 K) ermittelt. Noch kleinere Beiträge wurden von Williams and Searle mit Werten von 9 – 45 kJ / mol aus Schmelz- und Sublimationsstudien an Kohlenwasserstoffen und polaren organischen Molekülen gefunden (Searle & Williams, 1992), in Übereinstimmung mit Ergebnissen aus Assoziationsstudien rigider zyklischer Dipeptide in fester, flüssiger und gasförmiger Phase (Brady & Sharp, 1997a).

Bei der Bindung des Liganden an den Rezeptor kommt es darüber hinaus noch zur Einschränkung der konformativen Beweglichkeit, d.h. interner Freiheitsgrade um drehbare Bindungen. Der Beitrag zur Freien Bindungsenthalpie durch Entropieerniedrigung aufgrund Einschränkung eines Rotors wurde für 298 K allgemein zu 0.5 kJ / mol (Hossain & Schneider, 1999), 2.5 kJ / mol (Finkelstein & Janin, 1989; Searle & Williams, 1992) bzw. 4 – 6 kJ / mol (Page, 1977; Page, 1973; Page & Jencks, 1971) ermittelt. Im Fall von Aminosäuren wurden Wahrscheinlichkeiten für den jeweiligen rotameren Zustand aus beobachteten Verteilungen Lösemittel-exponierter Seitenketten in Proteinkristallen bestimmt und zur Abschätzung des Entropieverlustes bei Einschränkung ihrer konformativen Beweglichkeit verwendet (Doig & Sternberg, 1995; Pickett & Sternberg, 1993). Dabei ergaben sich Beiträge von 0 (für Ala, Gly, Pro) bis 8.7 kJ / mol (für Gln) bei einem Mittelwert von 3.7 kJ / mol pro Rest zur Freien Enthalpie.

Eine Reihe von experimentellen Funden zeigt, daß Liganden tatsächlich oft einen beträchtlichen Teil ihrer Mobilität im an den Rezeptor gebundenen Zustand behalten bzw. eine *Zunahme* der Beweglichkeit auf der Proteinseite sogar die Freie Bindungsenthalpie günstig beeinflussen kann (Forman-Kay, 1999). Ersteres zeigt sich z.B. bei der Bindung von Camphan, Adamantan bzw. Thiocampher an Cytochrom $P450_{cam}$ (Raag & Poulos, 1991). Ohne eine Wasserstoffbrücke auszubilden, rotieren die Liganden frei in der Bindetasche und werden daher nicht-stereoselektiv hydroxyliert. Ein anderes Beispiel dafür, daß Komplexbildung nicht ausschließlich mit Einschränkung molekularer Beweglichkeit einhergeht, ergibt sich bei der Bindung von DNA an die C-terminale Domäne von Topoisomerase I (Yu *et al*., 1996). Während ein Teil der Proteinreste hierbei eine höhere Ordnung aufweisen, werden andere hingegen stärker beweglich. Im Fall der Bindung eines hydrophoben Mauspheromons an *mouse major urinary* Protein konnte mit NMR-Relaxationsstudien sogar gezeigt werden (Zidek *et al*., 1999), daß eine Zunahme der Proteinrückgrats-Entropie einen beträchtlichen *günstigen*

Beitrag zur Freien Bindungsenthalpie liefert, der in der Größenordnung anderer Beiträge liegt. Analog fanden Weber *et al.* (Weber *et al.*, 1994; Weber *et al.*, 1992) bei kristallographischen und thermodynamischen Studien natürlicher und synthetischer Liganden gebunden an Streptavidin, daß das Molekül mit der höchsten Bindungsaffinität zugleich die größte Beweglichkeit im Komplex aufwies.

Ein alternativer Weg, den ungünstigen Einfluß der Einschränkung der Beweglichkeit des Liganden bei der Komplexbildung zu minimieren, besteht darin, seine konformative Vororientierung in Lösung zu erreichen. Dies ist z. B. im Fall des Thrombin-Inhibitors D-Phe-Pro-boro-Arg gelungen (Lim *et al.*, 1993), bei dem ein „hydrophober Kollaps“ (Wiley & Rich, 1993) zur Minimierung der hydrophoben Oberfläche der Seitenketten von D-Phe und Pro und somit zur Ausbildung einer Konformation führt, die der rezeptorgebundenen stark ähnelt. Interessanterweise wird ein inverser, „hydrophiler Kollaps“ der Immunsuppressoren CsA und FK506 und eine daraus resultierende Präorganisation als Grund für deren gute Permeabilität durch biologische Membranen angesehen und durch eine Formulierung der Darreichungsform mit Olivenöl unterstützt (Navia & Chaturvedi, 1996).

Neben entropischen haben auch enthalpische Unterschiede zwischen Lösungs- und rezeptorgebundenen Konformationen des Liganden einen Einfluß auf die Freie Bindungsenthalpie. Vergleiche zwischen Kraftfeldenergien von Konformationen proteingebundener Liganden mit denen der Konformationen des globalen Minimums ergaben Differenzen zwischen 0 und 167 kJ / mol bei 33 untersuchten Verbindungen und einen Mittelwert von 67 kJ / mol (Nicklaus *et al.*, 1995) bzw. ungünstige konformative Energien rezeptorgebundener Konformationen von 112 bis 296 kJ / mol für drei Dihydrofolatreduktase-Inhibitoren (Spark *et al.*, 1982). Diese hohen Werte sind allerdings darauf zurückzuführen, daß die rezeptorgebundenen Geometrien mit solchen aus Gasphasenensemblen verglichen werden. Bei analogen Studien unter Verwendung von mit einem Solvatationsmodell erzeugten Konformationsensemblen zeigt sich denn auch, daß die proteingebundene Geometrie hinsichtlich ihrer Konformationsenthalpie weniger als 12 kJ / mol ungünstiger ist (Bostrom *et al.*, 1998) bzw. daß die Position von „Ankerpunkten“ bei proteingebundenen Ligandkonformationen mit denen bei Minimumenergie-Strukturen festgestellten übereinstimmt (Vieth *et al.*, 1998a). Auch konnten in mehreren Fällen Konformationen für Liganden gefunden werden, die nur geringfügig von denen in der Kristallstruktur abwichen, aber dennoch eine wesentlich geringere Konformationsenergie aufwiesen. Zusätzlich ist zu bedenken, daß die anisotrope molekulare Umgebung des

Proteins die Energiebarrieren zwischen verschiedenen konformativen (Rotations-)Zuständen beeinflussen kann; ein Effekt, der in den verwendeten Kraftfeldern nicht berücksichtigt ist.

### 2.2.4 Additivität, Kooperativität und Enthalpie-Entropie-Kompensation

Für das Verständnis von Protein-Ligand-Wechselwirkungen und ihrer Vorhersage werden oft Ansätze basierend auf Gruppenadditivitäten (Andrews *et al.*, 1984) (Gl. 3) bzw. der Additivität Freier-Enthalpie-Komponenten verwendet (Lau & Pettitt, 1989) (Gl. 4).

$$\Delta G = \Delta G_{CH_3} + \Delta G_{OH} + \Delta G_{Phenyl} + ... \quad \textbf{Gl. 3}$$

$$\Delta G = \Delta G_{H-Brücke} + \Delta G_{Solvatation} + \Delta G_{Konformation} + ... \quad \textbf{Gl. 4}$$

Schon die Varianz der Zahlenwerte der in den vorherigen Kapiteln diskutierten Beiträge macht deutlich, daß dieses Vorgehen nicht ohne weiteres möglich ist. Eine strikte Betrachtung im Rahmen der statistischen Thermodynamik zeigt (Mark & van Gunsteren, 1994), daß die Freie Energie (Freie Enthalpie) eine globale Eigenschaft eines betrachteten Systems und als solche vom gesamten Konfigurations- bzw. Phasenraum des Systems abhängt. Während es also möglich ist, in guter erster Näherung die Energie eines Systems in (paarweise) Einzelbeiträge zu separieren, gilt dies *prinzipiell* nicht für die Entropie (Brady & Sharp, 1997b) sowie die Freie Energie (Freie Enthalpie) (Ben-Naim, 1997). Die Freie Energie (Freie Enthalpie) ist zwar als Zustandsfunktion wegunabhängig, dies gilt jedoch nicht für ihre Komponenten. Beispiele der Nichtadditivität aus Mutationstudien belegen dies (Ackers & Smith, 1985; Horovitz, 1987; Otzen & Fersht, 1999; Wells, 1990). Eine Zerlegung in einzelne Komponenten ist jedoch dann möglich, wenn das betrachtete Gesamtsystem sich in voneinander *unabhängige* Subsysteme zerlegen läßt (Mark & van Gunsteren, 1994). Letzteres ist allerdings für biologische Systeme mit schwachen, nicht-kovalenten Wechselwirkungen und daher wenig voneinander verschiedenen (makroskopischen) Zuständen fraglich (Dill, 1997).

Eine Alternative besteht darin, eine Zerlegung nur für den energiedominierten Teil der Freien Energie (Freien Enthalpie) vorzunehmen (Gl. 5) (Ben-Naim, 1997; Brady & Sharp, 1995; Dill, 1997):

$$\Delta G = \Delta H_{H-Brücke} + \Delta H_{Solvatation} + \Delta H_{Konformation} + ... + T\Delta S \quad \textbf{Gl. 5}$$

Dieses Vorgehen wurde z.B. zur Gewinnung „intrinsischer Bindungsenergien“ aus Freien Enthalpien der Bindung von Molekülen mit den Gruppen A, B bzw. A+B an ein Protein angewendet (Jencks, 1981).

Ein deutliches Beispiel für Nicht-Additivität – an anderer Stelle (Ackers *et al*., 1992) als „Kooperativität“ bezeichnet – zeigt sich beim Versuch der Korrelation der „hydrophoben Freien Enthalpie“ mit der dem Lösemittel nicht zugänglichen unpolaren Oberfläche. Ergebnisse aus Mutationsstudien an Proteinen und Studien zur Ligandbindung zeigen, daß der hydrophobe Effekt die Stabilität bzw. Bindung in wäßriger Lösung augenscheinlich stärker fördert, als es in Lösungsmittel-Transfer-Messungen bestimmt wurde (*vide supra*) (Williams & Bardsley, 1999). Es konnte jedoch gezeigt werden, daß das Vergraben eines hydrophoben Molekülteils in einer molekularen Erkennungsstelle zur gleichzeitigen, kooperativen Verstärkung benachbarter elektrostatischer Wechselwirkungen führen kann (Sharman *et al*., 1995; Williams *et al*., 1998).

Der Anteil der Standardenthalpie $\Delta H^0$ und Standardentropie $\Delta S^0$ an der Freien (Bindungs-)Enthalpie $\Delta G^0$ (Gl. 1) kann direkt aus mikrokalorimetrischen Messungen (Wieseman *et al*., 1989) oder über van’t Hoff-Auftragungen von Affinitätsmessungen bei unterschiedlichen Temperaturen (Hitzemann, 1988) bestimmt werden. Hierbei ergibt sich i.a. keine Korrelation zwischen $\Delta H^0$ und $\Delta G^0$, d.h. eine Interpretation und Vorhersage von Bindungseigenschaften eines Liganden an ein Protein allein auf enthalpischer Grundlage ist unzureichend (Böhm & Klebe, 1996). Eine Ausnahme besteht aber für Serien ähnlicher Liganden, innerhalb derer der entropische Anteil als nahezu konstant angesehen werden kann. Im Gegensatz dazu zeigt sich allerdings eine deutliche Korrelation zwischen $\Delta H^0$ und $\Delta S^0$. Diese „Enthalpie-Entropie-Kompensation“ ist eine intrinsische Eigenschaft schwacher intermolekularer Wechselwirkungen (Dunitz, 1995; Williams & Westwell, 1998) und wird allgemein beobachtet bei (supramolekularen) Wirt-Gast- (de Namor *et al*., 1991) und Rezeptor-Ligand-Komplexen (Gilli *et al*., 1994; Grunwald & Steel, 1995). Sie kann dahingehend interpretiert werden, daß eine Verstärkung einer intermolekularen Bindung gleichzeitig zu einem Verlust an Freiheitsgraden der Bewegung führt und umgekehrt. Ihr Auftreten ist von besonderer Bedeutung für die Vorhersage von Rezeptor-Ligand-Wechselwirkungen: während der enthalpische sowie entropische Anteil allein beträchtliche Größen annehmen kann, liegt der Wert der Freien Enthalpie häufig nahe bei Null. Schon kleine *relative* Fehler bei der Vorhersage von $\Delta H^0$ und $\Delta S^0$ können daher einen deutlichen Einfluß auf $\Delta G^0$ haben.

# 3 Ansätze zur Vorhersage von Bindungsaffinitäten aus der Literatur

Die Arbeiten zur Vorhersage von Bindungsaffinitäten lassen sich in zwei Hauptkategorien einteilen:

- Ist auf der einen Seite die *3D-Struktur des biologischen Zielmoleküles nicht bekannt*, beruht die (oft qualitative) Vorhersage der Bindungsaffinität neuer Liganden auf dem Vergleich mit bekannten Referenzstrukturen – etwa endogenen Liganden oder bereits synthetisierten Verbindungen (Klebe, 1998a; Kubinyi, 1993; Kubinyi, 1997; Kubinyi *et al.*, 1997).
- Andererseits erfolgt bei *Kenntnis der 3D-Struktur des Rezeptors* die Vorhersage der Bindungsaffinität aufgrund der geometrischen und chemischen Komplementarität der in das biologische Zielmolekül eingelagerten Liganden (Ajay & Murcko, 1995; Gilson *et al.*, 1997; Hirst, 1998; Knegtel & Grootenhuis, 1998; Kollman, 1994; McCammon, 1998; Oprea & Marshall, 1998; Reddy *et al*., 1998; Tame, 1999).

Der Schwerpunkt im folgenden Überblick liegt bei den Methoden unter Verwendung der Kenntnis der 3D-Struktur des Rezeptors. Nach einem Vergleich der Ansätze hinsichtlich Anwendbarkeit und Qualität der erzielten Ergebnisse erfolgt abschließend ein kurzer Überblick über bestehende Verfahren zur Generierung relativer Anordnungen von Protein und Ligand als Ausgangspunkt für Affinitätsvorhersagen neuer Moleküle im Rahmen des strukturbasierten Designs.

## *3.1 Ansätze ohne Kenntnis der Rezeptorstruktur*

Ausgangspunkt der Vorhersage der Bindungsaffinität eines Liganden ohne Kenntnis der Rezeptorstruktur ist die Annahme, daß sich biologische Ähnlichkeit in der molekularen (chemischen) Ähnlichkeit miteinander zu vergleichender niedermolekularer Verbindungen widerspiegelt (Klebe, 1998a). Einen Überblick über Definition, Berechnung sowie Anwendung molekularer Ähnlichkeit allgemein und im Rahmen des Wirkstoffdesigns geben Johnson und Maggiora (Johnson & Maggiora, 1990), Rouvray (Rouvray, 1995) und Dean (Dean, 1995). Auf Ansätze, die Moleküle auf eindimensionaler bzw. zweidimensionaler Basis (Brown & Martin, 1996) – etwa über die An- / Abwesenheit funktioneller Gruppen oder über Bitvektorrepräsentationen topologischer Deskriptoren (sog. *fingerprints*) – miteinander vergleichen, soll hier genauso wenig eingegangen werden wie auf 3D-Verfahren basierend auf Substruk-

turvergleichen (Humblet & Dunbar, 1993), Pharmakophorsuchen (Willett, 1995) und Überlagerungen niedermolekularer Verbindungen (Bures, 1997; Klebe, 1993). Diese Methoden können oft nur ein qualitatives Maß für die zu erwartende Bindungsaffinität liefern.

Quantitative Vorhersagen lassen sich dagegen aus ***Q**uantitative-**S**truktur-**A**ktivitäts-**B**eziehungs(QSAR)*-Ansätzen gewinnen (Kubinyi, 1993). Auf Basis physikochemischer bzw. struktureller Parameter wird eine Relation zwischen der Struktur einer Verbindung und ihrer Wirkung (im biologischen Sinn etwa Affinität und Selektivität) aufgestellt. Die klassischen *2D*-QSAR-Verfahren, beruhend auf den Arbeiten von Hansch und Fujita (Hansch & Fujita, 1964) bzw. Free und Wilson (Free & Wilson, 1964), weisen die Nachteile auf, daß nur Datensätze strukturell ähnlicher Moleküle untersucht werden können und daß für das Verständnis von Rezeptor-Ligand-Wechselwirkungen essentielle 3D-Strukturinformationen nur untergeordnet bzw. indirekt berücksichtigt werden (Kim, 1993).

Dieser letzte Punkt wird durch Anwendung von *3D*-QSAR-Verfahren (Greco *et al*., 1998; Oprea & Waller, 1997) umgangen: aus der räumlichen Struktur von Liganden werden relative Unterschiede der einzelnen Mitglieder eines Datensatzes ermittelt und mit bekannten Eigenschaften der Verbindungen, etwa der Bindungsaffinität zu einem Rezeptor, korreliert. Voraussetzung dafür ist aber, daß eine bioaktive Konformation für jeden Liganden bestimmt werden kann und diese Konformationen relativ zueinander so angeordnet werden, wie es für ihre Anordnung in der Bindetasche zu erwarten ist (Klebe, 1993). Ausgangspunkt für diese Überlagerungen können rigide Mitglieder des zu untersuchenden Datensatzes (Cramer III *et al*., 1988), Konformationen aus bekannten Protein-Ligand-Kristallstrukturen (Waller *et al*., 1993) oder einen Pharmakophor bildende funktionelle Gruppen sein (Marshall *et al*., 1979; Martin *et al*., 1993). Neben Atom- bzw. Gruppen-basierten Überlagerungsverfahren wird v.a. auch die (flexible) Alignierung von Molekülen durch Maximierung der Ähnlichkeit ihrer molekularen Felder angewendet (Klebe, 1993; Lemmen & Lengauer, 2000).

Im folgenden soll eine kurze Beschreibung gängiger 3D-QSAR-Verfahren gegeben werden; der Schwerpunkt wird auf der prototypischen CoMFA-Methode und davon ausgehenden Erweiterungen liegen. Vollständige Übersichten über dieses Gebiet bieten Bücher herausgegeben von Kubinyi (Kubinyi, 1993; Kubinyi *et al.*, 1997), Sanz (Sanz *et al*., 1995) und van de Waterbeemd (van de Waterbeemd, 1995).

- Das aus dem Vorläufer DYLOMMS (Cramer III & Bunce, 1987) hervorgegangene dreidimensionale, gitterbasierte CoMFA-Verfahren (***Co**mparative **M**olecular **F**ield **A**nalysis*) (Cramer III *et al.*, 1988) vergleicht molekulare Energiefelder einer Serie von

Molekülen und korreliert Unterschiede darin mit Unterschieden in abhängigen Zielgrößen, hier den Bindungsaffinitäten. Unter der Annahme, daß entropische Beiträge dazu innerhalb des betrachteten Datensatzes als konstant angesehen werden können, werden im klassischen Verfahren für jedes Molekül für die Punkte eines die gesamten Moleküle umschließenden Gitterkastens sterische und elektrostatische Wechselwirkungsenergien berechnet. Standardmäßig wird ein 6-12 Lennard-Jones-Potential und ein Coulomb-Potential sowie ein $sp^3$-hybridisiertes Kohlenstoff-Atom mit der Ladung +1, jeweils als Sonde lokalisiert an den Gitterpunkten, verwendet. Für jedes Molekül *n* ergibt sich damit die QSAR-Gleichung zu

$$Affinität_n = k + \alpha_1 S_{n,1} + ... + \alpha_M S_{n,M} + \beta_1 E_{n,1} + ... + \beta_M E_{n,M} \qquad \textbf{Gl. 6},$$

wobei die Indizes 1, 2, ..., *M* für den jeweiligen Gitterpunkt und $S_{n,1}$, ..., $S_{n,M}$ sowie $E_{n,1}$, ..., $E_{n,M}$ für die jeweilige sterische bzw. elektrostatische Energie an diesem Punkt stehen. Die Koeffizienten $\alpha_1$, ..., $\alpha_M$ sowie $\beta_1$, ..., $\beta_M$ werden aus dem gesamten, unterbestimmten linearen Gleichungssystem durch Anwendung einer PLS-Analyse (***P**artial **L**east **S**quares*) (Geladi & Kowalski, 1986; Wold *et al.*, 1993) erhalten. Mit dem so erhaltenen Modell, dessen Vorhersagekraft für die abhängige Größe durch Kreuzvalidierungs- oder *Bootstrapping*-Verfahren (Weisberg, 1985) getestet wurde, können nun Bindungsaffinitäten für neue, im Trainingsdatensatz nicht enthaltene Moleküle berechnet werden.

Während die sterischen und elektrostatischen molekularen Felder rein enthalpische Beiträge beschreiben, wird durch Charakterisierung hydrophober Eigenschaften der Moleküle versucht, entropische Beiträge zu berücksichtigen (Folkers & Merz, 1996). Hierunter fallen die Verwendung von Feldern basierend auf dem HINT(***H**ydrophobic **Int**eraction*)-Ansatz (Kellog & Abraham, 1991), molekulare Lipophilie-Potentiale (MLP) (Carrupt *et al.*, 1997), die durch das Programm GRID (Goodford, 1985) mit der $H_2O$- bzw. DRY-Probe erzeugten Felder sowie mit dem DelPhi-Programm (DelPhi/Solvation, 1995) berechnete Desolvatationsenergie-Felder.

- Ein alternativer Ansatz zur Berechnung molekularer Wechselwirkungsfelder wird in dem CoMSIA-Verfahren (***Co**mparative **M**olecular **S**imilarity **I**ndices **A**nalysis*) verfolgt (Klebe *et al.*, 1994). Bedingt durch die $r^{-n}$-Abhängigkeit des Lennard-Jones- ($n$ = 12) bzw. Coulomb-Potentials ($n$ = 1) kommt es beim klassischen CoMFA-Ansatz am Ort der Ligandatome ($r \rightarrow 0$) zu Singularitäten bzw. in der Nähe der Moleküloberfläche zu sehr steilen Potentialverläufen. Dieser Nachteil wird durch die Verwendung sterischer, elektrostatischer und hydrophober Ähnlichkeitsfunktionen sowie zusätzlicher

Funktionen zur Beschreibung von Wasserstoffbrücken-Donor- und –Akzeptor-Eigenschaften (Klebe, 1994; Klebe & Abraham, 1999) umgangen, die einen abstandsabhängigen Verlauf zwischen Molekülatom und Sondenatom aufweisen, der einer Gaussfunktion ($\exp\{-\alpha r^2\}$) entspricht. Damit können auch Ähnlichkeitsindizes in unmittelbarer Nähe von Molekülatomen beschrieben werden. Durch den insgesamt flacheren Potentialverlauf wird das Verfahren auch unempfindlicher gegenüber kleinen Variationen bei der Überlagerung der Moleküle bzw. Rotationen und Translationen des verwendeten Gitters (Böhm *et al*., 1999).

- Der HASL-Ansatz (***H**ypothetical **A**ctive **S**ite **L**attice*) (Doweyko, 1988) ist ebenfalls eine gitterbasierte 3D-QSAR-Technik, bei der allerdings partielle Aktivitäten auf Gitterpunkten innerhalb des van der Waals-Volumens der verwendeten Liganden verteilt werden. Die Summe der Werte an allen Punkten, die zu einem Molekül gehören, entspricht dann der zu korrelierenden Größe.
- Bei der Compass-Technik (Jain *et al*., 1994) werden molekulare Wechselwirkungsfelder nur in der Nähe der van der Waals-Oberfläche der jeweiligen Moleküle berechnet, womit eine Fokussierung auf den Bereich des (hypothetischen) Rezeptor-Ligand-Bindungsepitops erreicht wird. Zudem wird die Anzahl der verwendeten Deskriptoren stark reduziert. Aus ihnen wird dann unter Verwendung eines neuronalen *back-propagation* Netzes ein QSAR-Modell erzeugt. Dieses Modell kann weiterhin durch einen automatischen Prozeß der iterativen Konformationserzeugung und Überlagerung der verwendeten Verbindungen verbessert werden.
- Im Gegensatz zu den bisherigen Verfahren werden beim Expertensystem APEX-3D (Golender & Vorpagel, 1993) keine molekularen Wechselwirkungsfelder berechnet. Statt dessen wird versucht, den komplexen Mustererkennungsprozeß zur Aufstellung einer Beziehung zwischen strukturellen Moleküleigenschaften und beobachteter Aktivität in einzelne, automatisierte Schritte zu zerlegen. Moleküle mit ähnlicher Aktivität werden zunächst nach ähnlichen 2D-topologischen bzw. 3D-topographischen Mustern untersucht. Unter Verwendung logischer Programmierung werden damit Pharmakophore identifiziert, die als Basis zur Überlagerung der Moleküle verwendet werden. Abschließend wird ein 3D-QSAR-Modell aus physikochemischen Eigenschaften der (biophoren) Gruppen sowie globalen molekularen Eigenschaften wie Hydrophobizität und Molrefraktion erstellt.
- Die Methode YAK (Vedani *et al*., 1995) geht auf Ansätze von Höltje und Kier (Höltje & Kier, 1974) zurück, bei denen mit einem Satz von Liganden wechselwirkende Sei-

tenketten von Aminosäuren explizit in Form eines sogenannten Pseudorezeptors modeliert werden. YAK selektiert und positioniert diese Rezeptorseitenketten automatisch, wobei kristallographische Zusatzinformationen, Daten aus Sequenzanalysen homologer Proteine oder aus Mutationsstudien einfließen können. Die Auswahl und Positionierung hängt zudem davon ab, ob zwischen den Aminosäuren und den Liganden berechnete Wechselwirkungsenergien mit den Affinitätsdaten korrelieren. Aus den zunächst generierten, zu funktionellen Gruppen der Liganden komplementären Aminosäuren wird durch Verbrückung mit Poly-Gly-Fragmenten ein Pseudorezeptor konstruiert.

### *3.2 Ansätze mit Kenntnis der Rezeptorstruktur*

Der Erfolg von Docking- und *de-novo*-Design-Verfahren hängt maßgeblich von der Bewertung der erhaltenen geometrischen Orientierungen zwischen Ligand und Protein hinsichtlich der zu erwartenden Bindungsaffinität ab (Joseph-McCarthy, 1999; Knegtel & Grootenhuis, 1998; Oprea & Marshall, 1998). Die Grundlage der statistischen Thermodynamik zur Berechnung der Bindungsaffinitäten wird von Gilson *et al.* (Gilson *et al.*, 1997) und Ajay & Murcko (Ajay & Murcko, 1995) kritisch zusammengefaßt. Übersichten über angewendete Verfahren im allgemeinen werden gegeben von Tame (Tame, 1999), McCammon (McCammon, 1998), Reddy *et al.* (Reddy *et al.*, 1998), Knegtel und Grootenhuis (Knegtel & Grootenhuis, 1998), Oprea und Marshall (Oprea & Marshall, 1998) und Bamborough und Cohen (Bamborough & Cohen, 1996). Die Berechnung molekularer Wechselwirkungsfelder auf Grundlage der bekannten Rezeptorstruktur beschreibt Wade (Wade, 1998) im Überblick, während Weber und Harrison (Weber & Harrison, 1998) sowie Liljefors (Liljefors, 1998) im speziellen Anwendungen und Fortschritte bei kraftfeldbasierten Berechnungen behandeln. Einen Überblick über die rechnerische Behandlung der Elektrostatik bei makromolekularen Systemen wird von Honig und Nicholls (Honig & Nicholls, 1995), Warshel und Papazyan (Warshel & Papazyan, 1998), Gilson (Gilson, 1995) und Schaefer *et al.* (Schaefer *et al*., 1998) gegeben. Kollman (Kollman, 1993; Kollman, 1996), Kollman und Merz (Kollman & Merz Jr., 1990), Jorgensen (Jorgensen, 1989), Straatsma (Straatsma, 1996) und Meirovitch (Meirovitch, 1998) fassen die Berechnung von Freier Enthalpie und Entropie im Rahmen thermodynamischer Störungsrechnungen zusammen.

Im folgenden werden die Verfahren hinsichtlich ihrer methodischen Grundlagen klassifiziert und beschrieben. Allerdings sind die Grenzen dabei nicht immer scharf einzuhalten, ei-

nige Verfahren kombinieren außerdem mehrere Ansätze. In diesen Fällen folgt die Einteilung dem zentralen methodischen Element.

### 3.2.1 Freie-Energie-Störungsrechnungen und Lineare-Freie-Energie-Ansätze

Der vom thermodynamischen Standpunkt einzige korrekte Weg zur Vorhersage relativer Freier *Energien* der Bindung von Liganden an Proteine beruht auf der Anwendung von Freie-Energie-Störungsrechnungen (FEP, ***f**ree **e**nergy **p**erturbation*) (Gl. 7) (Kirkwood, 1935) bzw. den Verfahren der **t**hermodynamischen **I**ntegration (TI) (Gl. 8) (Zwanzig, 1954) unter expliziter Berücksichtigung von Solvensmolekülen und der Flexibilität von Rezeptor und Ligand. (Ajay & Murcko, 1995; Joseph-McCarthy, 1999; Tame, 1999).

$$\Delta F = F_1 - F_0 = -k_B T \ln\left\langle \exp\left( -\frac{H_1(\vec{X}) - H_0(\vec{X})}{k_B T} \right) \right\rangle_0 \qquad \textbf{Gl. 7}$$

$$\Delta F = F_1 - F_0 = \int_0^1 \left\langle \frac{\partial H_\lambda(\vec{X})}{\partial \lambda} \right\rangle_\lambda d\lambda \qquad \textbf{Gl. 8}$$

Grundlage hierfür ist die Beziehung zwischen der *Freien* Energie eines betrachteten Systems *F* und dem Ensemblemittelwert einer Energiefunktion, die dieses System beschreibt (Postma *et al.*, 1981; Straatsma, 1996; Tembe & McCammon, 1984). $H_\lambda(\vec{X})$ steht für einen von den Koordinaten ( $\vec{X}$ ) der Partikel im Phasenraum und einem Kopplungsparameter $\lambda$ abhängigen Hamiltonian des Systems, $k_B$ für den Boltzmann-Faktor, T für die absolute Temperatur. Die Indizes „0“ und „1“ stehen für $\lambda = 0$ bzw. $\lambda = 1$. $\langle ... \rangle$ beschreibt eine Mittelwertbildung. Die Konfigurationsensembles eines Systems können dabei durch Monte-Carlo(MC)- (Metropolis *et al.*, 1953) und Molekulardynamik(MD)-Simulationen (van Gunsteren & Berendsen, 1990) erhalten werden. Da die Differenz zwischen Freier *Enthalpie* und Freier *Energie*, das Produkt aus Druck und Änderung des Volumens während einer isothermen und isobaren Reaktion, für Vorgänge in Lösung vernachlässigbar ist, ergibt sich auf diesem Weg auch ein Zugang zu Freien Enthalpien.

Obwohl die Methode geeignet ist, einzelne Beiträge zur Freien Energie / Enthalpie auf atomarer Ebene bzw. auf der Ebene einzelner Subsysteme (z.B. des Liganden oder des Proteins) zu untersuchen (Boresch & Karplus, 1995; Gao *et al.*, 1989), stehen dem gegenüber jedoch Probleme hinsichtlich einer allgemeinen Anwendbarkeit bedingt durch die Frage der ausreichenden Durchmusterung des Phasenraumes, der Genauigkeit der verwendeten Hamiltonians (Kraftfelder) sowie der Abhängigkeit der Ergebnisse vom durchgeführten Berech-

nungsprotokoll (Beveridge & DiCapua, 1989; Kollman, 1993; Kollman, 1996). Dazu kommen lange Rechenzeiten sowie die Beschränkung auf lediglich kleine Veränderungen z.B. bei Liganden für die Vorhersage verläßlicher relativer Freier Energien / Enthalpien (Pearlman, 1994; Sen & Nilsson, 1999). Einige (klassische) Anwendungsbeispiele sowie neuere Ansätze sollen im folgenden aufgezeigt werden.

- Die auf Arbeiten von Postma *et al.* (Postma *et al.*, 1981) und Jorgensen und Ravimohan (Jorgensen & Ravimohan, 1985) zurückgehenden FEP-MD- und FEP-MC-Ansätze wurden u.a. zur Vorhersage relativer Freier Energiedifferenzen der Bindung von Benzamidin bzw. *p*-Fluorbenzamidin und Glycin bzw. Alanin an Trypsin (Wong & McCammon, 1986), von Folat-basierten Inhibitoren an Thymidylatsynthase (Reddy *et al*., 1992) und von Thermolysininhibitoren (Bash *et al*., 1987) verwendet. Im letzteren Fall wurde dabei eine bemerkenswerte Übereinstimmung zwischen *vorhergesagten* berechneten und anschließend experimentell ermittelten Werten erzielt (Merz & Kollman, 1989), die allerdings auf sehr kleinen strukturellen Veränderungen bei den untersuchten Liganden (Austausch von NH gegen O bzw. $CH_2$) beruht. Daß dagegen die Addition eines Phenylringes an ein Inhibitorgerüst selbst bei 400 ps Simulationszeit noch nicht zur vollständigen Konvergenz der berechneten Energien führt und die vorhergesagte relative Freie Energie sogar das falsche Vorzeichen in bezug auf den experimentell ermittelten Wert besitzt, macht die immer noch bestehenden Schwierigkeiten bei diesen Methoden deutlich (McCarrick & Kollman, 1999).
- In einem Ansatz von Ota und Brunger (Ota & Brunger, 1997) wird eine Kombination aus nicht-Boltzmann-bestimmter Durchmusterung des Phasenraumes und TI verwendet (NBTI). Der Vorteil der auch *umbrella sampling* genannten Methode besteht in der Erhöhung der Flexibilität der betrachteten Liganden durch Erniedrigung ihrer Rotationsbarrieren und einer damit einhergehenden verstärkten Durchmusterung des Konfigurationsraumes. Verglichen mit einer (klassischen) TI-Rechnung zeigten sich geringere Abweichungen zwischen berechneten und experimentell bestimmten relativen Freien Energien der Bindung von Benzamidin und Benzylamin an Trypsin (Ota *et al*., 1999).
- Während in (klassischen) Freie-Energie-Rechnungen für *jede* an einem Liganden durchgeführte Modifikation eine separate, zeitlich aufwendige Durchmusterung des Konfigurationsraumes erfolgen muß, wird durch ein Verfahren von Gerber *et al.* (Gerber *et al*., 1993) die Einbeziehung eines ganzen *Satzes* von Modifikationen in eine einzige MD-Simulation erreicht. Unter Annahme einer linearen Separierbarkeit einzelner Beiträge wird für jede spezifische Wechselwirkung des betrachteten Systems deren

Ableitung in Bezug auf den Kopplungsparameter $\lambda$ analytisch bestimmt. So kann aus den Anfangssteigungen der Freien Energiebeiträge bei einem initialen Zustand $\lambda = 0$ auf die Beiträge bei einem finalen Zustand $\lambda = 1$ geschlossen werden. Obwohl die vorgestellte Methode für das System Trimethoprim-basierter Inhibitoren / NADPH / Dihydrofolatreduktase den Rechenaufwand um den Faktor 1 / 1000 reduzierte, ergab sich allerdings keine signifikante Korrelation zwischen berechneten und experimentellen Ergebnissen.

- Guo *et al.* (Guo *et al*., 1998) verwenden einen Ansatz, bei dem der Kopplungsparameter $\lambda$ als dynamische Variable behandelt wird und sich zusammen mit den Atomkoordinaten eines Systems gemäß den Newton'schen Bewegungsgleichungen entwickelt. Damit kann für eine Serie von verwandten Liganden gleichzeitig unter Verwendung eines Satzes von $\lambda$'s relative Freie Energien der Bindung simuliert werden, wobei die verschiedenen Teile der Liganden zwar alle das umgebende Protein „spüren", nicht jedoch jeweils andere Ligandenteile. Der Vorteil ist ein effizienteres Durchmustern des Konfigurationsraumes und ein damit einhergehender reduzierter Rechenaufwand.
- Um die Probleme zu umgehen, die für große, strukturell verschiedene Sätze von Liganden bei der Verwendung (klassischer) Freie-Energie-Verfahren auftreten, wurde von Aquist *et al.* (Aquist *et al*., 1994; Hansson *et al*., 1998) eine semi-empirische Methode zur Berechnung absoluter Freier Energien der Bindung basierend auf der Simulationen zweier physikalischer Zustände entwickelt. Hierbei werden die polaren und nichtpolaren Beiträge zur Freien Energie aus Mittelwerten von MD-Simulationen des Liganden in Wasser bzw. des Protein-Ligand-Komplexes in Wasser linear approximiert. Benötigte Wichtungsparameter werden an Komplexen mit bekannten Bindungsaffinitäten kalibriert. Allerdings ist die Wahl der Energiebeiträge und die Größe ihrer Wichtungsparameter vom betrachteten System und den Simulationsbedingungen abhängig (Carlson & Jorgensen, 1995; Wang *et al*., 1999a), was die allgemeine Anwendbarkeit stark einschränkt (Wall *et al*., 1999).

### 3.2.2 Kraftfeldbasierte Verfahren und Ansätze beruhend auf additiven Freie-Enthalpie-Beiträgen

Die in diesem Kapitel beschriebenen Ansätze beruhen auf der Annahme, daß die Freie Enthalpie der Bindung eines Liganden an einen Rezeptor als Summe einzelner Beiträge dargestellt werden kann (Gl. 4) (bzgl. dieser Annahme s. Kap. 2.2.4) (Ajay & Murcko, 1995;

Gilson *et al.*, 1997; Joseph-McCarthy, 1999). Ausgehend von dieser „Zentralgleichung“ (engl. ***m**aster **e**quation, ME*) werden die einzelnen Terme auf physikochemischer Grundlage formuliert, wobei Kreuzkorrelationen zwischen ihnen zu vermeiden sind. Ein nennenswerter Haupttunterschied zwischen diesem und allen folgenden Kapiteln, verglichen mit dem vorhergehenden, ist außerdem der Ersatz von auf *Ensemble*-Mittelwerten beruhenden (Freien) Energiebeiträgen durch solche, die auf einer *einzigen* zugrundeliegenden Struktur beruhen (Ajay & Murcko, 1995).

- Die Modellierung intermolekularer Wechselwirkungen auf Grundlage Molekularmechanik-basierter Kraftfelder führt im einfachsten Fall der Berechnung für Protein-Ligand-Komplexe *in vacuo* zu einem rein enthalpischen Beitrag zur Freien Enthalpie der Bindung (Böhm & Klebe, 1996; Joseph-McCarthy, 1999). Unter Berücksichtigung von van der Waals- und elektrostatischen Wechselwirkungen sowie z. T. intramolekularer Energiebeiträge ergibt sich – allerdings nur für Serien eng verwandter Liganden, bei denen der entropische Beitrag zur Freien Enthalpie jeweils als konstant angesehen werden kann – eine Korrelation zu experimentell bestimmten Bindungsaffinitäten (Grootenhuis & van Galen, 1995; Grootenhuis & van Helden, 1994; Holloway *et al*., 1995; Joseph-McCarthy *et al*., 1997; Kurinov & Harrison, 1994). In einem Fall konnte dieses ohne explizite Berücksichtigung von Wasser erhaltene Ergebnis durch die Dominanz von van der Waals-Wechselwirkungen sowie lösemittelunabhängige elektrostatische Beiträge begründet werden (Checa *et al*., 1997).
- Ein einfacher Ansatz, den Einfluß des Lösemittels Wasser auf die Rezeptor-Ligand-Bindung in die „Hauptgleichung“ mit einzubeziehen, besteht in der Verwendung atombasierter Solvatations-Parameter (Eisenberg & McLachlan, 1986; Ooi *et al.*, 1987) im Zusammenhang mit der im Komplexbildungsprozeß vergrabenen Oberfläche von Protein und Ligand. Die Verfahren von Vajda *et al.* (Vajda *et al*., 1994), Weng *et al.* (Weng *et al*., 1996), Williams und Mitarbeitern (Searle *et al.*, 1992; Williams *et al*., 1991; Williams *et al.*, 1993), Krystek *et al.* (Krystek *et al*., 1993) und Novotny *et al.* (Novotny *et al*., 1989) berücksichtigen in weiteren Termen außerdem noch die der Bindungsaffinität abträglichen Beiträge bedingt durch die Einschränkung der translatorischen und rotatorischen Freiheitsgrade der Moleküle sowie die Verringerung ihrer intramolekularen Beweglichkeit (s. dazu Kap. 2.2.3). Die intermolekularen Wechselwirkungen werden bei Krystek *et al.*, Vajda *et al.* und Weng *et al.* durch Coulomb-Wechselwirkungen unter Verwendung einer distanzabhängigen Dielektrizitätskonstanten modelliert; Williams und Mitarbeiter verwenden stattdessen intrinsische Bin-

dungsbeiträge funktioneller Gruppen. Vajda *et al.* (Vajda *et al.*, 1994) ermittelt für flexible Liganden zudem noch den Energiebeitrag bedingt durch unterschiedliche *intra*-molekulare Wechselwirkungen des Moleküls in freiem und gebundenem Zustand.

- Der Beitrag elektrostatischer Wechselwirkungen in Gegenwart von Wasser kann im Rahmen einer Näherung des „gemittelten Feldes“ bzw. einer Kontinuumsrepräsentation des Lösemittels durch Lösung der linearisierten Poisson-Boltzmann-Gleichung (Honig & Nicholls, 1995) nach der Methode der finiten Differenzen (Warwicker & Watson, 1982) bzw. der Methode der finiten Elemente (Zauhar & Morgan, 1985) ermittelt werden. Hierbei werden die *polaren* Wechselwirkungsenergien von Rezeptor, Ligand bzw. Rezeptor-Ligand-Komplex untereinander und mit dem umgebenden Solvens ermittelt, indem die Moleküle (versehen mit diskreten Atomladungen) als Bereiche mit niedriger Dielektrizitätkonstante in ein umgebendes Medium mit hoher Dielektrizitätskonstanten eingebettet werden (Warshel & Papazyan, 1998). Der *nichtpolare* Beitrag der Desolvatation wird weiterhin proportional zur Größe der vergrabenen unpolaren Oberfläche beider Moleküle während der Komplexbildung angesetzt; entropische Beiträge durch Einschränkung der Mobilität und Flexibilität werden wie im vorherigen Absatz modelliert. Beispiele für auf diesen Prinzipien beruhende Verfahren zur Vorhersage der Freien Enthalpie der Bindung sind von Froloff *et al.* (Froloff *et al*., 1997), Zhang und Koshland (Zhang & Koshland, 1996), Massova und Kollman (Massova & Kollman, 1999), Hoffmann *et al.* (Hoffmann *et al*., 1999), Polticelli *et al.* (Polticelli *et al*., 1999) und Shoichet *et al.* (Shoichet *et al*., 1999) vorgestellt worden. Zou *et al.* (Zou *et al*., 1999) verwenden anstelle des Poisson-Boltzmann-Ansatzes das „Generalisierte-Born-Modell“ (GB/SA) von Still *et al.* (Still *et al*., 1990) zur Ermittlung der polaren Wechselwirkungsenergien, das eine von der jeweiligen molekularen Umgebung der betrachteten Atome abhängige Dielektrizitätskonstante verwendet.
- Alternativ dazu kann eine implizite Berücksichtigung der Beiträge durch Solvatation und Desolvatation auch direkt in einem Molekülmechanik-Kraftfeld erfolgen (Lazaridis & Karplus, 1999). Hierbei wird die Freie Solvatationsenergie einer funktionellen Gruppe bzw. eines Aminosäurerestes aus der Freien Solvatationsenergie dieser Gruppe in einem kleinen Molekül abzüglich eines Beitrags berechnet, der durch den Ausschluß von Lösemittelmolekülen durch die Anwesenheit anderer Atome im makromolekularen System bedingt ist.

### 3.2.3 Regressionsbasierte Ansätze

Regressionsbasierte Ansätze – auch „empirische Bewertungsfunktionen“ genannt - beruhen wie die im vorherigen Kapitel vorgestellten auf der Annahme der Additivität einzelner Freie-Enthalpie-Beiträge. Allerdings werden die Beiträge der einzelnen Terme einer Regressionsgleichung durch Bestimmung der Koeffizienten vor den unabhängigen (strukturbeschreibenden) Variablen in dieser Gleichung mittels Verfahren der multiplen linearen Regression, *Partial Least Squares*-Regression (Wold *et al.*, 1993) oder durch neuronale Netze (Gasteiger & Zupan, 1993) aus einem Trainingsdatensatz kristallographisch bekannter Rezeptor-Ligand-Komplexe mit experimentell bestimmten Bindungsaffinitäten berechnet. Basierend auf einem heuristischen Ansatz rechtfertigt die Plausibilität der erhaltenen Beiträge sowie ihre Eignung zur Vorhersage unbekannter Bindungsaffinitäten die anfänglich gewählte Separation der Freien Enthalpie. Wie bei allen regressionsbasierten Verfahren hängen die gewonnenen Ergebnisse und ihre Übertragbarkeit auf neue Fälle jedoch auch hier maßgeblich von der Zusammensetzung des Trainingsdatensatzes hinsichtlich der betrachteten Terme ab (Ajay & Murcko, 1995). Weiterhin ist es nur möglich, die Größe der Beiträge für in den experimentellen Daten auftretende (d.h. von sich aus *günstige*) Phänomene zu ermitteln, nicht jedoch für der Bindung abträgliche Beiträge.

- Der Archetyp einer empirischen Bewertungsfunktion für Protein-Ligand-Wechselwirkungen wurde von Böhm (Böhm, 1994) entwickelt (SCORE1). Die Summe aus Beiträgen für Wasserstoffbrücken, ionische Wechselwirkungen, vergrabene unpolare Oberflächenbereiche und die Einschränkung (intra-)molekularer Beweglichkeit ergibt bei der Regressionsanalyse an einem Trainingsdatensatz von 45 Protein-Ligand-Komplexen eine kreuzvalidierte Standardabweichung von 9.3 kJ / mol gegenüber den experimentell bestimmten Affinitäten. Unter Verwendung eines erweiterten Trainingsdatensatzes von 82 Komplexen sowie der Beachtung der Vergrabenheit von Wasserstoffbrücken im Bindungsepitop, der Hinzunahme von aromatischen Wechselwirkungen und der „Bestrafung“ ungünstiger elektrostatischer Wechselwirkungen wird für einen *Vorhersage*datensatz eine Standardabweichung von 8.8 kJ / mol erhalten (Böhm, 1998) (SCORE2). Die Abhängigkeit der (relativen) Beiträge einzelner Terme von der Zusammensetzung des Trainingsdatensatzes als auch der angewendeten Separation der Freien Enthalpie zeigt sich insbesondere beim Vergleich der auf beiden Wegen erhaltenen Parameter.
- Einen analogen Ansatz wie im obigen Fall beschrieben verfolgen auch die Arbeiten von Eldridge *et al.* (Eldridge *et al*., 1997) (82 Komplexe im Trainingssatz, ChemScore)

und Wang *et al.* (Wang *et al*., 1998) (170 Komplexe im Trainingssatz, SCORE). Der Hauptunterschied zu den Ansätzen von Böhm (Böhm, 1994; Böhm, 1998) liegt im ersten Fall (Eldridge *et al.*, 1997) bei der Beschreibung der Beiträge durch die Einschränkung der intramolekularen Flexibilität, im zweiten Fall (Wang *et al.*, 1998) bei der Unterteilung von Wasserstoffbrücken in „starke, moderate und schwache" und der Einbeziehung von Wassermolekülen als Vermittler von Wechselwirkungen. Unter Verwendung eines „evolutionären Testes" zeigt sich außerdem (Wang *et al.*, 1998), daß eine Konvergenz der erzielten Ergebnisse nur bei der Verwendung von mehr als 100 – 120, hinsichtlich des Charakters ihrer intermolekularen Wechselwirkungen ausreichend verschiedenen Protein-Ligand-Komplexen im Trainingsdatensatz erhalten werden kann. Murray *et al.* (Murray *et al*., 1998) verbessern die Vorhersagefähigkeit der in Eldridge *et al.* (Eldridge *et al.*, 1997) erhaltenen Bewertungsfunktion im Hinblick auf *ein spezielles* Protein noch durch Einbeziehung von Zusatzinformation im Rahmen Bayes'scher Statistik.

- Head *et al.* (Head *et al*., 1996) verwenden in ihrem „VALIDATE"-Ansatz eine auf dem AMBER-Kraftfeld (Weiner *et al*., 1984) basierende elektrostatische und sterische Wechselwirkungsenergie, einen HlogP-basierten (Hansch & Leo, 1979) Oktanol-Wasser-Verteilungskoeffizienten, polare und unpolare Kontaktoberflächen und einen Term zur Berücksichtigung intramolekularer Beweglichkeit. Allerdings werden die Koeffizienten hierbei unter Verwendung von 55 Protein-Ligand-Komplexen im Trainingsdatensatz mittels eines *Partial Least Squares*-Verfahrens (Wold *et al.*, 1993) bzw. eines neuronalen Netzes bestimmt.
- In den Arbeiten von Takamatsu und Itai (Takamatsu & Itai, 1998) (29 Avidin-Ligand-Komplexe im Trainingssatz), Venkatarangan und Hopfinger (Venkatarangan & Hopfinger, 1999) (23 Glycogen-Phosphorylase-Inhibitor-Komplexe im Trainingssatz) und Viswanadhan *et al.* (Viswanadhan *et al*., 1996) (11 HIV-1 Protease-Inhibitor-Komplexe im Trainingssatz) werden ebenfalls mit dem AMBER-Kraftfeld (Weiner *et al.*, 1984) berechnete Wechselwirkungsenergien zwischen Protein und Ligand mit zusätzlichen Termen zur Behandlung hydrophober Wechselwirkungen und weiterer entropischer Einflüsse kombiniert und die Koeffizienten durch multiple lineare Regression bestimmt.
- Auch Rognan *et al.* (Rognan *et al*., 1999) und Bohacek und McMartin (Bohacek & McMartin, 1994) verfolgen die Entwicklung empirischer Bewertungsfunktionen, die auf *ein* betrachtetes Protein zugeschnitten sind. So werden im ersten Fall Trainingssät-

ze von fünf kristallographisch bestimmten HLA-A*0201-Peptidinhibitor-Komplexen sowie von 37 modellierten H-2K$^k$-Peptidinhibitor-Komplexen verwendet, um eine an Eldridge *et al.* (Eldridge *et al.*, 1997) angelehnte Funktion zu reparametrisieren. Bohacek und McMartin verwenden sogar nur neun bekannte Thermolysin-Inhibitor-Komplexe zur Kalibrierung. Allerdings gehen in ihre Bewertungsfunktion auch nur die Anzahlen an Wasserstoffbrücken und hydrophoben Kontakten ein.

- Im Gegensatz zu allen bisher betrachteten empirischen Bewertungsfunktionen ist die von Jain (Jain, 1996) entwickelte Funktion kontinuierlich differenzierbar. Terme zur Beschreibung hydrophober und polarer Komplementarität zwischen Rezeptor und Ligand werden dazu durch eine Kombination einer Gauss- und einer sigmoiden Funktion modelliert und mit nur vom Liganden abhängigen Beiträgen für die entropischen Kosten kombiniert. Der Trainingsdatensatz umfaßt 34 Protein-Ligand-Komplexe.

### 3.2.4 Wissensbasierte Ansätze

Allgemein gehen wissensbasierte Systeme von der Annahme aus, daß auf Grundlage einer adäquat repräsentierten Sammlung von Wissen die Ableitung darin inhärent enthaltener Regeln bzw. Gesetzmäßigkeiten möglich ist (Bibel *et al.*, 1993). Im übertragenen Sinne läßt sich damit die Entwicklung einer wissensbasierten Bewertungsfunktion auf molekularer Ebene auf beobachtete Häufigkeitsverteilungen charakteristischer Wechselwirkungen in experimentell bestimmten Systemen zurückführen: nur diejenigen Wechselwirkungen im gerade zu bewertenden System werden als günstig erkannt, die in der Nähe von Häufigkeitsmaxima dieser Wechselwirkungen in der Wissensbasis liegen. Beispiele für eine erfolgreiche Anwendung dieser Verfahren resultieren insbesondere aus dem Bereich der Proteinfaltungsvorhersage (Jernigan & Bahar, 1996; Torda, 1997; Vajda *et al.*, 1997), wobei unter Anwendung des „inversen Boltzmann'schen Gesetzes" (Sippl, 1995) die aus kristallographisch bestimmten Proteinstrukturen erhaltenen Häufigkeitsverteilungen für interatomare Wechselwirkungen in „Freie-Energie-Beiträge" (engl. *potentials of mean force*) bzw. „wissensbasierte Potentiale" umgesetzt werden. Obwohl die thermodynamische Grundlage für dieses Vorgehen (Ben-Naim, 1997; Bürgi & Dunitz, 1988; Finkelstein *et al.*, 1995; Thomas & Dill, 1996) und die verwendete Terminologie (Koppensteiner & Sippl, 1998) kritisch angesehen wird, sind die damit erhaltenen Ergebnisse den mit Molekülmechanik-Kraftfeldern erzielten überlegen (Finkelstein, 1997; Lazaridis & Karplus, 1998; Moult, 1997; Sternberg *et al.*, 1999).

Bis 1999 waren für die Vorhersage von Protein-Ligand-Bindungsaffinitäten nur drei auf dieser Grundlage entwickelte Verfahren mit beschränkter Anwendbarkeit bekannt.

- Unter Verwendung eines Datensatzes von 30 HIV-1, HIV-2 und SIV-Protease-Inhibitor-Komplexen erzeugen Verkhivker *et al.* (Verkhivker *et al.*, 1995) distanzabhängige, wissensbasierte Paarpotentiale. Sie werden mit Termen für Desolvatationsbeiträge von Ligand und Protein unter Verwendung atombasierter Parameter (Sharp *et al.*, 1991) sowie einem auf Arbeiten von Pickett und Sternberg (Pickett & Sternberg, 1993) beruhenden Verfahren zur Abschätzung der Beiträge durch Einschränkung der konformativen Beweglichkeit von Proteinseitenketten kombiniert. Ebenfalls für HIV-1-Protease-Komplexe können damit Unterschiede in den Bindungsaffinitäten nachvollzogen werden.
- Wallquist *et al.* (Wallqvist *et al.*, 1995) ermitteln aus 38 Protein-Ligand-Kristallstrukturen Häufigkeitsverteilungen während der Komplexbildung vergrabener Oberflächenelemente für *Paare* wechselwirkender Atome. Normalisierung mit dem Produkt der vergrabenen Oberflächen der jeweils *einzelnen* Atome liefert daraus Atom-Atom-basierte statistische Präferenzen. Unter Verwendung zweier Parameter, die anhand experimentell bekannter Bindungsaffinitäten des Trainingssatz kalibriert werden müssen, können für 10 HIV-Protease-Inhibitor-Komplexe, die nicht im Trainingssatz enthalten waren, Bindungsaffinitäten mit einer Standardabweichung von 6.3 kJ / mol berechnet werden.
- DeWitte und Shaknovich (DeWitte & Shaknovich, 1996) verwenden 17 bzw. 109 Kristallstrukturen aus der Proteindatenbank PDB (Bernstein *et al.*, 1977) und entwickeln damit je einen Satz von „interatomaren Freien Wechselwirkungsenergien“ (SmoG-Score) für Liganden, die an der Oberfläche eines Proteins binden bzw. solche, die in Bindetaschen binden. Kombiniert mit einem Metropolis-Monte-Carlo-basierten (Metropolis *et al.*, 1953) Aufbauprozeß des Liganden in der Bindetasche werden Komplexe der Purinnucleosid-Phosphorylase, der SH3-Domäne sowie der HIV-1-Protease erzeugt und energetisch bewertet.
- Muegge und Martin (Muegge & Martin, 1999) erzeugten 1999 aus 697 kristallographisch bestimmten Protein-Ligand-Komplexen „Helmholtz’sche Freie Wechselwirkungsenergien“ („PMF“-Score; ***p**otential of **m**ean **f**orce*) unter Verwendung von 16 bzw. 34 Atomtypen für Proteine und Liganden. Die implizite Berücksichtigung der durch Wasser bedingten Beiträge wird durch die Verwendung eines Volumenkorrekturterms und maximale Abstände zwischen den betrachteten Atomen bis zu 12 Å bei

der Erzeugung der Paarverteilungsfunktionen begründet. Für 77 Protein-Ligand-Kristallstrukturen ergibt sich eine Abweichung von 1.8 logarithmischen Einheiten bzgl. der experimentell bestimmten Inhibitionskonstanten.

- Ebenfalls 1999 veröffentlichten Mitchell *et al.* (Mitchell *et al*., 1999b) Paar-Potentiale (BLEEP) aus 820 Protein-Ligand-Atompaarverteilungen unter Verwendung des „inversen Boltzmann-Ansatzes". Mit dem Programm HBPlus (McDonald & Thornton, 1994) regelbasiert beim Protein gesetzte Wasserstoffe wurden mit eingeschlossen. Als Referenzzustand verwenden sie ein von Ng *et al.* (Ng *et al*., 1979) entwickeltes semi-empirisches Ne-Ne-Paarpotential. Außerdem wird die Einbeziehung von Wassermolekülen als Teil des Proteins getestet. Für 90 gemischte Protein-Ligand-Komplexe ergibt sich ein Korrelationskoeffizient von 0.74 (eine Standardabweichung wird nicht aufgeführt) bzgl. experimentell bestimmter Affinitäten (Mitchell *et al*., 1999a). Außer für ein Beispiel in dieser Arbeit ist die Verwendung von wissensbasierten Ansätzen zur Auswahl korrekter Protein-Ligand-Komplex*geometrien* noch nicht herangezogen worden.

### 3.2.5 *Consensus-Scoring*, Filterfunktionen und Ansätze zur ortsaufgelösten Identifizierung von Wechselwirkungen

Ein pragmatischer, wenn auch vom wissenschaftlichen Standpunkt her nicht befriedigender Ansatz zur Steigerung der Verläßlichkeit berechneter Bindungsaffinitäten beruht auf der Zusammenfassung der von mehreren Methoden erhaltenen Ergebnisse. So zeigen Charifson *et al.* (Charifson *et al*., 1999) (Consensus-Scoring), daß durch die UND-Verknüpfung der Ergebnisse der ChemScore-Funktion (Eldridge *et al.*, 1997), der auf dem AMBER-Kraftfeld (Weiner *et al.*, 1984) basierenden Bewertungsfunktion des Docking-Programms DOCK (Meng *et al*., 1992) und dem „stückweisen linearen Potential" (*piecewise linear potential*) (Gehlhaar *et al*., 1995) für drei verschiedene Enzymsysteme die erzielten Anreicherungsraten im Rahmen eines virtuellen Screenings deutlich gesteigert werden können. Ergebnisse durch Mittelwertbildung aus Kombinationen von bis zu fünf verschiedenen QSAR-Methoden erweisen sich bei So und Karplus (So & Karplus, 1999) ebenfalls als den aus den Einzelverfahren erhaltenen Ergebnissen überlegen.

Regressionsbasierte Ansätze bewerten – bedingt durch die Art ihrer Ableitung (s. a. Kap. 3.2.3) – v.a. für den Bindungsprozeß *günstige* Wechselwirkungen, wie sie in Protein-Ligand-Kristallstrukturen auftreten. Um im Rahmen des virtuellen Screenings jedoch auch Protein-Ligand-Geometrien erkennen zu können, die *nicht* in Übereinstimmung mit experimentell zu

erwartenden sind, wurde von Stahl und Böhm (Stahl & Böhm, 1998) die Verwendung von „Filter-Funktionen“ vorgeschlagen.

Unter der Annahme, daß die Bindungsaffiniät in einzelne (additive) Beiträge separiert werden kann (s. a. Kap. 2.2.4), erhalten solche Methoden für die Ligand*optimierung* eine Bedeutung, die (günstige) Wechselwirkungspositionen von Ligandatomen bzw. –gruppen mit dem Protein in seiner Bindetasche identifizieren können.

- Der PROFEC(***p****ictorial* ***r****epresentation* ***o****f* ***f****ree* ***e****nergy* ***c****hanges*)-Ansatz von Radmer und Kollman (Radmer & Kollman, 1998) sowie die Weiterentwicklung von Pearlman (Pearlman, 1999) (OWFEG, ***o****ne-****w****indow* ***f****ree* ***e****nergy* ***g****rid*) beruhen auf FEP-Rechnungen und verwenden zwei Trajektorien aus MD-Simulationen, um die Änderung der Freien Enthalpie durch Hinzufügen eines Atoms / einer Gruppe an *verschiedenen* Positionen um den Inhibitor in Lösung bzw. im Protein zu bestimmen.
- Das archetypische Verfahren GRID von Goodford (Boobbyer *et al*., 1989; Goodford, 1985) basiert dagegen auf einem dafür parametrisierten Kraftfeldansatz und konturiert Bereiche innerhalb der Proteinbindetasche gemäß den Wechselwirkungsenergien einzelner Sonden auf festgelegten Gitterplätzen. Als Sonden stehen z.B. Wasser, eine Amino- und Carboxyl-Gruppe sowie eine hydrophobe Gruppe (DRY) zur Verfügung.
- Beim MCSS (***m****ultiple* ***c****opy* ***s****imultaneous* ***s****earch*)-Ansatz (Miranker & Karplus, 1991) wird dagegen das Kraftfeld CHARMM (Brooks *et al*., 1983) verwendet, um günstige Positionen für in die Bindetasche verteilte Acetamid-, Methanol-, Acetat-, Propan- u.a. Moleküle zu ermitteln.
- Eine GRID-ähnliche Idee, allerdings mit anderer Datenbasis, verfolgen die Methoden X-SITE (Laskowski *et al*., 1996) und SuperStar (Verdonk *et al*., 1999). X-SITE verwendet räumliche Kontaktverteilungen beruhend auf 163 dreiatomigen Fragmenten, um in der Bindetasche günstige Wechselwirkungsregionen zu identifizieren. Die Verteilungen werden aus 83 hochaufgelösten Proteinstrukturen (ohne Liganden) erhalten. SuperStar benutzt die in IsoStar (Bruno *et al.*, 1997) enthaltenen räumlichen Informationen über nichtbindende Wechselwirkungen aus Kristallstrukturen niedermolekularer Verbindungen der CSD (Allen *et al.*, 1991), um Wahrscheinlichkeitsdichten für Kontakte mit Atomen funktioneller Gruppen (etwa Ammoniumstickstoff, Carbonylsauerstoff, Methylkohlenstoff, ...) auf Gitterpositionen in einer Proteinbindetasche zu berechnen.

### 3.2.6 Vergleich der Ansätze

Ein Vergleich der existierenden Verfahren bezüglich Qualität der Ergebnisse und benötigter Laufzeit ist schwierig. Zum einen existiert nach wie vor kein einheitlich verwendeter Datensatz von Protein-Ligand-Komplexen, der als Grundlage des Vergleichs herangezogen werden könnte. Unterschiedlich sind auch die Hardwarevoraussetzungen und der Grad der benötigten Vorbereitungen für einzelne Methoden. Dazu kommt, daß häufig nur hinsichtlich Anzahl und zugrundeliegenden biologischen Systemen eng begrenzte Anwendungsbeispiele veröffentlicht werden, so daß eine zuverlässige Beurteilung des Verfahrens schwer möglich ist. Nichts desto trotz sollen die in den vorherigen Kapiteln 3.2.1 - 3.2.4 hinsichtlich methodischer Aspekte zusammengefaßten, literaturbekannten Verfahren zur Vorhersage von Bindungsaffinitäten dahingehend in Tab. 1 verglichen werden. Darüber hinaus werden Verknüpfungen zu anderen Arbeiten bzw. Methoden aufgeführt.

**Tab. 1:** Vergleich der in Kap. 3.2.1 - 3.2.4 aufgeführten Verfahren zur Vorhersage von Bindungsaffinitäten von Rezeptor-Ligand-Komplexen bei Kenntnis der 3D Rezeptorgeometrie.

| Erstautor (Methodenname) | Referenz | Methode[a)] | Anzahl der Testsysteme[b)] | SD[c)] | Laufzeit | Verknüpfungen[d)] | Kommentare |
|---|---|---|---|---|---|---|---|
| Wong | (Wong & McCammon, 1986) | FEP-MD | 2 | 2.2 | - | - | Ersatz von Benzamidin durch *p*-Fluorbenzamidin bzw. Mutation Gly216Ala in Trypsin |
| Reddy | (Reddy *et al.*, 1992) | FEP-MD | 2 | 3.6 | - | - | Ersatz einer Formyl- durch eine Propa-ragylgruppe |
| Bash | (Bash *et al.*, 1987) | FEP-MD | 2 | 0.5 | - | (Merz & Kollman, 1989) | Ersatz einer NH-Gruppe durch ein O-Atom |
| McCarrick | (McCarrick & Kollman, 1999) | FEP-MD | 3 | 8.4 | - | | Substitution von Phenylringen |
| Ota | (Ota *et al.*, 1999) | NBTI | 2 | 1.7 | - | (Ota & Brunger, 1997) | Bessere Durchmusterung des Konfigura-tionsraums; Ersatz von Benzamidin durch Benzylamin |
| Gerber | (Gerber *et al.*, 1993) | Ableitungen der Freien Energie | 2 x 36 | - | Beschleunigung um $1000^{-1}$ | - | Keine signifikante Korrelation zwischen experimentellen und berechneten Affini-täten |
| Guo | (Guo *et al.*, 1998) | λ-Dynamik-Ansatz | 4 | 2.1 | - | - | Ersatz von Benzamidin mit p-Aminobenzamidin, p-Methylbenzamidin, p-Chlorbenzamidin |
| Aquist | (Aquist *et al.*, 1994) | LIE | 18 | 3.9 | - | (Carlson & Jorgensen, 1995; Wall *et al.*, 1999; Wang *et al.*, 1999a) | SD-Wert von Model 6 in Tab.2 aus (Hansson *et al.*, 1998) |
| Grootenhuis | (Grootenhuis & van Galen, 1995) | CHARMM-Energie | 35 | 8.3 | 2-5 min je Verb. | (Grootenhuis & van Hel-den, 1994; Joseph-McCarthy *et al.*, 1997) | SD-Wert von Protokoll 8 aus Tab. 3 |
| Holloway | (Holloway *et al.*, 1995) | MM2X-Energie | 15 | 5.7 | - | - | SD für Testsatz aus Tab.2 |
| Vajda | (Vajda *et al.*, 1994) | ME-basiert | 9 + 3 + 5 + 9 | 5.4 | - | (Novotny *et al.*, 1989) | SD für Testsatz aus Tab. 1 |
| Wenig | (Weng *et al.*, 1996) | ME-basiert | 9 + 10 + 8 | ≈ 4.2 | - | (Vajda *et al.*, 1994) | Betrachtung von Protein-Protein-Komplexen |

Fortsetzung von Tab. 1:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Williams | (Williams *et al.*, 1991) | ME-basiert | 1 | ≈ 11.4 | - | (Searle *et al.*, 1992; Williams *et al.*, 1993) | SD aus Fehlern der Einzelbeiträge abgeschätzt |
| Krystek | (Krystek *et al.*, 1993) | ME-basiert | 9 | 16.7 | - | - | SD aus Fehlern der Einzelbeiträge abgeschätzt |
| Checa | (Checa *et al.*, 1997) | AMBER-Energie + PBE | 7 | 3.3 | - | - | AMBER-Energie alleine korreliert ebenfalls signifikant. |
| Froloff | (Froloff *et al.*, 1997) | PBE + ASP | 3 + 5 | > 42 | - | - | SD für Testsatz aus Tab. 2; systemat. Fehler |
| Zhang | (Zhang & Koshland, 1996) | PBE + ASP | 9 x 7 | 2.1 | - | - | 9 Mutanten der Isocitrat-Dehydrogenase als Proteinkomponenten |
| Massova | (Massova & Kollman, 1999) | PBE + ASP | 1 | - | - | (Srinivasan *et al*., 1998) | Betrachtung von Protein-Protein-Komplexen |
| Hoffmann | (Hoffmann *et al.*, 1999) | CHARMM + PBE + ASP | 10 | - | „mehrere h für 100 Verbindungen“ | (Rarey *et al*., 1995) | Verbesserung der Plazierung gedockter Geometrien als Ziel |
| Polticelli | (Polticelli *et al.*, 1999) | PBE + ASP | 4 | 56.4 | - | - | SD für Testsätze aus Tab. 1 und 2; systemat. Fehler |
| Shoichet | (Shoichet *et al.*, 1999) | Born-Gl. + ASP | 5 | 20.9 | - | (Kuntz *et al*., 1982; Meng *et al*., 1992) | SD für Testsatz in Tab. III; systemat. Fehler |
| Zou | (Zou *et al.*, 1999) | GB/SA | 6 | 6.3 | 10 s je Verbindung | (Kuntz *et al.*, 1982) | SD für Parametersatz 1 in Tab. 2 und 3 |
| Böhm (SCORE1) | (Böhm, 1994) | Regressionsbasiert | 45 | 9.3 | „mehrere Verbindungen je Sekunde“ | (Böhm, 1992; Rarey *et al*., 1995) | Kreuzvalidierte SD für Funktion #2 |
| Böhm (SCORE2) | (Böhm, 1998) | Regressionsbasiert | 82 + 12 | 8.8 | „mehrere Verbindungen je Sekunde“ | (Böhm, 1992; Rarey *et al*., 1995) | SD für Testsatz aus Tab. 3 |
| Eldridge (ChemScore) | (Eldridge *et al*., 1997) | Regressionsbasiert | 82 + 20 + 10 | 8.7 | - | (Murray *et al.*, 1998) | Kreuzvalidierte SD für gesamten Trainingssatz aus Tab. 8 |
| Wang (SCORE) | (Wang *et al.*, 1998) | Regressionsbasiert | 170 | 6.3 | - | - | Kreuzvalidierte SD für gesamten Trainingssatz aus Tab. 6 |
| Head (VALIDATE) | (Head *et al.*, 1996) | Regressionsbasiert | 51 + 14 + 13 + 11 | 6.3 | - | - | Kreuzvalidierte SD für gesamten Trainingssatz aus Tab. 2 |

Fortsetzung von Tab. 1:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Takamatsu | (Takamatsu & Itai, 1998) | Regressions-basiert | 29 | - | - | - | Kalibirierung ausschließl. an Avidin-Komplexen |
| Hopfinger | (Venkatarangan & Hopfinger, 1999) | Regressions-basiert | 15 | - | - | - | Kalibirierung ausschließl. an Glycogen-Phosphorylase-Komplexen |
| Viswanadhan | (Viswanadhan *et al.*, 1996) | Regressions-basiert | 11 | 2.4 | - | - | Kalibirierung ausschließl. an HIV1-Protease-Komplexen |
| Rognan | (Rognan *et al.*, 1999) | Regressions-basiert | 5 + 37 | 3.1 bzw. 5.1 | - | (Eldridge *et al.*, 1997) | Kalibrierung an HLA-A*0201- bzw. H-2K$^k$-Komplxen; SD jeweils dafür angeg. |
| Bohacek | (Bohacek & McMartin, 1994) | Regressions-basiert | 9 | 2.3 | - | - | Kalibrierung ausschließl. an Thermolysin-Inhibitoren |
| Jain | (Jain, 1996) | Regressions-basiert | 34 | 5.7 | - | - | Kreuzvalidierte SD für Funktion „F“ |
| Verkhivker | (Verkhivker *et al.*, 1995) | Wissens-basiert | 7 | - | - | - | Ableitung und Test der Funktion ausschließlich an HIV- und SIV-Proteasen |
| Wallquist | (Wallqvist *et al.*, 1995) | Wissens-basiert | 8 | 6.3 | - | (Wallqvist & Covell, 1996) | SD für Kalibrierungssatz in Tab. 3 |
| DeWitte (SMoG-Score) | (DeWitte & Shaknovich, 1996) | Wissens-basiert | 17 + 8 + 11 | - | - | - | Keine SD-Werte abgegeben |
| Muegge (PMF-Score) | (Muegge & Martin, 1999) | Wissens-basiert | 77 | 10.3 | - | (Muegge *et al.*, 1999) | SD für Testsatz 6 in Tab. 4 |
| Mitchell (BLEEP) | (Mitchell *et al.*, 1999b) | Wissens-basiert | 90 | - | - | (Mitchell *et al.*, 1999a) | Kein SD-Wert angegeben |

a) Für detaillierte Erläuterungen der aufgeführten Methoden siehe Text. Die Abkürzungen lauten: FEP-MD: Freie Energie Störungsrechnung / Molekular Dynamik, NBTI: Nicht-Boltzmann Thermodynamische Integration, LIE: Lineare Wechselwirkungs Energie, ME: „Hauptgleichung“ (***m****aster* ***e****quation*), PBE: Poisson-Boltzmann-Gleichung, ASP: Atomarer Solvatations Parameter, GB/SA: Generalisierter Born Ansatz. b) Angegeben sind die Anzahl der in einzelnen Testdatensätzen zur Validierung verwendeten Protein-Ligand-Komplexe. c) Angegeben ist die Standardabweichung zwischen berechneten und experimentell gefundenen Bindungsaffinitäten (in kJ / mol; für die Umrechnung in logarithmischen Einheiten angegebener Bindungsaffinitäten wurde eine Temperatur von 298 K angenommen). d) Aufgeführt sind Referenzen zu verwandten Arbeiten bzw. Anwendungen der Methoden in anderen Programmen.

## *3.3 Docking-Verfahren zur Generierung und Bewertung von Konfigurationen von Protein-Ligand-Komplexen*

Docking-Verfahren kombinieren Algorithmen zur Vorhersage möglicher 3D-Strukturen von Protein-Ligand-Komplexen mit Bewertungsverfahren zur Vorhersage der Affinität der so erhaltenen Anordnungen (Blaney & Dixon, 1993; Kuntz, 1992; Kuntz *et al.*, 1994; Lengauer & Rarey, 1996). Selbst bei der Beschränkung auf das Docking *rigider* Liganden in eine *rigide* Bindetasche verbleiben noch 6 Freiheitsgrade für das Suchproblem. Der zu durchsuchende Raum wird daher weiter eingeschränkt, indem ein diskretes Modell der Bindetasche erstellt wird (Vieth *et al.*, 1998b; Westhead *et al.*, 1997). Hierbei definiert i.a. ein Satz von Punkten ihre räumliche Ausdehnung; die Punkte können zudem mit physikochemischen Eigenschaften belegt werden. Ligandatome werden anschließend mit diesen Punkten in Übereinstimmung gebracht. Auch hierfür werden heuristische Suchstrategien eingesetzt, da selbst diese diskrete Formulierung des Docking-Problems zu komplex für eine systematische Suche ist.

Im folgenden wird eine nach den angewendeten Methoden getroffene Auswahl von Dokking-Verfahren beschrieben, wobei insbesondere auf die in dieser Arbeit verwendeten Programme DOCK und FlexX näher eingegangen wird.

- Das in der Gruppe von Kuntz vor fast 20 Jahren entwickelte Programm DOCK (Kuntz *et al.*, 1982) ist das erste dieser Art gewesen. In seiner ursprünglichen Version wurde die Bindetasche eines Proteins mit Kugeln variabler Radien ausgefüllt, wobei die Kugeln mindestens zwei Kontaktpunkte mit der Connolly-Oberfläche (Connolly, 1983) der Tasche haben mußten. Die nach Folgeschritten erhaltenen Kugelmittelpunkte (i.a. 10 bis 100) ergaben so potentielle Positionen für Ligandatome. Durch Vergleich interner Distanzen innerhalb des Kugelsatzes bzw. des Liganden unter Anwendung von Algorithmen für bipartite Graphen (Shoichet *et al.*, 1992) wurden anschließend ähnliche Cluster von mind. vier Kugeln bzw. Ligandatomen gefunden. Aus ihnen läßt sich eine Translations- und Rotationsmatrix bestimmen, die auf alle Atome des Liganden angewendet wird und so eine Transformation des Liganden in die Bindetasche bewirkt. Abschließend wird die Position des als starren Körper behandelten Liganden in der Proteinbindetasche durch Energieminimierung optimiert (Meng *et al.*, 1993). Zur Steigerung der Effiziens der Überlagerung von Ligandatomen und Kugelzentren verwendet die aktuelle DOCK-Version eine einfachere Graphenrepräsentation des Suchraumes sowie eine Cliquen-Such-Technik. Zusätzlich zu der von den Kugeln dargestellten sterischen Information können auch noch chemische Eigenschaften von ihnen repräsen-

tiert werden (Oshiro & Kuntz, 1998). Damit wird der Suchraum nochmals eingeschränkt, indem nur noch komplementäre Kugel-Ligandatom-Paare betrachtet werden.

Das Docking von Liganden unter Berücksichtigung ihrer Flexibilität kann mit DOCK auf zwei Wegen erfolgen. Zum einen können vorab für einen Liganden verschiedene Konformationen erzeugt werden und diese dann, wie vorher beschrieben, rigide in die Bindetasche eingepaßt werden (Wang *et al*., 1999b). Ein alternativer Ansatz beruht auf einem inkrementellen Ligandenaufbau in der Proteintasche (Makino & Kuntz, 1997): nach Auswahl, Plazierung und geometrischer Optimierung eines Ankerfragmentes werden die verbleibenden Fragmente schrittweise angefügt, wobei ein *Backtracking*-Algorithmus den zu durchsuchenden Konformationsraum begrenzt. Zum Abschluß erfolgt eine flexible Simplex-Minimierung der erhaltenen Ligandkonfiguration.

Zur Bewertung der erhaltenen Protein-Ligand-Anordnungen kann der ursprünglich entwickelte und auf sterischer Komplementarität beruhende *contact score* (Kuntz *et al.*, 1982), ein auf dem AMBER-Kraftfeld beruhender *energy score* (Meng *et al.*, 1992) sowie ein durch Modifikation der van der Waals-Energy davon abgewandelter *chemical score* (Makino & Kuntz, 1997) verwendet werden. Obwohl DOCK seit seiner Einführung erfolgreich zur Generierung von Leitstrukturen für verschiedene biologische Zielmoleküle eingesetzt worden ist (siehe (Joseph-McCarthy, 1999) für eine Übersicht), gibt es bislang keine groß angelegte Validierungsstudie, die das Zusammenspiel der verschiedenen Ligandkonfigurations-Generierungsmöglichkeiten mit den zur Verfügung stehenden Berwertungsfunktionen für verschiedene Rezeptor-Ligand-Komplexe untersucht. DOCK benötigt etwa 1 - 3 min für das flexible Docking eines Liganden.

- Der Ansatz, Ligandatome mit in der Bindetasche lokalisierten Bereichen mit komplementären physikochemischen Eigenschaften zur Deckung zu bringen, bildet auch die Grundlage des Programmes FlexX (Rarey *et al.*, 1995; Rarey *et al*., 1996a; Rarey *et al*., 1996b). FlexX analysiert zunächst die Bindetasche, indem es Wechselwirkungsbereiche des Proteins mit definierten Eigenschaften (Wasserstoffbrücken, Salzbrücken, Metallkontakte, hydrophobe Konatkte) identifiziert. Diese Bereiche können dabei die Form von Kugeln, Kugelkappen, Kugelschichten oder sphärischen Rechtecken haben. Davon ausgehend werden für alle Paare von Wechselwirkungszentren die Distanzen bestimmt und geordnet in einer *hash*-Tabelle abgelegt. Für aus drei Ligandatomen gebildete Suchanfragen wird diese Tabelle nach geometrisch und physikochemisch pas-

senden Kombinationen aus Wechselwirkungszentren des Proteins durchsucht und – bei Erfolg – der Ligandenteil entsprechend transformiert.

Docking flexibler Liganden erfolgt auch hier nach einem inkrementellen Konstruktionsalgorithmus: Zerlegung des Liganden an drehbaren Bindungen, Auswahl eines Ankerfragmentes (Rarey *et al.*, 1997), Plazierung des Ankerfragmentes und Anbau der verbleibenden Fragmente. Der Konformationssuchraum wird hierbei durch die Verwendung empirischer Torsionspotentiale (Klebe & Mietzner, 1994) zusätzlich eingeschränkt.

Zur Bewertung der im Aufbau befindlichen als auch abschließenden Protein-Ligand-Orientierungen verwendet FlexX eine modifizierte Version der regressionsbasierten Bewertungsfunktion von Böhm (Böhm, 1994). Für einen Testsatz von 200 Rezeptor-Ligand-Komplexen aus der PDB können für 46.5 % der Fälle eine Ligandorientierung mit weniger als 2 Å mittlerer quadratischer Abweichung in den kartesischen Koordinaten (***r**oot **m**ean-**s**quare **d**eviation, rmsd*) gegenüber der Kristallgeometrie und *bester* Bewertung gefunden werden. Betrachtet man jeweils alle für eine Protein-Ligand-Kombination erzeugten Anordnungen *unabhängig* von ihrer Bewertung, haben sogar 70 % der Fälle einen *rmsd* < 2.0 Å (Kramer *et al.*, 1999). Die mittlere Laufzeit beträgt dabei etwa 1.5 min pro Ligand.

- Das Programm GOLD (***G**enetic **O**ptimisation for **L**igand **D**ocking*) (Jones *et al.*, 1995) verwendet einen genetischen Algorithmus zur Plazierung von Liganden in einer Bindetasche. Ligandkonformationen wie auch relative Orientierungen bzgl. der Rezeptortasche werden dabei als Bit- bzw. Ganzzahlrepräsentationen in sog. Chromosomen dargestellt. Genetische Operatoren wie Mutation oder Überkreuzungen zusammen mit einer *fitness*-Funktion entscheiden dann über die Entwicklung der Population aus Protein-Ligand-Anordnungen. Während die oben erwähnten Programme DOCK und FlexX die Bindetasche des Rezeptors als völlig starr ansehen, werden bei GOLD terminale Donor- und Akzeptor-Gruppen des Proteins als drehbar behandelt. Die abschließende Bewertungsfunktion besteht aus einem Term für van der Waals-Beiträge und einem für Wasserstoffbrückenbeiträge. Für einen Testdatensatz aus 100 Komplexen aus der PDB werden für 67 % *rmsd*-Werte < 2.0 Å auf dem bestbewerteten Rang gefunden (Jones *et al.*, 1997).
- Der QXP-Algorithmus (McMartin & Bohacek, 1997) verwendet einen Monte-Carlo-Minimierungsansatz als Grundlage für das Einpassen der Liganden in die Bindetasche. Hierbei werden sowohl Torsionswinkel des Liganden als auch seine relative Orientie-

rung zum Protein durch zufällige Variationen verändert; die damit erhaltene neue Anordnung wird mit dem AMBER-Kraftfeld (Weiner *et al.*, 1984) in Gegenwart des Proteins im kartesischen Raum minimiert. Das Metropolis-Kriterium (Metropolis *et al.*, 1953) entscheidet dann, ob die so erhaltene Konfiguration im Gegensatz zu ihrer Ausgangskonfiguration als Startpunkt für einen neuen Zyklus verwendet wird. Für einen Testsatz aus 12 Komplexen werden in mehr als 80 % der Fälle *rmsd*-Werte < 1.0 Å bzgl. einer minimierten Kristallstruktur für Ligandorientierungen auf dem bestbewerteten Rang gefunden; die Rechenzeiten betragen dabei 3 min bis 3 h je Ligand (für 7 - 24 drehbare Bindungen).

- AutoDock (Morris *et al.*, 1996) basiert auf einem *simulated annealing*-Verfahren, d.h. einem temperaturgesteuerten Monte-Carlo-Algorithmus. Wie beim Abkühlen einer Schmelze erlaubt man dem Protein-Ligand-System dabei zunächst, sich frei, d.h. durch zufällige Veränderungen seiner geometrischen Variablen, über die multidimensionale Energiehyperfläche zu bewegen. Mit abnehmender Temperatur sinkt dabei auch seine mittlere Energie, Bewegungen in Richtung abnehmender Energie werden zunehmend stärker begünstigt. Der Endzustand entspricht dann – bei richtiger Durchführung - dem globalen Energieminimum. Auch AutoDock verwendet das AMBER-Kraftfeld (Weiner *et al.*, 1984) zur Bewertung der erhaltenen Protein-Ligand-Anordnungen. Für sechs Protein-Ligand-Komplexe mit 0 – 10 drehbaren Bindungen in den Liganden werden in allen Fällen *rmsd*-Werte < 1.5 Å auf dem Rang niedrigster Energie gefunden. Je nach Flexibilität des Liganden werden dafür 5 - 18 min an Rechenzeit benötigt.

# 4 Theorie und Methoden

Im folgenden werden die Theorie sowie die angewendeten Methoden zur Ableitung und Validierung der wissensbasierten Bewertungsfunktion für Protein-Ligand-Komplexe beschrieben. Aufbauend auf den von Tanaka und Scheraga (Tanaka & Scheraga, 1976), Sippl (Sippl, 1990; Sippl, 1993) und Jernigan *et al.* (Bahar & Jernigan, 1997; Miyazawa & Jernigan, 1985; Miyazawa & Jernigan, 1996) begründeten Formalismen wird hierbei besonders auf die Modifikationen eingegangen, die zur Anpassung des Verfahrens an das Struktur- und Affinitätsvorhersageproblem bei Rezeptor-Ligand-Komplexen eingeführt wurden (Kap. 4.1 – 4.5). Nach einer Aufstellung von Kriterien zur Bewertung der Funktion (Kap. 4.6) wird ein auf ihrer Grundlage beruhender Ansatz zur Identifikation von Wechselwirkungszentren in einer Proteinbindetasche beschrieben (Kap. 4.7) und ein Weg zu ihrer proteinspezifischen Adaptierung unter Einbeziehung zusätzlicher struktureller und energetischer Informationen vorgestellt (Kap. 4.8). Die Beschreibung der Aufbereitung der verwendeten Testdatensätze erfolgt zum Schluß (Kap. 4.9).

## *4.1 Motivation des wissensbasierten Ansatzes und Begriffsbestimmungen*

Die Betrachtungen in Kap. 2.2 machen deutlich, daß zur adäquaten Beschreibung der Bindung eines Liganden an ein Protein nicht nur Wechselwirkungs*energien* (bzw. *-enthalpien*) zwischen beiden betrachtet werden können, sondern zusätzlich Effekte durch (De-)Solvatation der Reaktionspartner bzw. Reorganisation im Lösemittel beachtet werden müssen. Eine *explizite* Beschreibung dieser (auch Entropie-Änderungen bedingenden) Effekte in Form von Paar-, Triplett- oder Wechselwirkungen höherer Ordnung bei nur *einer* gegebenen Anordnung von Protein und Ligand zueinander ist nicht möglich (Meirovitch, 1998). Eine *implizite* Berücksichtigung dieser komplexen Lösemittelbeiträge zusammen mit enthalpischen Beiträgen durch intermolekulare Wechselwirkungen ist allerdings in dem hier verwendeten wissensbasierten Formalismus enthalten, bei dem „Potentiale“ aus bekannten Molekülstrukturen extrahiert werden (Jernigan & Bahar, 1996; Sippl, 1995) (Abb. 5).

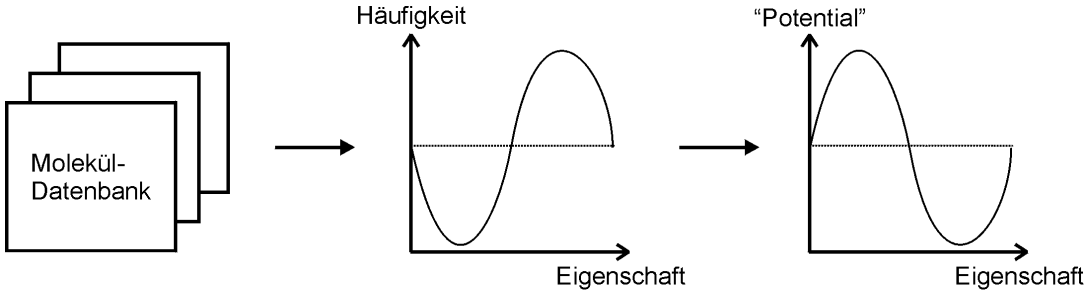



**Abb. 5:** Vorgehensweise bei wissensbasierten Ansätzen. Aus einer Wissensbasis (hier: Moleküldatenbank) werden Häufigkeiten interessierender Eigenschaften extrahiert. Bezogen auf einen Referenzzustand (gepunktete Linie) lassen sich daraus „Potentiale“ ableiten.

Die physikochemische Grundlage der so erhaltenen „Potentiale“ wie auch die für sie verwendete Terminologie ist allerdings keineswegs eindeutig (Koppensteiner & Sippl, 1998). Der häufig dafür gebrauchte Ausdruck *potential of mean force* geht auf die auf der statistischen Mechanik beruhende Theorie von Flüssigkeiten zurück (Ben-Naim, 1987; Hill, 1956). Dort ergibt sich ein *potential of mean force* aus einer *n*-Teilchen-Korrelationsfunktion $g^{(n)}(\vec{X}_1,...,\vec{X}_n)$ mit $\vec{X}_i$ als Ortskoordinaten des *i*-ten Teilchens gemäß:

$$W^{(n)}(\vec{X}_1,...,\vec{X}_n) = -RT \ln g^{(n)}(\vec{X}_1,...,\vec{X}_n) \qquad \textbf{Gl. 9}$$

Für den Fall zweier wechselwirkender Teilchen unter der Annahme sphärischer Symmetrie entsteht aus der Paarkorrelationsfunktion $g^{(2)}(\vec{X}_1,\vec{X}_2)$ die radiale Paarverteilungsfunktion $g^{(2)}(r_{12})$, wobei $r_{12}$ gleich $|\vec{X}_1 - \vec{X}_2|$ ist. Grundsätzlich läßt sich diese Funktion auch für Paare wechselwirkender Atome aus (Protein-)Kristallstrukturen bestimmen, indem für Distanzintervalle zwischen $r_{12}$ und $r_{12} + dr$ die Häufigkeit des Auftretens für zwei bestimmte Atome ausgezählt wird. Allerdings ist die häufig zugrundegelegte Annahme der Anwendbarkeit des inversen Boltzmann’schen Gesetzes gemäß Gl. 9 für auf diesem Weg erhaltene Verteilungsfunktionen nicht haltbar (Ben-Naim, 1997; Bürgi & Dunitz, 1988; Thomas & Dill, 1996).

Das Boltzmann’sche Gesetz gilt streng nur für Ensembles *gleicher* Teilchen im thermodynamischen Gleichgewicht. Eine Datenbank von Proteinstrukturen dagegen besteht aus vielen *verschiedenen* Teilchen, die sich nicht notwendigerweise im gleichen globalen Energieminimum aufhalten. Darüber hinaus ist noch nicht einmal bekannt, ob jede Proteinstruktur sich in *ihrem* globalen Energieminimum befindet. Es kommt noch hinzu, daß die Verwendung *einer* absoluten Temperatur *T* in Gl. 9 ebenfalls nicht zu begründen ist (Finkelstein *et al.*, 1995): Proteinkristalle werden über einen weiten Temperaturbereich gezüchtet und vermessen.

In dem hier vorgestellten Ansatz geht es allerdings – im Gegensatz zu vergleichbaren Anwendungen bei Proteinfaltungsvorhersagen – nicht um die Eigenschaft eines Proteins als Ganzem. Vielmehr geht es um die Eigenschaft eines speziell interessierenden „Bereiches“ (etwa eines Atom-Atom-Paarkontakts), der in die molekulare Umgebung verschiedener Proteinstrukturen eingebettet ist. Die dem Ansatz zugrundeliegende Annahme ist daher, daß dieser „Bereich“ in der Datenbasis für alle relevanten Zustände mit statistischer Signifikanz repräsentiert wird.

Unter diesen Voraussetzungen ist der hier gewählte Ansatz als heuristisch motiviert zu betrachten und kann – wie jedes empirische Verfahren - nur durch die damit erzielbaren Ergebnisse hinsichtlich Reproduzierbarkeit und Vorhersagbarkeit experimenteller Daten gerechtfertigt werden. Nach Koppensteiner und Sippl (Koppensteiner & Sippl, 1998) sind daher die im folgenden gemäß

$$"Potential_X" = -\ln g_X \qquad \textbf{Gl. 10}$$

erhaltenen Terme (mit $g_X$ als Verteilungsfunktion eines „Bereiches unter spezieller Betrachtung“ $X$) als „wissensbasierte Größen“ oder „statistische Präferenzen“ zu bezeichnen, und der im weiteren verwendete Ausdruck „Potential“ ist in diesem Sinn zu verstehen.

## *4.2 Distanzabhängige Paarpotentiale*

### 4.2.1 Definition

Die exakte paarweise Addivität von Wechselwirkungen in einem Vielteilchensystem (d.h. das Superpositionsprinzip) gilt nur beim Auftreten von *Zentral*kräften zwischen den betrachteten Teilchen, etwa zwischen Punktladungen in einem homogenen Medium (Dransfeld *et al*., 1992), nicht jedoch für Protein-Ligand-Systeme in wäßriger Umgebung (Ben-Naim, 1997). Die Anwendung einer empirischen Funktion aus *Paar*potentialen für diese Systeme geht auf eine zuerst von Kirkwood (Kirkwood, 1935) eingeführte Näherung dieses Superpositionsprinzips zurück und wird durch experimentelle Befunde gestützt. So haben Crippen und Snow (Crippen & Snow, 1990) bei einer statistischen Analyse von 22 Proteinen in der PDB zeigen können, daß der Verzicht auf Drei- und Mehrteilchen-Wechselwirkungsterme für Proteine gerechtfertigt ist. Dies unterstreichen auch die Erfolge bei der Verwendung derartiger Potentialmodelle zur Erkennung von nativ-ähnlichen Strukturen gegenüber nicht-nativen im Rahmen der Proteinfaltungsvorhersage (Bowie *et al*., 1991; Hendlich *et al*., 1990; Jones *et al*., 1992; Kocher *et al*., 1994; Miyazawa & Jernigan, 1996).

Im Sinne dieser physikalischen Vereinfachung soll daher ein Teil der Wechselwirkungen zwischen Protein und Ligand mit Hilfe von wissensbasierten Paarpotentialen beschrieben werden. Da mit zunehmender Detailliertheit des Ansatzes auch der aus den Daten extrahierbare Informationsgehalt wächst (Bahar & Jernigan, 1997; Park & Levitt, 1996), werden die Potentiale Atompaar-basiert und distanzabhängig abgeleitet. Dabei werden folgende grundlegende Annahmen gemacht:

1. Die Distanzen zwischen Ligand- und Proteinatomen sind paarweise unabhängig voneinander.
2. Die Verteilung interatomarer Distanzen ist für verschiedene molekulare Umgebungen des betrachteten Atompaares ähnlich.
3. Die Verteilungen interatomarer Distanzen sind für verschiedene Atom-Atom-Paarkombinationen ausreichend scharf und verschieden voneinander.

Ausgehend von dem von Sippl (Sippl, 1990; Sippl, 1993) entwickelten Formalismus ergibt sich für ein Ligandatom *l* und ein Proteinatom *p* jeweils mit den Ortskoordinaten $\vec{X}_l$ bzw. $\vec{X}_p$ die statistische Nettopräferenz $\Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$ in Abhängigkeit des (diskretisierten) Abstandes zwischen beiden Atomen innerhalb eines Intervalls *i* mit den Grenzen [$r_i$, $r_i + dr$) zu

$$\Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) = W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) - W(r_i) = -\ln \frac{g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)}{g(r_i)} \qquad \textbf{Gl. 11}$$

Mit T(*x*) sei dabei eine Funktion bezeichnet, die den Typ des Atoms *x* definiert. Beschränkt man den (kontinuierlichen) Abstand $d_{l,p} = \left|\vec{X}_l - \vec{X}_p\right|$ generell auf ein Intervall [$r_{min}$, $r_{max}$), so ergibt sich bei festgelegter Intervallweite *dr* der dazugehörige Wert $r_i$ zu

$$r_i = r_{\min} + \left\lfloor \frac{d_{l,p} - r_{\min}}{dr} \right\rfloor dr \qquad \textbf{Gl. 12}.$$

$\lfloor ... \rfloor$ steht für die „größte ganze Zahl kleiner als ...“. $g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$ ist hierbei eine normalisierte radiale Paarverteilungsfunktion für Atome des Typs T(*l*) und T(*p*), die sich in einem Abstand $d_{l,p}$ innerhalb [$r_i$, $r_i + dr$) befinden.

$$g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) = \frac{N_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) \big/ 4\pi r_i^{\,2}}{\sum_{i \in I} (N_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) \big/ 4\pi r_i^{\,2})} \qquad \textbf{Gl. 13}$$

Die Summation läuft dabei über alle Intervalle *i* in dem Bereich [$r_{min}$, $r_{max}$). Die von den Arbeiten von Sippl (Sippl, 1990; Sippl, 1993) abweichende Skalierung mit $4\pi r_i^2 dr$ berücksichtigt das Volumen der Kugelschale mit dem Radius $r_i$ und der Dicke *dr* und führt zu einer

schnelleren Konvergenz der Potentialwerte gegen Null bei großen Distanzen (Bahar & Jernigan, 1997).

Die Anzahl der Atompaare $N_{T(l),T(p)}(r_i)$ mit den Typen T($l$) und T($p$) innerhalb [$r_i$, $r_i$ + $dr$) ergibt sich durch Auszählen der Häufigkeiten ihres Auftretens über alle Ligandatome $L_k$ und Proteinatome $P_k$ eines Komplexes $k$ für alle (nativen) Komplexe $K_n$ in der betrachteten Datenbank:

$$N_{T(l),T(p)}(r_i) = \sum_{k \in K_n} \sum_{l \in L_k} \sum_{p \in P_k} \delta(d_{l,p}, r_i) \qquad \textbf{Gl. 14}$$

Die Delta-Funktion $\delta(d_{l,p},r_i)$ ergibt 1, wenn $d_{l,p} \in [r_i, r_i + dr)$, sonst 0.

$g(r_i)$ ist die normalisierte radiale Paarverteilungsfunktion für zwei Atome beliebigen Typs ebenfalls mit einem Abstand innerhalb des Intervalls [$r_i$, $r_i$ + $dr$). Sie entspricht einem Referenzzustand und beinhaltet alle nicht-spezifischen Informationen, die typisch für ein beliebiges Atompaar in einer Protein-Ligand-Umgebung sind. Damit ergibt sich nach Gl. 11 die statistische Nettopräferenz $\Delta W_{T(l),T(p)}(r_i)$ als Differenz der statistischen Präferenz des betrachteten Subsystems $W_{T(l),T(p)}(r_i)$ und des Referenzzustandes $W(r_i)$.

Der auf die wissensbasierten Paarpotentiale zurückgehende Teil der Wechselwirkungen für eine gegebene Konfiguration von Protein und Ligand resultiert schließlich zu

$$\Delta W_{Paar} = \sum_{l \in L_k} \sum_{p \in P_k} \Delta W_{T(l),T(p)}(r_i) \qquad \textbf{Gl. 15}$$

mit $r_i \le d_{l,p} < r_i + dr$.

Eine alternative, aber zu dem gleichen Ergebnis führende Formulierung ergibt sich im Rahmen probabilistischer Methoden der Wissensrepräsentation durch Anwendung Bayesscher Statistik (Jensen, 1996). Parallel zu dieser Arbeit ist dieses von Samudrala und Moult für ein Verfahren zur Proteinstrukturvorhersage gezeigt worden (Samudrala & Moult, 1998). Ausgangspunkt ist die *bedingte* Wahrscheinlichkeit, daß für eine gegebene Menge von Distanzen zwischen Ligand- und Proteinatomen $D = \{d_{l,p} \mid l \in L_k, p \in P_k, k \in K\}$ ein Komplex $k$ Element der Menge nativer Komplexe $K_n$ ist, geschrieben $P(k \in K_n \mid D)$. $K_n$ ist dabei Teilmenge einer Menge gegebener Komplexe $K$, die auch nicht-native Protein-Ligand-Komplexe enthält.

Die Anwendung des Satzes von Bayes liefert nun:

$$P(k \in K_n \mid D) = \frac{P(D \mid k \in K_n) \cdot P(k \in K_n)}{P(D)} \qquad \textbf{Gl. 16}$$

$P(D \mid k \in K_n)$ ist dabei die Wahrscheinlichkeit für eine Menge von Distanzen zwischen Ligand- und Proteinatomen $D$ in einem nativen Komplex $k \in K_n$, $P(k \in K_n)$ ist die Wahrschein-

lichkeit, daß Komplex *k* Element der Menge $K_n$ ist und $P(D)$ die Wahrscheinlichkeit für das Auftreten eines Satzes von Distanzen *D*.

Benutzt man obige Annahme, daß alle Distanzen paarweise *unabhängig* voneinander sind, so läßt sich die Wahrscheinlichkeit ihres gleichzeitigen Auftretens in *D* als Produkt der Wahrscheinlichkeiten ihres einzelnen Auftretens schreiben (Satz über statistische Unabhängigkeit):

$$P(D) = \prod_{l \in L_k} \prod_{p \in P_k} P(d_{l,p}) \qquad \textbf{Gl. 17}$$

sowie

$$P(D \mid k \in K_n) = \prod_{l \in L_k} \prod_{p \in P_k} P(d_{l,p} \mid k \in K_n) \qquad \textbf{Gl. 18}$$

Einsetzen in Gl. 16 und Verwendung des negativen natürlichen Logarithmus liefert

$$-\ln P(k \in K_n \mid D) = -\sum_{l \in L_k} \sum_{p \in P_k} \ln \frac{P(d_{l,p} \mid k \in K_n)}{P(d_{l,p})} - \ln P(k \in K_n) \qquad \textbf{Gl. 19},$$

d.h. eine Bewertung (*log likelihood*) dafür, ob der Komplex *k* in der gegebenen Konfiguration *D* nativ ist.

Für $P(d_{l,p} \mid k \in K_n)$ können dabei Werte aus diskreten Wahrscheinlichkeitsverteilungen über die Atomtypen T(*l*) und T(*p*) aus einem Satz bekannter (nativer) Komplexstrukturen eingesetzt werden, wobei $d_{l,p}$ auch hier in einem Intervall $[r_i, r_i + dr)$ liegt.

$$P(d_{l,p} \mid k \in K_n) \rightarrow P(\mathrm{T}(l), \mathrm{T}(p), r_i) \qquad \textbf{Gl. 20}$$

$P(d_{l,p})$ ist im Sinne Bayes'scher Statistik die *a priori* Wahrscheinlichkeit, zwei Atome *l* und *p* im Abstand $d_{l,p}$ in einem beliebigen (nativen oder nicht-nativen) Komplex zu finden. Sie spiegelt damit das Wissen über interatomare Distanzen *ohne* Kenntnisse über Strukturen nativer Komplexe wieder. Im Rahmen diskreter Wahrscheinlichkeitsverteilungen folgt damit

$$P(d_{l,p}) \rightarrow P(r_i) \qquad \textbf{Gl. 21}$$

Setzt man Gl. 11 in Gl. 15 ein und vergleicht das Ergebnis mit Gl. 19 unter Verwendung von Gl. 20 und Gl. 21, so lassen sich folgende Terme miteinander identifizieren:

$$\Delta W_{Paar} \cong -\ln P(k \in K_n \mid D) \quad ; \quad g_{\mathrm{T}(l)\mathrm{T}(p)}(r_i) \cong P(\mathrm{T}(l), \mathrm{T}(p), r_i) \quad ; \quad g(r_i) \cong P(r_i) \qquad \textbf{Gl. 22}$$

Der Term $-\ln P(k \in K_n)$ in Gl. 19 ist für ein gegebenes *k* eine von dessen Konfiguration *D* unabhängige Konstante. Für den Vergleich verschiedener Konfigurationen *eines* Komplexes *k* ist er also unerheblich. Ein analoger Ausdruck tritt in dem ursprünglichen, auf das inverse Boltzmann'sche Gesetz gestützten Formalismus von Sippl (Sippl, 1990; Sippl, 1993) ebenfalls auf und wird dort unter Annahme der Gleichheit zweier Zustandssummen auf Null gesetzt. Dieser Term ist von Bedeutung beim Vergleich *verschiedener* Komplexe *k*. Nur wenn angenommen werden kann, daß er für diese Komplexe ähnliche Größen annimmt, können

auch die mit Gl. 15 bzw. Gl. 19 erhaltenen Bewertungen der Wechselwirkungen miteinander in Relation gesetzt werden.

### 4.2.2 Wahl des Referenzzustandes

Die Wahl des Referenzzustandes ($g(r_i)$ bzw. $P(r_i)$) ist nach Gl. 11 bzw. Gl. 16 bestimmend für den Verlauf der Paarpotentiale und damit für die Ausnutzung des zur Verfügung stehenden Wissens. Eine denkbare Annahme für den Referenzzustand wäre die vollständiger Separation von Protein und Ligand. Gegen eine solche Definition sprechen jedoch nicht nur die Schwierigkeiten, diesen Zustand ohne explizite Berücksichtigung des Lösemittels im Rahmen des obigen Formalismus‘ darzustellen. Im Hinblick auf die Anwendung der Potentiale für die Erkennung einer nativ-ähnlichen Komplexgeometrie aus einer Menge, die auch nicht-native Strukturen enthält, wäre dagegen ein Referenzzustand vorteilhaft, der alle (sicher nicht-spezifischen) Informationen auch über nicht-native Geometrien enthält. Mit anderen Worten: da die in der Menge vorgeschlagener Komplexstrukturen enthaltenen Geometrien mehr oder weniger alle kompakt sein werden, ist diese Information nicht-spezifisch für eine native Komplexgeometrie und trägt daher nicht wesentlich zur Unterscheidung zwischen nativen und nicht-nativen Geometrien bei. Eine gute Annahme eines Referenzzustandes (oder nach Gl. 21 einer *a priori*-Verteilung) ergibt sich daher aus der Betrachtung kristallographisch bekannter (kompakter) Geometrien mit Atomen willkürlich angenommenen Typs an den jeweiligen Atompositionen. Eine Mittelung über die verschiedenen Atomtypen käme dabei ihrer zufälligen Anordnung gleich.

Zwei Wege der Mittelung sind dabei denkbar: zum einen unter Verwendung der normierten radialen Paarverteilungsfunktion (Gl. 24) und zum anderen unter Verwendung der ausgezählten Häufigkeiten des Auftretens der Atom-Atom-Kontakte (Gl. 23). Der letztere Fall

$$\tilde{g}(r_i) = \frac{\sum_{\mathrm{T}(l)} \sum_{\mathrm{T}(p)} (N_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) / 4\pi r_i^{\,2})}{\sum_{\mathrm{T}(l)} \sum_{\mathrm{T}(p)} \sum_{i \in I} (N_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) / 4\pi r_i^{\,2})} \qquad \textbf{Gl. 23}$$

entspricht dem von Sippl (Sippl, 1990; Sippl, 1993) verfolgten Verfahren und führt zu einer normierten radialen Paarverteilungsfunktion $\tilde{g}(r_i)$, die von zahlenmäßig häufig auftretenden Fällen $N_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$ dominiert wird. Dies bedeutet im Zusammenhang mit Gl. 11, daß Verteilungen mit vielen Beobachtungen nur geringe statistische Nettopräferenzen aufweisen (s.a. Kap. 5.4.5).

Im Gegensatz dazu wird der Referenzzustand unabhängig von der Anzahl auftretender Fälle in einzelnen Paarverteilungen, wenn man über die normierten radialen Paarverteilungsfunktionen mittelt:

$$g(r_i) = \frac{\sum_{\mathrm{T}(l)} \sum_{\mathrm{T}(p)} g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)}{\left\|\mathrm{T}(l)\right\| \cdot \left\|\mathrm{T}(p)\right\|} \qquad \textbf{Gl. 24}$$

Mit $\left\|\mathrm{T}(l)\right\| \cdot \left\|\mathrm{T}(p)\right\|$ als der Anzahl aller möglichen Kombinationen von Ligand- und Proteinatomtypen wird dieser letzte Weg in dem hier beschriebenen Ansatz verfolgt.

### 4.2.3 Wahl der Intervallparameter und Glättung der Rohdaten

Die aus Kristallstrukturen von Protein-Ligand-Komplexen gewonnenen Distanzverteilungen werden auf Atom-Atom-Abstände innerhalb des Intervalls [$r_{min}$, $r_{max}$) beschränkt. Im oben vorgestellten Formalismus ist es möglich, in Proteinen vorkommende Kofaktoren bzw. Metallatome als Teil des Proteins aufzufassen und in die Distanzverteilungen mit einzubeziehen. Die untere Grenze $r_{min}$ wird daher auf 1 Å gesetzt, da Wechselwirkungen zwischen Metallatomen und Sauerstoff- bzw. Stickstoffatomen des Liganden Abstände von etwa 1.8 Å aufweisen (Harding, 1999). Die Festlegung der oberen Abstandsgrenze $r_{max}$ bestimmt maßgeblich den Informationsgehalt und damit die Form der mit Gl. 11 erhaltenen statistischen Nettopräferenzen (Godzik *et al*., 1995). Atompaarverteilungen für kurze obere Abstände (etwa 6 Å) betonen spezifische Wechselwirkungen zwischen funktionellen Gruppen des Liganden mit denen benachbarter Gruppen des Proteins. Eine Erweiterung auf dazu größere Abstände (bis zu 12 Å wie in der Arbeit von Muegge und Martin (Muegge & Martin, 1999)) bezieht in verstärktem Maße den Einfluß eines gemittelten Lösemittelbeitrages mit ein, der hauptsächlich entropiebedingt ist (DeWitte & Shaknovich, 1996). Im Hinblick auf das Erkennen korrekter Ligandbindungsmoden aus einem Satz erzeugter Rezeptor-Ligand-Konfigurationen wird daher die obere Grenze $r_{max}$ auf 6 Å festgelegt, um Unterschiede in spezifischen, kurzreichweitigen Wechselwirkungen zu repräsentieren. Die Distanz von 6 Å ist dabei gerade so groß, daß ein Wassermolekül nicht als gegenseitiger Mediator von Wechselwirkungen zwischen Protein- und Ligandatom auftreten kann. Ergebnisse für davon abweichende Größen von $r_{max}$ werden in Kap. 5.4.5 diskutiert. Die Weite der Intervalle *dr* wird als Kompromiß hinsichtlich einer ausreichend guten Auflösung und der zur Verfügung stehenden Menge an Daten für Atom-Atom-Abstände in der Datenbank zu 0.1 Å festgelegt.

Um den Effekt der Diskretisierung der Distanzverteilungen zu berücksichtigen und der Tatsache Rechnung zu tragen, daß Atompositionen in Kristallstrukturen inhärent mit einer Unsicherheit in den Ortskoordinaten von etwa 1/6 der maximalen Auflösung behaftet sind (bei 2.5 Å Auflösung also etwa 0.4 Å Ortsunschärfe) (Dauber-Osguthorpe *et al.*, 1988; Kossiakoff *et al.*, 1992; Wlodawer *et al.*, 1987), wird auf die erhaltenen Rohdaten eine Glättungsfunktion angewendet. Sie besteht aus einer gleichschenkligen Dreiecksfunktion, die an der Basis eine Breite von 0.4 Å besitzt und eine Höhe, die so gewählt ist, daß der Flächeninhalt unter der Funktion 1 wird. Das Zentrum der Funktion befindet sich jeweils beim Abstand $d_{l,p}$ eines betrachteten Atom-Atom-Kontaktes. Dieser wird nun anteilsmäßig so über das umschließende Intervall [$r_i$, $r_i + dr$) und die angrenzenden Nachbarintervalle verteilt, wie es dem von den Intervallen überstrichenen Flächeninhalt der Dreiecksfunktion entspricht (Abb. 6).

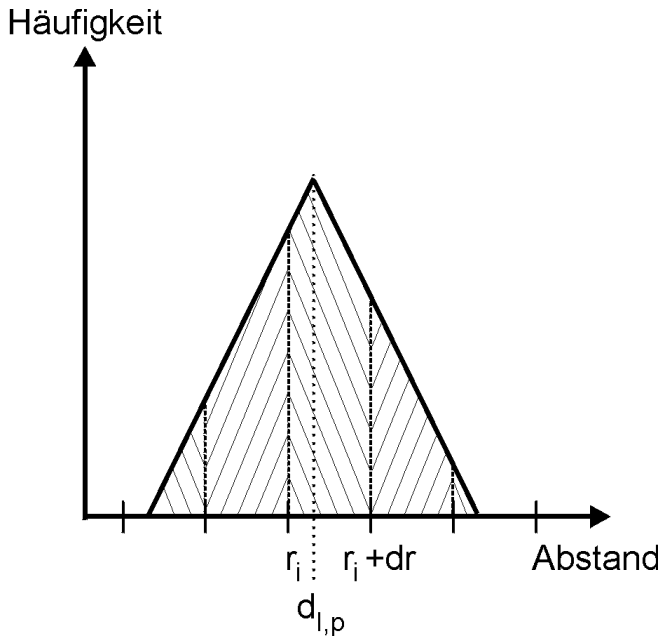



**Abb. 6:** Schematische Darstellung der zur Glättung der Rohdaten verwendeten gleichschenkligen Dreiecksfunktion. Eine in der Datenbank gemachte Beobachtung eines Atom-Atom-Kontaktes im Abstand $d_{l,p}$ wird anteilsmäßig so in dem diesen Abstand umschließenden Intervall [$r_i$, $r_i + dr$) und den angrenzenden Nachbarintervallen gezählt, wie es dem von den Intervallen überstrichen Flächeninhalt (jeweils schraffiert dargestellt) entspricht. Die Gesamtfläche unter der Kurve ist auf 1 normiert.

### 4.2.4 Behandlung von Verteilungen geringer Datenanzahl und Volumenkorrekturterm

Für eine Beurteilung der Signifikanz eines statistischen Paarpotentials muß die Anzahl von Atom-Atom-Kontakten über die gesamte Verteilung bzw. genaugenommen sogar pro Intervall einer Verteilung betrachtet werden. Nur im Grenzfall einer (unendlich) großen Daten-

menge kann davon ausgegangen werden, daß die *beobachteten* Paarverteilungsfunktionen auch den *tatsächlichen* entsprechen. An diesem Punkt stellt sich also die Frage, ob Verteilungen mit zu geringer Anzahl an Beobachtungen (etwa kleiner 500 Beispiele) aus der Berechnung der Wechselwirkungen zwischen Rezeptor und Ligand nach Gl. 15 herausgenommen werden sollten. Andererseits bedeutet *jede* Beobachtung eines Atom-Atom-Kontaktes (*d*,*l*) einen Zugewinn an Information für die Nettopräferenz $\Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$, der genutzt werden sollte. In der vorliegenden Arbeit wurden daher verschiedene Verfahren zur Behandlung von Verteilungen geringer Beobachtungszahl eingesetzt.

1. Sippl (Sippl, 1990) schlug ein von der Anzahl der Beobachtungen in jeder Verteilung $N_{\mathrm{T}(l),\mathrm{T}(p)}$ abhängiges „Mischen“ der normalisierten Paarverteilungsfunktion für ein spezielles Atompaar $g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$ mit der Paarverteilungsfunktion des Referenzzustandes $g(r_i)$ vor:

   $$g'_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) = \frac{1}{1+N_{\mathrm{T}(l),\mathrm{T}(p)}\sigma}\, g(r_i) + \frac{N_{\mathrm{T}(l),\mathrm{T}(p)}\sigma}{1+N_{\mathrm{T}(l),\mathrm{T}(p)}\sigma}\, g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) \qquad \textbf{Gl. 25}$$

   Mit adäquat gewähltem $\sigma > 0$ (etwa dem Reziproken der geforderten Mindestbeobachtungen je Verteilung) nähern sich so Verteilungen mit geringer Anzahl an Beobachtungen der Verteilung des Referenzzustandes an, Verteilungen mit großem $N_{\mathrm{T}(l),\mathrm{T}(p)}$ dagegen werden von $g(r_i)$ kaum beeinflußt. Sippl geht hierbei allerdings davon aus, daß $g(r_i)$ über alle Distanzbereiche statistisch signifikant ist.

2. In den hier verwendeten distanzabhängigen Verteilungen von Protein-Ligand-Atomkontakten sind – nicht überraschend – allerdings gerade Distanzbereiche unterhalb der Summe der van der Waals-Radien des betrachteten Atompaares schwach populiert. Die daraus herrührende statistische Unsicherheit tritt auch bei $g(r_i)$ in diesem Bereich auf. Eine von der Anzahl der Beobachtungen in einer Verteilung $N_{\mathrm{T}(l),\mathrm{T}(p)}$ unabhängige Reduzierung „lokaler Unsicherheit“ gelingt, indem eine konstante Größe $\chi > 0$ gleichermaßen zu $g_{T(l),T(p)}(r_i)$ und $g(r_i)$ addiert wird:

   $$g''_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) = g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) + \chi \quad ; \quad g''(r_i) = g(r_i) + \chi \qquad \textbf{Gl. 26}$$

   Eingesetzt in Gl. 11 geht damit für allgemein schwach populierte Distanzbereiche ($g_{T(l),T(p)}(r_i) \rightarrow 0$ *und* $g(r_i) \rightarrow 0$) die statistische Nettopräferenz $\Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$ gegen Null.

3. In dem unter Punkt 2. angegebenen Verfahren werden allerdings alle Verteilungen gleich behandelt, unabhängig von der Anzahl ihrer Beobachtungen. Kombination mit dem Ansatz von Sippl unter Punkt 1. umgeht dieses:

$$g'''_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) = \frac{1}{1+N_{\mathrm{T}(l),\mathrm{T}(p)}\sigma}(g(r_i)+\chi) + \frac{N_{\mathrm{T}(l),\mathrm{T}(p)}\sigma}{1+N_{\mathrm{T}(l),\mathrm{T}(p)}\sigma}(g_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)+\chi) \qquad \textbf{Gl. 27}$$

Natürlich ist in Gl. 11 jetzt auch $g''(r_i)$ aus Gl. 26 anstelle $g(r_i)$ zu verwenden.

4. Das unter 2. vorgestellte Verfahren kann auch für Größen $\chi$ angewendet werden, die sich mit dem Abstand ändern. Um in dem i.a. gut populierten Bereich oberhalb der Summe der van der Waals-Radien der betrachteten Atompaare ein kleines $\chi$ benutzen zu können, im Bereich der van der Waals-Abstoßung auftretende Unsicherheiten in den Daten aber ausreichend dämpfen zu können, wurde ein sich mit dem Abstand linear änderndes $\chi$ eingeführt:

$\chi = \chi_{Start}$ für $d_{l,p} > VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)}$

$$\chi = \chi_{Ende} + (d_{l,p} - (VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} - dist))\frac{\chi_{Start} - \chi_{Ende}}{dist}$$

für $VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} - dist \le d_{l,p} \le VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)}$

$\chi = \chi_{Ende}$ für $d_{l,p} < VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} - dist$. **Gl. 28**

Hierbei ist $VDW_{\mathrm{T}(x)}$ der van der Waals-Radius des Atoms *x*, *dist* ist die Strecke, innerhalb der $\chi$ von $\chi_{Start}$ auf $\chi_{Ende}$ steigt.

5. Ein analoges Verfahren wurde verwendet, um den Effekt schneller gegen Null gehender statistischer Nettopräferenzen für Distanzen $d_{l,p} > VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} + dist$ zu untersuchen:

$\chi = \chi_{Start}$ für $d_{l,p} \le VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} + dist$

$$\chi = \chi_{Start} + (d_{l,p} - (VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} + dist))\frac{\chi_{Ende} - \chi_{Start}}{r_{\max} - (VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} + dist)}$$

für $d_{l,p} > VDW_{\mathrm{T}(l)} + VDW_{\mathrm{T}(p)} + dist$ **Gl. 29**

*dist* steht hier für eine wählbare Distanz.

Bei der Bestimmung distanzabhängiger, radialer Paarverteilungsfunktionen von Protein Ligand-Kontakten muß beachtet werden, daß das zum Intervall *i* mit Grenzen $[r_i, r_i + dr)$ gehörige freie Kugelschalenvolumen $4\pi r_i^2 dr$ durch umgebende Atome des betrachteten Moleküls verringert wird und so nur noch eingeschränkt für wechselwirkende Atome des Gegenmoleküls zur Verfügung steht. Diesbezügliche Auswirkungen auf die Form von Paarvertei-

lungsfunktionen, gewonnen aus MD-Trajektorien von Kohlenhydrat-Wasser-Systemen, haben Astley *et al.* (Astley *et al.*, 1998) beschrieben. Um den Effekt für die hier abgeleiteten Paarpotentiale zu untersuchen, wurde für jeden Atom-Atom-Kontakt im Abstand $d_{l,p}$ das durch andere Ligandatome *unbesetzte* Volumen der Kugelschale mit $d_{l,p} \in [r_i, r_i + dr)$ bestimmt. Hierzu wurde das gesamte Kugelschalenvolumen in 81 Elemente mit dem Volumeninhalt

$$V_{Element} = dr \cdot r_i \; d\phi \cdot r_i \sin\theta \; d\theta \qquad \textbf{Gl. 30}$$

mit $dr$ = 0.1 Å, $d\phi = 2\pi / 9$ und $d\theta = \pi / 9$ unterteilt und das unbesetzte Volumen durch Summation über alle unbesetzten $V_{Element}$ ermittelt. Ein Kugelschalenelement gilt dabei als unbesetzt, wenn der Abstand seines Mittelpunktes zum nächsten Ligandatom größer als der van der Waals-Radius dieses Atoms ist. Die so ermittelten Volumina wurden für alle Ligandatomtypen jeweils für alle Intervalle $i$ innerhalb $[r_{min}, r_{max})$ gemittelt, man erhält $V_{unbesetzt}(\mathrm{T}(l), r_i)$. Anstatt in Gl. 13 jeweils durch das gesamte Kugelschalenvolumen $4\pi r_i^2 dr$ zu teilen, wird hierzu nun $V_{unbesetzt}(\mathrm{T}(l), r_i)$ verwendet.

## *4.3 Von der Lösemittel-zugänglichen Oberfläche abhängige statistische Präferenzen*

### 4.3.1 Definition

Im Zusammenhang mit der Anwendung von (distanzabhängigen) Paarpotentialen im Rahmen der Proteinfaltungsvorhersage ist bekannt, daß durch Lösemittel-bedingte Effekte unterschiedlich durch diese repräsentiert werden können (Godzik *et al.*, 1995; Skolnick *et al.*, 1997). Für die Wahl der oberen Grenze der Paarpotentiale wurde dieses in Kap. 4.2.3 diskutiert. Darüber hinaus ist zunächst von Sippl selbst (Sippl, 1993) und später auch von Miyazawa und Jernigan (Miyazawa & Jernigan, 1999) angemerkt worden, daß bei der Anwendung von Paarpotentialen nach dem Sippl-Formalismus Lösemitteleffekte nicht ausreichend berücksichtigt werden. Um dennoch diese für Protein-Ligand-Komplexe wichtigen Wechselwirkungen einzubeziehen, wird hier zusätzlich ein von der Lösemittel-zugänglichen Oberfläche (***S**olvent **A**ccessible **S**urface*, SAS) von Protein- und Ligandatomen abhängiges wissensbasiertes Einteilchenpotential abgeleitet.

Für *getrennt* betrachtete Atome $x$ der Ligand- oder Proteinmoleküle ($L$ bzw. $P$) ergibt sich die statistische Präferenz $\Delta W^{L/P}_{\mathrm{T}(x)}(S_i^{Kpl}, S_j^{Frei})$ als Funktion der SAS im komplexierten Zustand $SAS_x^{Kpl}$ innerhalb eines Intervalls $i$ mit den Grenzen $[S_i^{Kpl}, S_i^{Kpl} + dS)$ und der SAS im freien,

d.h. ungebundenen Zustand $SAS_x^{Frei}$ innerhalb eines Intervalls $j$ mit den Grenzen $[S_j^{Frei}, S_j^{Frei} + dS)$ zu

$$\Delta W_{T(x)}^{L/P}(S_i^{Kpl}, S_j^{Frei}) = W_{T(x)}^{L/P}(S_i^{Kpl}) - W_{T(x)}^{L/P}(S_j^{Frei}) = -\ln \frac{g_{T(x)}^{L/P}(S_i^{Kpl})}{g_{T(x)}^{L/P}(S_j^{Frei})} \quad \textbf{Gl. 31}$$

Beschränkt man $SAS_x^{Kpl/Frei}$ generell auf ein Intervall $[S_{min}, S_{max})$, so ergibt sich bei festgelegter Intervallweite $dS$ der Wert $S_i^{Kpl/Frei}$ zu

$$S_i^{Kpl/Frei} = S_{\min} + \left\lfloor \frac{SAS_x^{Kpl/Frei} - S_{\min}}{dS} \right\rfloor dS \quad \textbf{Gl. 32}$$

$g_{T(x)}^{L/P}(S_i^{Kpl})$ ist hierbei die normalisierte Verteilungsfunktion bezüglich der SAS für Atome des Typs T($x$) entweder der Liganden oder der Proteine im an das Gegenmolekül gebundenen Zustand, wohingegen $g_{T(x)}^{L/P}(S_j^{Frei})$ die analoge Verteilungsfunktion bzgl. der SAS im freien, d.h. solvatisierten Zustand ist.

$$g_{T(x)}^{L/P}(S_i^{Kpl/Frei}) = \frac{N_{T(x)}^{L/P}(S_i^{Kpl/Frei})}{\sum_{i \in I} N_{T(x)}^{L/P}(S_i^{Kpl/Frei})} \quad \textbf{Gl. 33}$$

Die Summation läuft hierbei jeweils über alle Intervalle $i$ der diskreten Verteilungen. Hierbei ist zu beachten, daß die normierte Verteilungsfunktion im Nenner von Gl. 31, $g_{T(x)}^{L/P}(S_j^{Frei})$, nicht – im Gegensatz zu der analogen Formel Gl. 11 bei den Paarpotentialen - über alle Atomtypen gemittelt wird, sondern sich auf einen spezifischen Atomtyp bezieht. $\Delta W_{T(x)}^{L/P}(S_i^{Kpl}, S_j^{Frei})$ beschreibt also den Beitrag, der durch Unterschiede in der SAS eines Atoms im freien bzw. komplexierten Zustand entsteht.

Die Anzahl der Atome $x$ – getrennt nach Protein- und Ligandatomen - des Typs T($x$) innerhalb des Intervalls $[S_i^{Kpl/Frei}; S_i^{Kpl/Frei} + dS)$, bezeichnet als $N_{T(x)}(S_i^{Kpl/Frei})$, ergibt sich durch Auszählen ihrer Auftrittshäufigkeiten über alle Atome $L_k$ bzw. $P_k$ für Ligand- oder Proteinmoleküle eines Komplexes $k$ für alle (nativen) Komplexe $K_n$ in der betrachteten Datenbank.

$$N_{T(x)}^{L/P}(S_i^{Kpl/Frei}) = \sum_{k \in K_n} \sum_{x \in L_k / P_k} \delta(SAS_x^{Kpl/Frei}, S_i^{Kpl/Frei}) \quad \textbf{Gl. 34}$$

Die Delta-Funktion $\delta(SAS_x^{Kpl/Frei}, S_i^{Kpl/Frei})$ ergibt 1, wenn $SAS_x^{Kpl/Frei} \in [S_i^{Kpl/Frei}, S_i^{Kpl/Frei} + dS)$, sonst 0. Beachtet werden hier allerdings nur Atome, für die $SAS_x^{Kpl} < SAS_x^{Frei}$ gilt, d.h. die während der Komplexbildung (partiell) vergraben werden.

Der auf die wissensbasierten Einteilchenpotentiale zurückgehende Anteil der Wechselwirkungen für eine gegebene Konfiguration von Protein und Ligand im Komplex $k$ resultiert damit zu

$$\Delta W_{Einzel} = \sum_{l \in L_k} \Delta W^{L}_{T(l)}(S_i^{Kpl}, S_j^{Frei}) + \sum_{p \in P_k} \Delta W^{P}_{T(p)}(S_k^{Kpl}, S_l^{Frei}) \qquad \textbf{Gl. 35}$$

mit $SAS_i^{Kpl} \in [S_i^{Kpl}, S_i^{Kpl} + dS)$ und analog für alle anderen Oberflächen.

### 4.3.2 Berechnung der Lösemittel-zugänglichen Oberfläche

Die Berechnung der für das Lösemittel zugänglichen Oberfläche eines Atoms wird mit einem schnellen, approximierenden, auf einem kubischen Gitter beruhenden Algorithmus durchgeführt. Böhm (Böhm, 1994) hat für eine ähnliche Vorgehensweise zeigen können, daß die damit für die Oberfläche erhaltenen Werte gut mit denen aus dem MS-Programm von Connolly (Connolly, 1983) korrelieren. Für die Ermittlung der SAS im gebundenen Zustand ist zu beachten, daß die Oberflächenbereiche *polarer* Atome (d.h. aller Sauerstoff und Stickstoffatome) dann als *nicht* vergraben gelten, wenn sie sich in einer *polaren* Umgebung des Gegenmoleküls befinden (Koehl & Delarue, 1994). Dies beruht auf der Annahme, daß ein Beitrag zur Freien Enthalpie bedingt durch Lösemitteleffekte für *polare* Atome nur dann erwartet werden kann, wenn sie aus dem polaren Lösemittel Wasser in eine *unpolare* Molekülumgebung gelangen. Für einen Transfer polarer Gruppen zwischen polaren Umgebungen wird der Lösemittelbeitrag daher vernachlässigt.

Als Näherung wird in dieser Arbeit die Konformation des im Komplex mit dem Protein gebundenen Liganden ebenfalls verwendet, um die SAS für den freien, d.h. solvatisierten Zustand zu bestimmen. Mit diesem Modell werden daher Konformationsänderungen zwischen freiem und gebundenem Zustand - bedingt etwa durch einen „hydrophoben Kollaps“ des Liganden (Wiley & Rich, 1993) (s.a. Kap. 2.2.2) - nicht berücksichtigt.

Der im folgenden angegebene Algorithmus (Alg. 1) zur Bestimmung der SAS im freien bzw. an das Protein gebundenen Zustand für einzelne Atome des Liganden kann analog auch zur Bestimmung der SAS von Proteinatomen angewendet werden. Eingabeparameter sind zum einen die Vektoren $P = (x, y, z, t)^{n_P}$ und $L = (x, y, z, t)^{n_L}$ der Nichtwasserstoff-Atome von Protein (Anzahl: $n_P$) und Ligand (Anzahl: $n_L$), wobei $x, y, z \in$ IR die kartesischen Koordinaten und $t$ ein Atomtyp aus einer Menge $T$ ist. Ein kubisches Gitter $G = \{(x, y, z, f) \mid x, y, z \in \text{IR}, f \in \{leer, lig, prot\}\}$ mit den Koordinaten $x$, $y$, $z$ der Mittelpunkte der Würfel und einer Variablen zur Kennzeichnung, ob der jeweilige Würfel *leer*, *lig*andbesetzt oder *prot*einbesetzt ist, umschließt den Liganden. Zur analogen Bestimmung der SAS von Proteinatomen wird das Gitter um den Bereich der Bindetasche gelegt. Die hier

verwendete Gitterweite beträgt 1 Å. Ausgegeben wird ein Vektor $SAS^{Kpl} \in \mathrm{IN}^{n_L}$ sowie ein Vektor $SAS^{Frei} \in \mathrm{IN}^{n_L}$, die die zu den jeweiligen Ligandatomen dazugehörigen Anzahlen von angrenzenden Oberflächenwürfeln enthalten.

Nach der Initialisierung aller Einträge in $SAS^{Kpl}$ und $SAS^{Frei}$ (Zeile 1) werden in einem ersten Schritt alle Würfel, deren Zentren näher als die Summe aus dem van der Waals-Radius eines Ligandatoms (s.a. Tab. 2, S. 65) plus dem Radius eines Wassermoleküls (1.4 Å) an diesem Ligandatom liegen, als *lig*andbesetzt markiert (Zeile 2 - 8). Nach einem analogen Kriterium werden alle Kuben in der Nähe von Proteinatomen anschließend als *prot*einbesetzt markiert, sofern die beiden nächsten Atome von Ligand bzw. Protein nicht polar (d.h. Sauerstoff oder Stickstoff) sind (Zeile 9 - 23). Für alle nicht *lig*andbesetzten Würfel, die benachbart zu einem anderen *lig*andbesetzten Kubus sind, wird das nächstliegende Ligandatom bestimmt und dessen $SAS^{Frei}$ um eins erhöht (Zeile 24 - 27). Ist der betrachtete Würfel auch nicht *prot*einbesetzt, so wird auch $SAS^{Kpl}$ um eins erhöht (Zeile 28 - 32).

### 4.3.3 Wahl der Intervallparameter, Glättung der Rohdaten und Behandlung von Verteilungen geringer Datenanzahl

Bei der Berechnung der Lösemittel-zugänglichen Oberfläche nach dem im vorherigen Kapitel angegebenen Algorithmus entspricht die vollständige Vergrabung eines Atoms der unteren Grenze $S_{min}$ von Null zu einem Atom zugeordneten Kuben. Für die obere Grenze ($S_{max}$) wird eine Anzahl von 40 Kuben festgelegt, wobei selten auftretende Fälle mit diesbezüglich größeren Werten auf diesen Grenzwert zurückgesetzt werden. Die Intervallweite *dS* wird auf einen Kubus festgelegt.

Die Glättung der Rohdaten erfolgt analog zu dem in Kap. 4.2.3 beschriebenen Vorgehen unter Verwendung einer gleichschenkligen Dreiecksfunktion. Ihre Breite an der Basis wird auf 3 Kuben festgelegt. Die Fläche unter der Kurve wird durch die Wahl ihrer Höhe auf 1 normiert.

Für die Behandlung der Verteilungen geringer Datenanzahl wird die in Kap. 4.2.4 unter Punkt 2 beschriebene Methode gewählt (mit den in Gl. 33 angegebenen Verteilungsfunktionen anstelle der in Gl. 26 aufgeführten), die eine Reduzierung „lokaler Unsicherheit“ bewirkt (hinsichtlich der Wahl des $\chi$-Wertes s.a. Kap. 5.2).

SAS-Approximation für Ligandatome

EINGABE:

| | |
|---|---|
| $P$: | Vektor von Proteinatomen |
| $L$: | Vektor von Ligandatomen |
| $G$: | Kubisches Gitter |

AUSGABE:

| | |
|---|---|
| $SAS^{Kpl}$: | Vektor der SAS der Ligandatome im gebundenen Zustand |
| $SAS^{Frei}$: | Vektor der SAS der Ligandatome im freien Zustand |

ALGORITHMUS:

```
1.  ∀i = 1, ..., n_L: SAS^Kpl[i] ← 0, SAS^Frei[i] ← 0
2.  for ∀g ∈ G do
3.      g.f ← leer
4.      for i ← 1 to n_L do
5.          mindist ← VDW(L[i].t) + 1.4
6.          dist ← |(g.x – L[i].x, g.y – L[i].y, g.z – L[i].z)|
7.          if dist ≤ mindist then
8.              g.f ← lig
9.          else
10.             for j ← 1 to n_P do
11.                 mindist ← VDW(P[j].t) + 1.4
12.                 dist ← |(g.x – P[j].x, g.y – P[j].y, g.z – P[j].z)|
13.                 if dist ≤ mindist then
14.                     k ← NEAREST_ATOM(g, L)
15.                     l ← NEAREST_ATOM(g, P)
16.                     if NONPOLAR(L[k].t) or NONPOLAR(P[l].t) then
17.                         g.f ← prot
18.                     fi
19.                 fi
20.             od
21.         fi
22.     od
23. od
24. for ∀g ∈ G: g.f ≠ lig do
25.     if ∃h ∈ G: h.f = lig ∧ NEIGHBORS(g,h) then
26.         k ← NEAREST_ATOM(g, L)
27.         SAS^Frei[k] ← SAS^Frei[k] + 1
28.         if g.f = leer then
29.             SAS^Kpl[k] ← SAS^Kpl[k] + 1
30.         fi
31.     fi
32. od
```

**Alg. 1:** Bestimmung der SAS für Ligandatome im freien, d.h. vollständig solvatisierten Zustand und gebunden an ein Protein. Die SAS wird approximiert unter Verwendung eines Gitteralgorithmus. Für Erläuterungen siehe Text.

### *4.4 Ableitung der Potentiale*

Die für die distanzabhängigen Paar- und die von der Lösemittel-zugänglichen Oberfläche abhängigen Einteilchenpotentiale benötigten Daten werden aus der Rezeptor-Ligand-Datenbank ReLiBase (Hemm *et al*., 1995; Hendlich, 1998) extrahiert. ReLiBase in der Version vom Januar 1998 enthält 6026 Protein-Ligand-Komplexe, die auch in der PDB (Bernstein *et al.*, 1977) enthalten sind. Im Unterschied zur PDB verfügt ReLiBase allerdings über eine hierarchische, objektorientierte Datenstruktur sowie eine Benennung der Atom- und Bindungstypen der enthaltenen Liganden, die der SYBYL-Notation (SYBYL) folgt. Beide Aspekte sind für die hier vorgestellte Verwendung von Bedeutung.

Zur Ableitung der Potentiale wurden in zwei Auswertungen nur kristallographisch bestimmte Protein-Ligand-Komplexe mit einer Auflösung $\leq 2.0$ Å bzw. $\leq 2.5$ Å verwendet. Damit soll der Einfluß der Qualität der verwendeten Daten auf den Verlauf der Potentiale ermittelt werden. Komplexe mit kovalent gebundenen Liganden oder mit Liganden mit weniger als sechs oder mehr als 50 Nichtwasserstoffatomen wurden aus der Datenmenge ausgeschlossen. Dieses letzte Kriterium dient der Beschränkung auf Liganden mit einer Größe, die typisch für Wirkstoffmoleküle ist. Begründet durch Aspekte der Bioverfügbarkeit liegt eine obere Grenze bei etwa 600 Dalton für ein typisches organisches Molekül (Lipinski *et al*., 1997; Pfeifer *et al*., 1995).

Zusätzlich werden alle Komplexe ausgeschlossen, die nachher zur Validierung der Potentiale verwendet werden, um jegliche Redundanzen oder ein Übertrainieren der Methode zu vermeiden. Außerdem wurden in zusätzlichen Auswertungen für einzelne Protein-Ligand-Komplexe (129l, 1rga, 4phv) solche mit einer Sequenzhomologie von größer als 30 % dazu ebenfalls aus der zur Ableitung verwendeten Datenmenge ausgeschlossen, um den Einfluß der Zusammensetzung der verwendeten Daten auf den Potentialverlauf zu untersuchen.

Bei den im folgenden angegebenen Atomtypen (Tab. 2), für die die Paar- und Einteilchenpotentiale bestimmt wurden, ist zu beachten, daß prosthetische Gruppen wie Eisen-Häm-Zentren, Flavin-adenin-mononukleotid, Flavin-adenin-dinukleotid, Nikotinamid-adenin-dinukleotid, N-Acetyl-glucosamin und Eisenschwefelcluster nicht als Ligand, sondern als Teil des Proteins angesehen wurden. Im Rahmen des oben vorgestellten Formalismus‘ für die Paarpotentiale läßt sich diese Vorgehensweise natürlich auch auf Wassermoleküle erweitern, wodurch sich auch wasservermittelte Wechselwirkungen beschreiben lassen. Innerhalb dieser Arbeit wurde hierauf jedoch nur in Form von Vorversuchen eingegangen (s. a. Kap. 6.2). Aufgrund der Datenlage ist an einigen Stellen eine von der SYBYL-Notation abweichende Zusammenfassung einzelner Subtypen unter einem übergeordneten Typ notwendig.

Ferner werden keine Potentiale für den Typ „Wasserstoff“ abgeleitet. Bedingt durch die geringe Auflösung von kristallographisch bestimmten Proteinstrukturen können in den meisten Fällen die Positionen von beweglichen Wasserstoffatomen in diesen Strukturen nicht angegeben werden (s. a. Kap. 2.1). Außerdem ist eine *a priori* Angabe von Protonierungszuständen von an Proteine gebundenen Liganden nur beschränkt möglich, da sich der $pK_a$-Wert funktioneller Gruppen durch Einfluß der Proteinumgebung z. T. drastisch verändern kann (Antosiewicz *et al.*, 1996; Bashford & Karplus, 1990; van Vlijmen *et al.*, 1998). Unter diesen Umständen beruht die Beschränkung auf Nichtwasserstoffatome auf der Annahme, daß die abgeleiteten statistischen Präferenzen diese Einflüsse im Zusammenhang mit der Beschreibung von Wechselwirkungen zwischen funktionellen Gruppen implizit berücksichtigen.

Die für die vorgestellten Atomtypen verwendeten van der Waals-Radien (Tab. 2) zur Berechnung der SAS gelten für „vereinigte Atome“ (*united atoms*), d.h. der Radius für ein Schweratom berücksichtigt eine zusätzliche Ausdehnung durch gebundene Wasserstoffatome. Sie wurden dem Tripos-Kraftfeld (Clark *et al.*, 1989) entnommen. Dabei wurden alle Sauerstoff- und Stickstoffradien um 0.2 Å verringert, um einer Verkürzung des Atom-Atom-Abstandes bei Ausbildung einer Wasserstoffbrücke zwischen diesen Atomen Rechnung zu tragen (Li & Nussinov, 1998). Der Radius des Metallatoms ist nachträglich eingefügt worden und wurde so gewählt, daß die Summe aus ihm und dem Radius eines wechselwirkenden Atoms der Distanz beim Häufigkeitsmaximum in den radialen Paarverteilungen entspricht. Bedingt durch den koordinativen Charakter der Metall-X-Bindung (X: insbesondere Stickstoff, Sauerstoff, Schwefel) führt dieses Vorgehen zu dem im Vergleich geringen Radius.

**Tab. 2:** Zur Ableitung der Paar- und Einteilchenpotentiale verwendete Atomtypen für Ligand- und Proteinatome mit den dazugehörigen van der Waals-Radien

| **Bedeutung**[a)] | **Atomtyp**[b)] | $r_{vdW}$[c)] |
|---|---|---|
| $sp^3$-hybrid. Kohlenstoff | C.3 | 1.70 |
| $sp^2$- und $sp^1$-hybrid. Kohlenstoff | C.2 (C.1) | 1.70 |
| Kohlenstoff in aromatischen Ringen | C.ar | 1.70 |
| Kohlenstoff in Amidino- und Guanidino-Gruppen | C.cat | 1.70 |
| $sp^3$-hybrid. Stickstoff | N.3 (N.4) | 1.35 |
| Stickstoff in aromatischen Ringen und $sp^2$-hybrid. Stickstoff | N.ar (N.2) | 1.35 |
| Stickstoff in Amidbindungen | N.am | 1.35 |
| Stickstoff in Amidino- und Guanidino-Gruppen | N.pl3 | 1.35 |
| $sp^3$-hybrid. Sauerstoff | O.3 | 1.32 |
| $sp^2$-hybrid. Sauerstoff | O.2 | 1.32 |
| Sauerstoff in Carboxylgruppen | O.co2 | 1.32 |
| $sp^3$- und $sp^2$-hybrid. Schwefel | S.3 (S.2) | 1.60 |
| $sp^3$-hybrid. Phosphor | P.3 | 1.60 |
| Fluor | F | 1.27 |
| Chlor | Cl | 1.55 |
| Brom | Br | 1.65 |
| Calcium, Zink, Eisen, Nickel | Met | 0.40[d)] |

a) Bei der hier verwendeten Unterteilung der Atomtypen werden $sp^2$-hybridisierte Kohlenstoff-, Stickstoff- und Sauerstoffatome weitergehend nach ihrem Auftreten in funktionellen Gruppen unterschieden. b) Die Benennung der Atomtypen folgt der SYBYL-Notation (SYBYL). In Klammern sind die ebenfalls unter die aufgeführte Bedeutung fallenden ursprünglichen SYBYL-Typen angegeben. c) Die van der Waals-Radien sind dem Tripos-Kraftfeld (Clark *et al.*, 1989) entnommen und in Å angegeben. Die Radien von Sauerstoff- und Stickstoffatomen wurden um 0.2 Å verringert (Li & Nussinov, 1998). d) Der Radius von Metall wurde nachträglich eingeführt. Für Erläuterungen siehe Text.

### *4.5 Bewertungsfunktion für Protein-Ligand-Wechselwirkungen*

Unter der Annahme, daß die durch die distanzabhängigen Paarpotentiale sowie die von der Lösemittel-zugänglichen Oberfläche abhängigen Einteilchenpotentiale beschriebenen Wechselwirkungen jeweils voneinander unabhängig und für verschiedene molekulare Umgebungen gleichermaßen gültig sind, ergibt sich die gesamte statistische Präferenz für eine gegebene Protein-Ligand-Konfiguration durch Summation über alle individuellen Beiträge. Hieraus

folgt für einen Komplex *k* mit $L_k$ Ligandatomen und $P_k$ Proteinatomen unter Verwendung von Gl. 15 und Gl. 35:

$$\Delta W = \gamma \sum_{l \in L_k} \sum_{p \in P_k} \Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i) + (1-\gamma)\left[ \sum_{l \in L_k} \Delta W^{L}_{\mathrm{T}(l)}(S_i^{Kpl}, S_j^{Frei}) + \sum_{p \in P_k} \Delta W^{P}_{\mathrm{T}(p)}(S_k^{Kpl}, S_l^{Frei}) \right] \quad \textbf{Gl. 36}$$

Bei der Ermittlung der Präferenz werden Kofaktoren und Metallatome als Teil des Proteins angesehen, wohingegen Wassermoleküle – bis auf im einzelnen beschriebene Fälle – ausgeschlossen werden. Dabei gilt für den Abstand $d_{l,p}$ zwischen *l* und *p*, daß er jeweils innerhalb des Intervalls $[r_i, r_i + dr)$ liegt und für die Lösemittel-zugänglichen Oberflächen $SAS_x^{Kpl/Frei}$ im komplexierten (*Kpl*) bzw. ungebundenen (*Frei*) Zustand der Atome *x* von Protein- oder Ligand, daß sie im Intervall $[S_i^{Kpl/Frei}, S_i^{Kpl/Frei} + dS)$ liegen. $\gamma$ ist ein empirisch zu bestimmender Wichtungsparameter. Ansonsten gelten die Bezeichnungen wie in Kap. 4.2.1 und 4.3.1 eingeführt.

Außer den in Gl. 36 aufgeführten gehen keine weiteren Beiträge in die Bewertung ein, wie etwa Einflüsse durch Änderungen der Translations- und Rotationszustände oder Einschränkungen der Flexibilität beider Reaktionspartner. Unter Hinweis auf Untersuchungen von Boström *et al.* (Bostrom *et al.*, 1998), wonach sich die Energien von Konformationen von Liganden im Protein-gebundenen sowie im vollständig solvatisierten Zustand um weniger als 12 kJ / mol unterscheiden, wird außerdem von der Berücksichtigung von intramolekularen Wechselwirkungsenergien (v.a. von van der Waals und Torsionspotentialen) abgesehen. Dies gilt umso mehr, als die zu untersuchenden Ligandkonformationen entweder aus Kristallstrukturen stammen oder aber von Dockingprogrammen wie FlexX (Rarey *et al.*, 1996a), DOCK (Kuntz *et al.*, 1982) oder GOLD (Jones *et al.*, 1995) erzeugt wurden, die Regeln zur Erzeugung günstiger Konformationen berücksichtigen. Beachtet man weiter die Steilheit kraftfeldbasierter Potentiale, so müßte für eine verläßliche Energieevaluierung außerdem eine Geometrieoptimierung des in der Bindetasche plazierten Liganden vorgeschaltet werden.

### *4.6 Bewertung der Güte der Bewertungsfunktion*

Die Bewertung der Güte der mit Gl. 36 erhaltenen Funktion erfolgt unter Anwendung von drei Kriterien, die die in Kap. 1 gegebenen Problemstellungen zur Struktur- und Affinitätsvorhersage von Protein-Ligand-Komplexen und die Fragestellungen im Verlauf eines strukturbasierten Designs widerspiegeln.

### 4.6.1 Bestimmung nativ-ähnlicher Ligandenkonformationen

Als erstes Kriterium für die Güte einer Bewertungsfunktion soll ihre Fähigkeit zur Unterscheidung zwischen einer nativen Anordnung von Protein und Ligand und einer großen Anzahl falsch angeordneter Konfigurationen herangezogen werden. Dies zielt auf die Auswahl der korrekten Geometrie eines Rezeptor-Ligand-Komplexes aus einer Menge gegebener Alternativen, wobei jeweils nur *ein* Protein und *ein* Ligand betrachtet werden. Als Maß für die Abweichung einer erzeugten Protein-Ligand-Konfiguration von der Geometrie der nativen, experimentell bestimmten Kristallstruktur wird i.a. die Wurzel aus der mittleren quadratischen Abweichung in den kartesischen Koordinaten (*rmsd*) der Nichtwasserstoffatome $L_k$ eines Liganden verwendet (Kramer *et al.*, 1999):

$$rmsd = \sqrt{\frac{\sum_{l \in L_k} \left( \left| \vec{X}_l^{\,nativ} - \vec{X}_l^{\,generiert} \right| \right)^2}{\left\| L_k \right\|}} \qquad \textbf{Gl. 37}$$

Für eine Anzahl von Testfällen zeigt sich, daß generierte Konfigurationen bis zu einem *rmsd*-Wert von 2.0 Å dem nativen Bindungsmodus ausreichend ähneln. Darauf basierend werden innerhalb einer Menge von erzeugten Protein-Ligand-Geometrien $S$ solche als „gut“ (d.h. zu $S_{2.0}$ gehörig) definiert, die einen *rmsd*-Wert < 2.0 Å von ihrer Kristallstruktur aufweisen. Dies stimmt auch mit von Jones *et al.* (Jones *et al.*, 1997) und Kramer *et al.* (Kramer *et al.*, 1999) gewählten Kriterien überein. Die Kristallstruktur selbst soll mit $s_{ideal}$ bezeichnet werden. Damit lassen sich drei Subkriterien mit ansteigenden Anforderungen an die Bewertungsfunktion definieren.

I. *Eine der „guten“ Lösungen soll nach der Bewertung einen der ersten Z Ränge einnehmen:*

$$\exists s_{2.0} \in S_{2.0} : \mathrm{Rang}(s_{2.0}) < Z \qquad \textbf{Gl. 38}$$

Dieses Kriterium ist im Rahmen von virtuellen Screening-Verfahren in anbetracht der dabei zu untersuchenden Datenmengen (für jede Protein-Ligand-Kombination bis zu $Z$ Konfigurationen) allerdings nicht sinnvoll anwendbar.

II. *Eine der „guten“ Lösungen soll nach der Bewertung den ersten Rang einnehmen:*

$$\exists s_{2.0} \in S_{2.0} \quad \forall s \in S \setminus \{s_{2.0}\} : \mathrm{Rang}(s_{2.0}) < \mathrm{Rang}(s) \qquad \textbf{Gl. 39}$$

III. *Die experimentell bestimmte Kristallstruktur soll nach der Bewertung auf dem ersten Rang liegen:*

$$\forall s \in S \setminus \{s_{\mathrm{ideal}}\} : \mathrm{Rang}(s_{\mathrm{ideal}}) < \mathrm{Rang}(s) \qquad \textbf{Gl. 40}$$

Hinter diesem Kriterium steckt die Annahme, daß die experimentell bestimmte Geometrie dem wahrscheinlichsten Zustand entspricht und die Bewertungsfunktion – abgeleitet an einer Vielzahl experimenteller Strukturen – diese Information beinhaltet. Allerdings ist zu beachten, daß auch die Geometrie der Referenz $s_{ideal}$ von der begrenzten Auflösung bei Proteinkristallen und den damit einhergehenden Unsicherheiten in den Koordinaten der Atompositionen betroffen ist. Aus diesem Grund wird zusätzlich ein schwächeres Kriterium formuliert. Hierbei wird gefordert, daß *nach der Bewertung besser als die Kristallstruktur plazierte Protein-Ligand-Strukturen nur solche aus der Menge der „guten" sein dürfen:*

$$\forall s \in S \setminus (S_{2.0} \cup \{s_{ideal}\}) : \text{Rang}(s_{ideal}) < \text{Rang}(s) \qquad \textbf{Gl. 41}$$

Die Bewertung der entwickelten wissensbasierten Funktion hinsichtlich ihrer Eignung zur Vorhersage der Geometrie von Protein-Ligand-Komplexen wird im Rahmen dieser Arbeit mit den in Gln. 39, 40 und 41 beschriebenen Kriterien vorgenommen.

### 4.6.2 Priorisierung unterschiedlicher Liganden gegenüber einem Protein, Vorhersage der Selektivität von Liganden in Bezug auf mehrere Rezeptoren und Berechnung von Bindungsaffinitäten

Die Untersuchung der in Gl. 36 entwickelten Funktion hinsichtlich ihrer Eignung zur Reihung *mehrerer* Liganden gegenüber *einem* Protein oder gar zur Selektivitätsvorhersage *mehrerer* Liganden in Bezug auf *mehrere* Rezeptoren ergibt nicht nur zwei weitere Kriterien zur Bewertung ihrer Güte, sondern auch Hinweise auf die Gültigkeit von im Verlauf ihrer Ableitung getroffenen Annahmen. So hängt die Anwendbarkeit der Funktion für paarweise verschiedene Protein-Ligand-Komplexe maßgeblich davon ab, wie ähnlich die bei der Ableitung der Paarpotentiale verbleibende Konstante für unterschiedliche Komplexe ist (s.a. Kap. 4.2.1) und wie schwer die Vernachlässigung expliziter Terme für die Einschränkung von Mobilität und Flexibilität wiegt. Andererseits sollte das Auftreten von Korrelationen zwischen berechneten Bewertungen und experimentell bestimmten Bindungsaffinitäten zumindest einen Hinweis geben, daß entweder die vernachlässigten Beiträge bei der Komplexbildung nur eine untergeordnete Rolle spielen oder aber implizit durch den Formalismus der Potentialableitung mit berücksichtigt wurden.

Die mit der wissensbasierten Bewertungsfunktion erhaltenen Präferenzen $\Delta W$ werden durch Multiplikation mit einem zu bestimmenden Parameter $c_S$ mit experimentell bestimmten

Bindungsaffinitäten (hier in Form von $pK_i$-Werten) in Beziehung gesetzt. $c_S$ ist die Steigung einer abschnittslosen Korrelationsgeraden, die unter Anwendung eines *regula falsi*-Verfahrens (Harris & Stocker, 1998) durch Skalierung der $\Delta W$-Werte iterativ so bestimmt wird, daß deren Standardabweichung von der Korrelationsgeraden gleich der Standardabweichung der $pK_i$-Werte davon ist.

$$pK_i = c_S \, \Delta W \qquad \textbf{Gl. 42}$$

Die Allgemeingültigkeit dieser Beziehung hängt davon ab, ob der ermittelte Skalierungsparameter $c_S$ auf verschiedene Protein-Ligand-Systeme übertragbar ist. Um eine Abhängigkeit der Methode von der Generierungsart der eingesetzten Rezeptor-Ligand-Geometrien zu untersuchen, werden nicht nur Datensätze von kristallographisch bestimmten Komplexgeometrien zur Validierung verwendet, sondern auch solche, die mit Docking-Verfahren erzeugt wurden.

Eine mit der Vorhersage von Bindungsaffinitäten bzw. der Priorisierung unterschiedlicher Liganden gegenüber *einem* Protein unmittelbar in Zusammenhang stehende Anwendung der entwickelten Bewertungsfunktion ergibt sich bei Ansätzen zum virtuellen Screening. Bei der Validierung entwickelter Vohersagemethoden werden hierzu Datensätze von Liganden mit bekannter Aktivität gegenüber einem Rezeptor (sog. „Aktive“) mit solchen Ligandensätzen kombiniert, für die zwar keine Aktivität bzgl. des Rezeptors bekannt ist, die aber als Datengrundlage für ein real durchzuführendes virtuelles Screening-Verfahren herangezogen werden (sog. „Inaktive“). Nach der Erzeugung möglicher Rezeptor-Ligand-Anordungen mit Docking-Verfahren werden diese dann hinsichtlich ihrer zu erwartenden Affinität gemäß Gl. 36 bewertet und sortiert. Ausgehend von dieser Reihung werden Anreicherungsfaktoren (*enrichment factors, ef*) berechnet, die eine Funktion des schon betrachteten Anteils $F$ der insgesamt prozessierten Liganden sind:

$$ef(F) = \frac{\|Aktive(F)\| \; / \; \left(\|Aktive(F)\| + \|Inaktive(F)\|\right)}{\|Aktive_{gesamt}\| \; / \left(\|Aktive_{gesamt}\| + \|Inaktive_{gesamt}\|\right)} \qquad \textbf{Gl. 43}$$

||...|| steht hierbei jeweils für „Anzahl von ...“. Der Index *gesamt* umfaßt alle Aktiven bzw. Inaktiven in der gesamten Datenbank. Für $F \rightarrow 1$ folgt auch $ef(F) \rightarrow 1$, d.h. eine vollständige Betrachtung der Datenbank ohne Kenntnis einer Priorisierung käme dem Ergebnis einer zufälligen Auswahl von Liganden gleich. Werte von $ef(F) > 1$ ergeben sich dagegen, wenn der herausgefilterte Datenbankanteil $F$ mehr Aktive enthält als vom Datenbankmittelwert her zu erwarten wären. Bei Gl. 43 ist zu beachten, daß die erhaltenen $ef(F)$-Werte nicht skaliert sind, d.h. vom Verhältnis $||Aktive_{gesamt}|| \; / \; ||Inaktive_{gesamt}||$ abhängen. Sie sollten daher jeweils mit dem maximal erreichbaren Anreicherungsfaktor $ef_{max}$ verglichen werden:

$$ef_{\max} = \frac{\left\|Aktive_{gesamt}\right\| + \left\|Inaktive_{gesamt}\right\|}{\left\|Aktive_{gesamt}\right\|} \qquad \textbf{Gl. 44}$$

## *4.7 Untersuchung der impliziten Berücksichtigung von Direktionalität in Paar-Potentialen*

Die Stärke von Wechselwirkungen - besonders zwischen polaren funktionellen Gruppen aber auch z.B. zwischen aromatischen Systemen - hängt sowohl von deren Entfernung als auch von ihrer gegenseitigen relativen Orientierung ab, wobei auch die Position von Wasserstoffatomen von Bedeutung ist (Cole *et al.*, 1998; Murray-Rust & Glusker, 1984). Die statistischen Paar- und Einteilchenpräferenzen der in dieser Arbeit entwickelten Bewertungsfunktion werden allerdings nur für Nichtwasserstoffatome abgeleitet (s. a. Kap. 4.4). Zudem weist ein *einzelnes* distanzabhängiges Paarpotential ausschließlich einen kugelsymmetrischen Potentialverlauf auf, der keinerlei Winkelabhängigkeit der Wechselwirkungsstärke beschreibt. An diesem Punkt stellt sich also die Frage, inwieweit Direktionalität von Wechselwirkungen zwischen (Paaren von) Atomen implizit durch die gleichzeitige Betrachtung *mehrerer* distanzabhängiger Paarpotentiale in Form einer zusammengesetzten Repräsentation beschrieben wird.

Um diese Eigenschaft für die Gesamtheit aller Paarpotentiale zu untersuchen, wird in der Bindetasche von kristallographisch bestimmten Protein-Ligand-Komplexen jeweils ein kubisches Gitter mit einer Weite von 0.5 Å und einem Abstand um alle Ligandatome von mindestens 8 Å erzeugt. An jedem Gitterpunkt $g$, dessen Abstand $d_{g,p}$ von einem Proteinatom $p \in P_k$ größer als der van der Waals-Radius dieses Atoms ist und innerhalb des Intervalls $[r_i, r_i + dr)$ liegt, wird anschließend für alle zur Verfügung stehenden Ligandatomtypen $T$ unter Verwendung eines Sondenatoms eine Bewertung mit den Paarpotentialen berechnet:

$$\forall t \in T: \quad \Delta P_{g,t} = \sum_{p \in P_k} \Delta W_{t,\mathrm{T}(p)}(r_i) \qquad \textbf{Gl. 45}$$

Zur Visualisierung werden die erhaltenen Gitterwerte anschließend für individuelle Ligandatomtypen konturiert, wobei die Isoplethen – sofern nicht anders angegeben – alle Potentialwerte bis zu 10 % über dem globalen Minimum des betrachteten Typs einschließen.

Für eine statistische Untersuchung werden in der Kristallstruktur beobachtete Ligandatomtypen mit den von den Paarpotentialen als am günstigsten in dieser Region der Bindetasche vorhergesagten verglichen. Hierzu werden Potentialwerte jeweils für ein Sondenatom des Typs C.3, N.3, O.3, O.2 und O.co2 an zu einem betrachteten Ligandatom nächsten Gitterpunkt berechnet und der Sondenatomtyp mit dem günstigsten Wert bestimmt. Der Einfluß des

Lösemittels auf die Anwesenheit eines Ligandatoms bestimmten Typs wird untersucht, indem vorhergesagte und tatsächlich gefundene Atomtypen sowohl nur für vollständig vergrabene als auch für alle Atome (unabhängig von ihrem Grad der Vergrabung) verglichen werden. Die Lösemittel-zugängliche Oberfläche wird hierzu mit dem in Kap. 4.3.2 erläuterten Algorithmus bestimmt.

## *4.8 Proteinspezifische Adaptierung der statistischen Paarpräferenzen durch Einbeziehung von Zusatzinformation*

Die in Kap. 4.2 eingeführten distanzabhängigen Paarpotentiale werden gemäß Kap. 4.4 aus einer Datenbank von Protein-Ligand-Komplexen abgeleitet. Bei einer *ausreichenden* Anzahl *verschiedener* Komplexe kann daher angenommen werden, daß die *beobachteten* Paarverteilungen den *tatsächlichen* entsprechen und die daraus berechneten statistischen Präferenzen allgemeingültige Informationen über Protein-Ligand-Wechselwirkungen beinhalten (s. a. Kap. 5.1.4). Allerdings ist zu berücksichtigen, daß diese Potentiale durch *Mittelung* von Beobachtungen über eine Vielzahl von Kristallstrukturen erhalten werden. Das durch sie repräsentierte „mittlere“ Bild von Protein-Ligand-Wechselwirkungen gilt daher nur begrenzt für von diesem Bild abweichende Extreme. Zudem unterliegt ihre Anwendung der Annahme, daß die Verteilungen der intermolekularen Distanzen (und damit die erhaltenen Paarpräferenzen selbst) für verschiedene molekulare Umgebungen des jeweiligen Atompaares ähnlich sind. An diesem Punkt stellt sich daher die Frage, ob und wie die erhaltenen „mittleren“ (d.h. allgemeinen) Informationen der Paarpräferenzen unter Einbeziehung weiterer bekannter *struktureller* und *energetischer* Informationen für Komplexe *eines* Proteins so adaptiert werden können, daß die *proteinspezifische* Vorhersage von Struktur und Bindungsaffinität dieser Komplexe verläßlicher wird. Hierbei ist allerdings zu erwarten, daß die so adaptierte Methode eine geringere *generelle* Vorhersagekraft besitzt. Insbesondere gilt zu untersuchen, wie*viel* Zusatzinformation für eine Adaptierung benötigt wird. Weiterhin sollte die zu entwickelnde Methode eine Möglichkeit bieten, den Grad zwischen *Spezifität* und *Generalität* variieren zu können, so daß bei Vorliegen von wenig Zusatzinformation die ursprünglich abgeleiteten, allgemeinen Paarpräferenzen einen größeren Einfluß auf das Ergebnis haben und umgekehrt. Der letzte Punkt entspräche auch den Anforderungen im Rahmen des strukturbasierten Designs von Liganden: das mit jedem (Optimierungs-)Zyklus zunehmende Wissen könnte in dem darauffolgenden verstärkt berücksichtigt werden und würde so zu einer Zunahme der Verläßlichkeit der vorhergesagten Komplexeigenschaften des spezifischen Proteins führen.

### 4.8.1 Berechnung von Wechselwirkungsfeldern

Nach Gl. 15 ergibt sich der auf die wissensbasierten Paarpotentiale zurückgehende Teil der Wechselwirkungen für eine gegebene Konfiguration von Protein und Ligand durch Summation über alle Paarwechselwirkungen. Nach Gl. 42 ergibt sich daraus ein Zusammenhang zu Bindungsaffinitäten (hier ausgedrückt als $pK_i$-Werte) durch Skalierung mit einer (zu bestimmenden) Konstanten $c_{S,Paar}$. Ein denkbarer Ansatz zur Einbeziehung zusätzlicher struktureller und enthalpischer Information bestünde nun darin, anstelle einer allgemeinen Konstanten $c_{S,Paar}$ Gewichtungsfaktoren $\eta_{\mathrm{T}(l),\mathrm{T}(p)}$ für die Atom-Atom-basierten statistischen Nettopräferenzen einzuführen (hinsichtlich der Bedeutung der verwendeten Symbole s. a. Kap. 4.2.1) und diese problemspezifisch - aber unter Berücksichtigung bereits bekannter Information - anzupassen.

$$pK_i = \sum_{l \in L_k} \sum_{p \in P_k} (\eta_{\mathrm{T}(l),\mathrm{T}(p)} \cdot \Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)) \qquad \textbf{Gl. 46}$$

Unter Verwendung der Bayes'schen Regression (O'Hagan, 1994) ist solch ein Vorgehen auch von Murray *et al.* (Murray *et al.*, 1998) für eine „regressionsbasierte" Bewertungsfunktion mit bis zu 6 Koeffizienten angewendet worden. Hierbei ist jedoch zu beachten, daß in dem vorliegenden Fall der Paarpräferenzen die Anzahl der zu adaptierenden Parameter gleich dem Produkt aus der Anzahl der auftretenden Atomtypen für Ligand- und Proteinatome ist, bei 17 Atomtypen auf beiden Seiten also im Extremfall 289 beträgt. Bei Verwendung klassischer linearer Regressionsmethoden (wie im Fall der Bayes'schen Regression) *muß* die Anzahl der zur Anpassung herangezogenen Datenpunkte aber mindestens genauso groß sein, um eine Unterbestimmtheit des Gleichungssystems zu verhindern, und *sollte* mindestens das fünffache der anzupassenden Parameter betragen, um die Wahrscheinlichkeit von Zufallskorrelationen zu reduzieren. Da nicht davon ausgegangen werden kann, daß diese Menge an Zusatzinformation in allen Fällen zur Verfügung steht, wird dieser Ansatz hier *nicht* weiter verfolgt.

Der in Gl. 46 aufgeführte Ansatz weist zusätzlich zu dem formal bedingten Nachteil noch einen methodischen auf. Die danach gewonnenen Gewichtungsfaktoren drücken ebenfalls nur einen *gemittelten* Einfluß der *strukturellen* Zusatzinformation aus – für einen gegebenen Abstand $d_{l,p} \in [r_i, r_i + dr)$ zwischen zwei betrachteten Ligand- bzw. Proteinatomen $l$ und $p$ ergibt sich nach wie vor ein einheitlicher, ortsunabhängiger Beitrag $\Delta W_{\mathrm{T}(l),\mathrm{T}(p)}(r_i)$. Für ein betrachtetes Protein wäre jedoch ein vom jeweiligen Ort in der Bindetasche abhängig vorhergesagter Beitrag einer Atom-Atom-Wechselwirkung wünschenswert. Die im folgenden erläuterte Methode paßt daher die in den Paarpotentialen enthaltene Information *ortsabhängig* an die in

Form der Komplexe *eines* Proteins bekannter Geometrie und Bindungsaffinität vorhandenen Zusatzinformationen an.

Ausgangspunkt sind die überlagerten Protein-Ligand-Komplexgeometrien *eines* Proteins (hinsichtlich der hier gewählten Überlagerungsmethode sowie der Verbindungen des Trainings- und Testdatensatzes siehe Kap. 4.9.5). Um die so erhaltene Anordnung von Liganden wird ein kubisches Gitter *G* in der Bindetasche des Proteins erzeugt, das die Ligandmoleküle in allen drei Raumrichtungen mit einem Überstand von mindestens 4 Å umschließt. Die an dieser Stelle untersuchten Gitterweiten lagen bei 1.0, 1.5 und 2.0 Å. Für ausgewählte Ligandatomtypen $T' \subseteq T$ (hinsichtlich der getroffenen Auswahl s.a. Kap. 5.7.2) wird nun an jedem Punkt des Gitters unter Verwendung eines *Sonden*atoms eine Bewertung mit den zuvor abgeleiteten Paarpotentialen berechnet (Gl. 15). Dieses Vorgehen gleicht der in Gl. 45 beschriebenen Methode, wobei ein künstlicher Abstoßungsterm zu den Paarpotentialen hinzugefügt wird (s.u.). Er dient dazu, auch für die Gitterpunkte eine Bewertung zu erhalten, die näher als die Summe der van der Waals-Radien von Ligand- und Proteinatomen am Protein liegen, wobei für diese betrachteten Distanzbereiche aufgrund mangelnder Beobachtungen in der zur Ableitung verwendeten Komplexdatenbank keine verläßlichen Paarpräferenzwerte herangezogen werden können. Die so erhaltenen Bewertungen an den Gitterpunkten können in Analogie zur Elektrostatik als *Potentialfeld* für ein Ligandatom gegebenen Typs bezeichnet werden.

Der Abstoßungsterm wird folgendermaßen ermittelt (Abb. 7): zunächst wird - von größeren Distanzen kommend – die Lage des ersten Maximums $d_{max}$ im Verlauf der Paarpotentiale für Atom-Atom-Abstände kleiner einer Kontaktdistanz bestimmt. Die Kontaktdistanz ergibt sich für Wechselwirkungen zwischen polaren Atomen aus der mit 0.75 multiplizierten Summe der van der Waals-Radien gemäß Tab. 2 (S. 65), für alle anderen Wechselwirkungen nur aus der Summe der van der Waals-Radien. Anschließend werden die Koeffizienten einer beim Abstand der Atome von 0 zentrierten Gauss-Funktion iterativ so bestimmt, daß die aus dem ursprünglichen Potential und der Gauss-Funktion zusammengesetzte Funktion am Ort $d_{max} + dr$ stetig ist. Für die Höhe der Gauss-Funktion beim Atom-Atom-Abstand von 0 werden Werte von 10, 20 und 40 getestet; sie sind in Bezug auf die Größenordnung der abgeleiteten statistischen Präferenzen zu sehen (siehe dazu auch Abb. 11, S. 110). Für Abstände zwischen 0 und $d_{max} + dr$ werden nun die Funktionswerte der so ermittelten Gauss-Funktion verwendet, für Abstände von $d_{max} + dr$ bis $r_{max}$ gelten die ursprünglichen Werte der Paarpräferenzen.

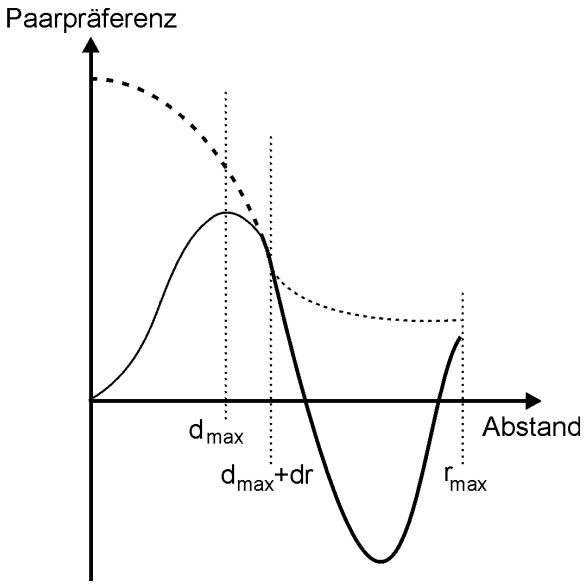



**Abb. 7:** Schematische Darstellung des künstlich zu den Paarpräferenzen (durchgezogene Linie) hinzugefügten Abstoßungsterms (gestrichelte Linie) in Form einer Gauss-Funktion. Für Abstände von 0 bis $d_{max}$ + $dr$ werden die Werte dieser Gauss-Funktion verwendet, für Abstände von $d_{max}$ + $dr$ bis $r_{max}$ die Werte der Paarpräferenz (breite Linien). Zusätzlich mit angegeben sind die nicht weiter berücksichtigten Bereiche der beiden Funktionen (schmale Linien).

Die Wechselwirkungen von jeweils in der Bindetasche plazierten Liganden mit dem Protein werden in nachfolgend beschriebener Form auf die Gitterpunkte abgebildet (Abb. 8). Für jeden Gitterpunkt $g$ wird der distanzabhängige Beitrag $B_{g,l}$ eines Ligandatoms $l$ eines Typs $T(l) \in T$‘ unter Anwendung einer Gaussfunktion ermittelt

$$B_{g,l} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{d_{g,l}^{\,2}}{2\sigma^2}\right\} \qquad \textbf{Gl. 47}.$$

Hierbei werden $\sigma$-Werte von 0.55, 0.7, ... , 1.3 getestet; die Wendepunkte der Gauss-Funktion liegen bei der hier verwendeten Form jeweils bei einem Abstand $d_{g,l} = \sigma$.

Der Gesamtwechselwirkungsbeitrag für einen Ligandatomtyp $t \in T$‘ an einem Gitterpunkt $g$ ergibt sich sodann durch Summation der Produkte des normierten distanzabhängigen Beitrags $B_{g,l}$ aller Ligandatome $l \in L$ mit $T(l) = t$ mit dem Wert des Potentialfeldes an diesem Punkt $\Delta P_{g,t}$ (s. o. und Gl. 45)

$$\Delta W_{g,t} = \sum_{l \in L:T(l)=t} \frac{B_{g,l}}{B_l} \Delta P_{g,t} \qquad \textbf{Gl. 48}$$

Die Normierungskonstante berechnet sich als Summe über alle Gitterpunkte $g \in G$ aller distanzabhängigen Beiträge $B_{g,l}$ eines Ligandatoms:

$$B_l = \sum_{g \in G} B_{g,l} \qquad \textbf{Gl. 49}$$

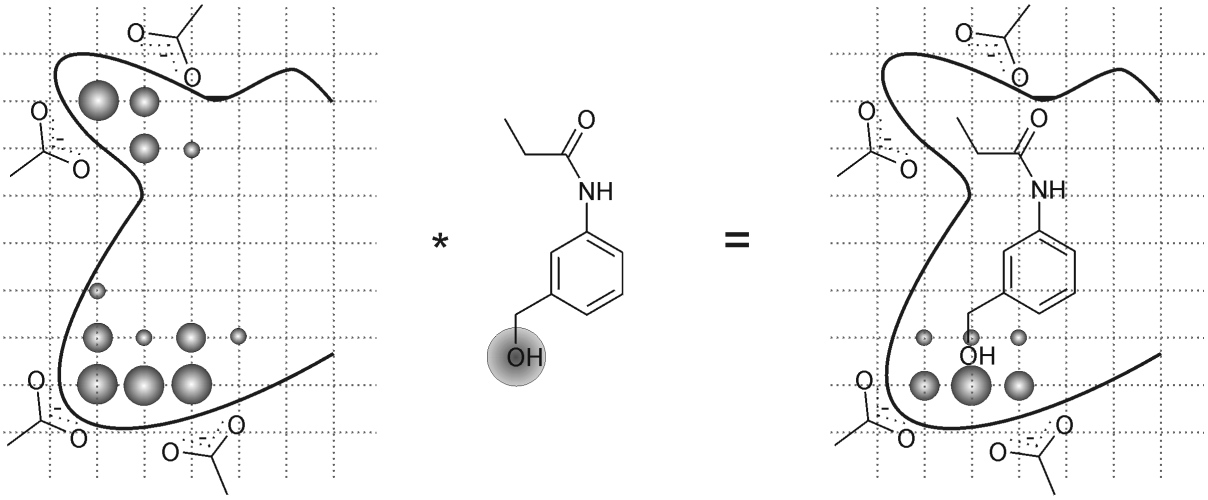



**Abb. 8:** Schematische Darstellung der Berechnung der Wechselwirkungsfelder nach Gl. 48 für den Atomtyp O.3. Ausgehend von nach Gl. 45 in der Bindetasche für diesen Typ berechneten Potentialfeldern (linker Teil) erfolgt die Abbildung der durch das Hydroxylsauerstoffatom des Liganden (Mitte) bedingten Wechselwirkungen auf die benachbarten Gitterpunkte (rechter Teil) unter Verwendung einer räumlichen Gauss-Funktion (Gl. 47).

Die auf diese Weise für jeden Gitterpunkt $g$ erhaltenen Werte $\Delta W_{g,t}$ ergeben insgesamt das *Bewertungsfeld* (in Analogie zur Elektrostatik auch *Energiefeld*) bzgl. eines Atomtyps $t$ für einen gegebenen Liganden in der Bindetasche. Letzteres wird insbesondere bei Summation über alle Gitterpunkte $g \in G$ deutlich:

$$\Delta W_t = \sum_{g \in G} \Delta W_{g,t} = \sum_{g \in G} \sum_{\substack{l \in L: \\ \mathrm{T}(l)=t}} \frac{B_{g,l}}{B_l} \Delta P_{g,t} = \sum_{\substack{l \in L: \\ \mathrm{T}(l)=t}} \frac{\sum_{g \in G} B_{g,l} \; \Delta P_{g,t}}{B_l} \qquad \textbf{Gl. 50}$$

Unter Beachtung von Gl. 45 ergibt sich hiernach der auf die statistischen Paarpräferenzen zurückgehende Beitrag aller Ligandatome des Typs $t$ zu den Protein-Ligand-Wechselwirkungen als vom Ort des einzelnen Atoms (distanz-)abhängig gewichtete Mittelung der Beiträge der umliegenden Gitterpunkte. Im Grenzfall eines Gitters der Weite 0 und der Verwendung von $\sigma \to 0$ (die Gauss-Funktion geht dann in eine Delta-Funktion über) zur Ermittlung von $B_{g,l}$ (Gl. 47), sowie nach Summation über alle Ligandatomtypen $t \in T'$ (für $T' = T$), stimmt das damit erhaltene Ergebnis mit dem direkt unter Verwendung der Paarpräferenzen gemäß Gl. 15 erhaltenen überein.

### 4.8.2 Korrelation der Wechselwirkungsfelder mit experimentell bestimmten Bindungsaffinitäten und Vorhersage unbekannter Bindungsaffinitäten

Die Summation aller nach Gl. 48 erhaltenen Wechselwirkungsbeiträge über alle Gitterpunkte $g \in G$ sowie über alle Atomtypen $t \in T'$ liefert gemäß Gl. 50 dann lediglich einen *Anteil* $\Delta W'_{Paar}$ des nach Gl. 15 auf die wissenbasierten Paarpräferenzen zurückgehenden Beitrags der Protein-Ligand-Wechselwirkungen, wenn $T'$ eine *echte* Teilmenge von $T$ ist.

$$\Delta W'_{Paar} = \sum_{t \in T'} \sum_{g \in G} \Delta W_{g,t} \qquad \textbf{Gl. 51}$$

Mit Gl. 42 und einer (unter Verwendung *aller* Atomtypen in $T$ bestimmten) Konstanten $c_{S,Paar}$ ließe sich damit der auf die Atome mit $t \in T'$ zurückgehende *Anteil* an der Bindungsaffinität für eine gegebene Anordnung von Protein und Ligand vorhersagen.

$$pK'_i = c_{S,Paar}\ \Delta W'_{Paar} \qquad \textbf{Gl. 52}$$

Ligandatome mit $t \in T \setminus T'$ liefern den verbleibenden Beitrag zur Bindungsaffinität in Analogie zu Gl. 42.

$$pK''_i = c_{S,Paar}\ \Delta W''_{Paar} \qquad \textbf{Gl. 53}$$

$\Delta W''_{Paar}$ wird hierbei gemäß Gl. 15 für diese Ligandatome berechnet. Letztendlich ergibt sich die vorhergesagte Bindungsaffinität als Summe beider Einzelbeiträge aus Gl. 52 und 53.

$$pK_i = pK'_i + pK''_i \qquad \textbf{Gl. 54}$$

Darüber hinaus können jedoch die *einzelnen* Beiträge zu $\Delta W'$ (nach Gl. 51) und damit zu $pK_i'$ (nach Gl. 52) unter Verwendung eines Trainingsdatensatzes aus Protein-Ligand-Komplexen mit bekannter Geometrie und Bindungsaffinität *ortsabhängig* für den Bereich der Bindetasche *eines* betrachteten Proteins angepaßt werden, indem adäquate Koeffizienten $c_{g,t}$ ermittelt werden.

$$pK'_i = \sum_{t \in T'} \sum_{g \in G} c_{g,t}\ \Delta W_{g,t} \qquad \textbf{Gl. 55}$$

Hierzu müssen die Koeffizienten $c_{g,t}$ durch Lösung des linearen Gleichungssystems $\vec{y} = \hat{X}\ \vec{c}$ bestimmt werden. $\vec{y}$ steht hierbei für einen (zentrierten) $n$-dimensionalen Spaltenvektor der aus den Bindungsaffinitäten $pK_i$ der $n$ Komplexe des Trainingsdatensatzes durch Abziehen der Beiträge $pK_i''$ (Gl. 53) erhaltenen Restbeiträge $pK_i'$ (gemäß Gl. 54). Die Verwendung dieser Restbeiträge anstelle der gesamten Bindungsaffinitäten eröffnet die Möglichkeit, nicht alle Atomtypen der Verbindungen des Trainingsdatensatzes in die Bestimmung der Koeffizienten $c_{g,t}$ mit einzubeziehen. Dies gilt insbesondere für nur selten in diesen Molekülen repräsentierte Typen (s.a. Kap. 4.8.3). $\hat{X}$ ist eine $n$-mal-$m$-Matrix, wobei jede der $n$ Rei-

hen aus $m = \|G\| \cdot \|T'\|$ (zentrierten) Spalten („Deskriptoren“) besteht, mit ||*G*|| als Anzahl der Gitterpunkte und ||*T*‘|| als Anzahl der betrachteten Atomtypen. Die Werte der Spalten entsprechen den nach Gl. 48 berechneten Wechselwirkungsbeiträgen der Atome des Typs *t* am Ort des Gitterpunktes *g*. $\bar{c}$ steht für den *m*-dimensionalen Spaltenvektor der Koeffizienten $c_{g,t}$. Da für *m* Werte in der Größenordnung von $10^3$ bis $10^5$ zu erwarten sind, allerdings nur maximal etwa 100 Trainingsdaten zur Verfügung stehen, ist das aufgestellte Gleichungssystem stark unterbestimmt. Zudem kann davon ausgegangen werden, daß die einzelnen Deskriptoren untereinander korreliert sind. Beides würde bei der Anwendung von multipler linearer Regression zur Bestimmung von $\bar{c}$ zu einem Modell zwischen den unabhängigen Variablen $\hat{X}$ und den abhängigen Größen $\bar{y}$ führen, das übertrainiert ist. Solch ein Modell wäre dann zwar ideal an die Trainingsdaten angepaßt, könnte jedoch nicht für verläßliche Vorhersagen der Bindungsaffinität neuer, nicht im Trainingsdatensatz enthaltener Verbindungen verwendet werden.

Die von Wold *et al.* entwickelte PLS-Analyse (***P**artial **L**east **S**quares*) (Geladi & Kowalski, 1986; Wold *et al.*, 1984) verhindert jedoch gerade dieses Übertrainieren, indem sie die Deskriptoren auf eine geringe Anzahl von daraus gebildeten, zueinander orthogonalen Linearkombinationen („Komponenten“) reduziert, die die Gesamtheit aller Deskriptoren möglichst gut erklären und die abhängigen Größen möglichst gut vorhersagen können (Wold *et al.*, 1993). Hierbei werden in einem iterativen Verfahren diese Komponenten nacheinander generiert (NIPALS (***N**onlinear **I**terative **P**artial **L**east **S**quares*)-Algorithmus (Geladi & Kowalski, 1986)). Der hier zur Implementierung verwendete Algorithmus folgt dem von Bush und Nachbar (Bush & Nachbar, 1993) beschriebenen Ansatz, der auf die Arbeiten von Wold *et al.* (Wold *et al.*, 1984) zurückgeht. Die damit erhaltene erste Komponente (auch als „Trendvektor“ bezeichnet) erklärt einen Teil des beobachteten Trends der abhängigen Größen im Raum der unabhängigen Größen, die zweite erklärt einen Teil des verbliebenen Trends in einer zur ersten orthogonalen Richtung usw. Wird der Prozeß fortgesetzt, bis die Anzahl der erhaltenen Komponenten *h* gleich min(*n*,*m*) ist, so ist eine *Full Least Squares*-Analyse durchgeführt worden. Es ist allerdings zu beachten, daß mit jedem Zyklus das Signal-zu-Rauschen-Verhältnis in den erhaltenen Regressionsgrößen abnimmt, so daß zwar die Anpassung an die Trainingsdaten verbessert wird, nicht jedoch die Vorhersagekraft dieser Größen. Aus diesem Grund wird die Anzahl der verwendeten Komponenten *h* i.a. (deutlich) kleiner als *n* (für $n \leq m$) gewählt.

Die Anzahl der optimalen Komponenten wird hier unter Verwendung einer „*Leave-One-Out*“-Kreuzvalidierung bestimmt, womit die statistische Signifikanz wie auch die Vorhersage-

fähigkeit des Modells überprüft wird. Hierzu werden für *n* Trainingsdaten *n* PLS-Analysen durchgeführt, wobei jeweils nur *n* – 1 Trainingsdaten zur Aufstellung des Modells verwendet werden. Mit diesem Modell wird sodann die abhängige Größe des nicht in der PLS-Analyse verwendeten Datums vorhergesagt ($pK'_{i,pred}$) und mit der tatsächlichen ($pK'_{i,act}$) verglichen. Anstelle der PLS-Analyse wird hierzu die ebenfalls in Bush & Nachbar (Bush & Nachbar, 1993) beschriebene SAMPLS-Methode (***SAM****ple-distance* ***PLS***) implementiert. Hierbei muß jedoch darauf geachtet werden, daß die in Gl. 12‘ bzw. Gl. 16 in (Bush & Nachbar, 1993) erhaltenen Größen zusätzlich zentriert werden müssen. Die SAMPLS-Methode liefert die gleichen Ergebnisse wie eine PLS-Analyse, allerdings in wesentlich kürzerer Rechenzeit. Anstelle die *n*-mal-*m*-Matrix der unabhängigen Variablen nach jedem Zyklus anpassen zu müssen, wie in einer PLS-Analyse verlangt, arbeitet die SAMPLS-Methode mit der *n*-mal-*n*-Kovarianzmatrix $\hat{X}\,\hat{X}^T$. Den Namen dieser Methode bedingt schließlich, daß anstelle der Kovarianzmatrix auch die Matrix der Distanzen der Trainingsdaten im Deskriptorraum verwendet werden kann.

Die Qualität der Vorhersagen läßt sich in zwei Größen zusammenfassen. Die zu einem Regressionskoeffizienten analoge Größe wird in diesem Zusammenhang als $q^2$-Wert (bzw. kreuzvalidierter $r^2$-Wert) bezeichnet.

$$q^2 = 1 - \frac{PRESS}{SSD} = 1 - \frac{\sum_{j=1}^{n} (pK'_{i,pred}(j) - pK'_{i,act}(j))^2}{\sum_{j=1}^{n} (pK'_{i,act}(j) - pK'_{i,mean})^2} \qquad \textbf{Gl. 56}$$

*PRESS* (***P****redictive* ***RE****sidual* ***S****um of* ***S****quares*) ist die Summe der quadrierten Differenzen zwischen vorhergesagtem und tatsächlichem Bindungsaffinitätbeitrag, *SSD* ist die Summe der quadrierten Differenzen aus tatsächlichem $pK_i$‘-Wert der Verbindung und dem Mittelwert über alle Bindungsaffinitätsbeiträge des Trainingsdatensatzes. Ein $q^2$-Wert von 1 bedeutet, daß ein perfektes Modell vorliegt (*PRESS* = 0), ein $q^2$-Wert von 0 zeigt „kein“ Modell an, denn anstelle der $pK'_{i,pred}$-Werte könnte man hierfür auch den Mittelwert $pK'_{i,mean}$ des Datensatzes zur Vorhersage verwenden. Ein Ergebnis wird dagegen als statistisch signifikant angesehen, wenn $q^2 > 0.3$ (Cramer III *et al.*, 1993), ein „gutes“ Modell besitzt $q^2$-Werte größer als 0.5.

Die Abweichung $s_{PRESS}$ gibt ein Maß für die erwartete Unsicherheit der vorhergesagten Bindungsaffinität bei der Kreuzvalidierung an. Sie hängt dabei von der Skalierung der abhängigen Variablen ab.

$$s_{PRESS} = \sqrt{\frac{PRESS}{n-h-1}} = \sqrt{\frac{\sum_{j=1}^{n} (pK'_{i,pred}(j) - pK'_{i,act}(j))^2}{n-h-1}} \qquad \textbf{Gl. 57}$$

Der Term *n*-*h*-1 gibt die Anzahl der Freiheitsgrade an; bedingt durch die Einbeziehung der Anzahl der Komponenten *h* ergibt sich für ein Modell mit gleichem *PRESS*-Wert, aber weniger verwendeten Komponenten, ein geringerer $s_{PRESS}$-Wert. Dies folgt dem *parsimony*-Prinzip† (Thibaut *et al*., 1993), nach dem das Ziel einer PLS-Analyse ein Modell mit möglichst wenigen Komponenten bei optimaler Vorhersagekraft ist. In der vorliegenden Arbeit wird die Anzahl optimaler Komponenten so gewählt, daß $s_{PRESS}$ minimal wird.

Unter Verwendung der so bestimmten Anzahl optimaler Komponenten wird anschließend eine PLS-Analyse mit allen *n* Trainingsdaten durchgeführt. Hierbei werden alle Werte einer Deskriptorspalte dann nicht berücksichtigt, wenn die Standardabweichung dieser Spalte kleiner als $5 \cdot 10^{-4}$ ist. Dieser Wert wurde so gewählt, daß etwa 90 % aller Spalten auf diesem Weg eliminiert werden, was zu einer 10-fachen Steigerung der Rechengeschwindigkeit führt (Cramer III *et al.*, 1993). Die statistischen Parameter sind analog zu den in Gl. 56 und 57 angegebenen.

$$r^2 = 1 - \frac{RSS}{SSD} = 1 - \frac{\sum_{j=1}^{n} (pK'_{i,fit}(j) - pK'_{i,act}(j))^2}{SSD} \qquad \textbf{Gl. 58}$$

$$S = \sqrt{\frac{RSS}{n-h-1}} = \sqrt{\frac{\sum_{j=1}^{n} (pK'_{i,fit}(j) - pK'_{i,act}(j))^2}{n-h-1}} \qquad \textbf{Gl. 59}$$

*RSS* ist die Summe der quadrierten Differenzen zwischen angepaßtem und tatsächlichem Bindungsaffinitätsbeitrag. $r^2$ und *S* geben hierbei an, wie gut die berechneten Bindungsaffinitäten an die experimentell bestimmten *angepaßt* wurden. Für eine Aussage über die Güte eines Modells ist jedoch der $q^2$-Wert entscheidend, da er (im Gegensatz zum $r^2$-Wert) eine Abschätzung über die *Vorhersagefähigkeit* mit diesem Modell erlaubt. Zusätzlich wird noch der Fischersche F-Wert berechnet. Er definiert das unter der Beachtung der Anzahlen der verwendeten Komponenten *h* sowie der abhängigen Variablen *n* gewichtete Verhältnis zwischen erklärten und unerklärten Ergebnissen (Harnett & Murphy, 1975).

† Auch als *Occam's Razor* (William of Occam, engl. Philosoph, ca. 1285-1349) bezeichnet: „Bevorzuge die einfachste Hypothese, die konsistent mit allen Beobachtungen ist."

$$F = \frac{r^2}{1-r^2} \cdot \frac{n-h-1}{h} \qquad \textbf{Gl. 60}$$

Der Anteil der einzelnen Ligandatomtyp-Felder an der Beschreibung der beobachteten Bindungsaffinitäten wird als sog. „erklärender Beitrag“ (*contribution*) angegeben:

$$\forall t \in T' : \quad contribution_t = \sum_{g \in G} \frac{SD_{g,t} \cdot \left| c_{g,t} \right|}{SDY} \qquad \textbf{Gl. 61}$$

*SDY* ist die Standardabweichung der abhängigen Variablen, $SD_{g,t}$ die Standardabweichung der Spalte *g* des Feldes *t*, $|c_{g,t}|$ der Betrag des errechneten Koeffizienten für diese Spalte.

Für die in Gl. 56 bis 59 aufgeführten Größen ist zu beachten, daß jeweils die $pK_i$‘-Größen durch die nach Gl. 54 unter Verwendung des in dem Modell nicht erfassten Beitrags $pK_i$‘‘ erhaltenen, analogen Größen $pK_i$ ersetzt werden können. Hierbei verändern sich die Werte für $s_{PRESS}$ und *S* nicht. Die für $q^2$ und $r^2$ abweichenden Werte beruhen dagegen auf jeweils veränderten *SSD*-Werten.

Die $pK_i$‘-Werte des Trainingsdatensatzes wurden zentriert. Der dazu verwendete Mittelwert ($<pK_i'>$) aller $pK_i$‘-Werte entspricht dem Achsenabschnitt der Regressionshyperfläche und muß bei der Vorhersage jeweils berücksichtigt werden. Die einzelnen Spalten der Deskriptormatrix werden ebenfalls zentriert ($<\Delta W_{g,t}>$). Da die Spaltenwerte auf die abgeleiteten Paarpräferenzen zurückzuführen sind, diese aber wiederum gemäß Kap. 4.2 in einem inneren Zusammenhang stehen, kann angenommen werden, daß dieses auch für die Gewichte der einzelnen Spalten gilt. Aus diesem Grund wurden die einzelnen Spalten grundsätzlich nicht skaliert. Ergebnisse, die nach Autoskalierung (d.h. Division der Spaltenwerte durch ihre jeweilige Standardabweichung $\overline{\Delta W_{g,t}}$ entsprechend einer Einheitsgewichtung aller Spalten) erhalten wurden, sind zum Vergleich in Tab. 31 (S. 182) mit aufgeführt. Hierbei ist zu beachten, daß während der einzelnen Kreuzvalidierungsläufe die Spalten neu *zentriert*, aber *nicht reskaliert* werden, was im Einklang mit Empfehlungen von Wold (Wold *et al.*, 1993), Kubinyi (Kubinyi & Abraham, 1993) und Bush (Bush & Nachbar, 1993) steht. Die Zentrierungen und Skalierungen des Trainingsdatensatzes müssen analog auf die Daten des Testsatzes angewendet werden.

Um zu überprüfen, daß das durch die PLS-Analyse mit dem Trainingsdatensatz erhaltene Modell nicht auf einer Zufallskorrelation beruht, wurden die $pK_i$‘-Werte der einzelnen Verbindungen mehrfach zufällig vertauscht und anschließend die Ableitung des Modells wiederholt.

### 4.8.3 Graduelle Variation des Beitrags der angepaßten Wechselwirkungsfelder auf die Vorhersage unbekannter Bindungsaffinitäten

Bei nur gering vorhandener, proteinspezifischer Zusatzinformation sollte die Vorhersage von Bindungsaffinitäten unter Anwendung von Gl. 55 mit großer Unsicherheit behaftet, die Verwendung der aus der Protein-Ligand-Datenbank abgeleiteten, allgemeingültigen Paarpräferenzen gemäß Gl. 53 dagegen zuverlässiger sein. Mit zunehmender Signifikanz der Zusatzinformationen für *ein* Protein in Form verschiedener Protein-Ligand-Strukturen mit bekannten Bindungsaffinitäten ist allerdings eine Umkehrung der genannten Verhältnisse zu erwarten. Dementsprechend sollte eine graduelle Variation des Beitrags der angepaßten Wechselwirkungsfelder auf die Vorhersage unbekannter Bindungsaffinitäten in Abhängigkeit des zur Verfügung stehenden Umfangs an Zusatzinformationen möglich sein. Dies gelingt durch „Mischen“ der allgemeinen, auf die aus der Datenbank aller Komplexe abgeleiteten Paarpräferenzen zurückgehenden Wechselwirkungsfelder mit den durch die spezifische Anpassung an ein Protein erhaltenen Felder. Für Ligandatome *l*, deren Typ T(l) in *T‘* enthalten ist und die sich innerhalb der Grenzen des zur Anpassung verwendeten Gitters befinden, folgt damit:

$$
\begin{aligned}
pK_i' &= (1-\theta)\, c_{S,Paar} \sum_{g\in G} \sum_{\substack{l\in L:\\ \mathrm{T}(l)\in T'}} \Delta W_{g,\mathrm{T}(l)} + \\
&\quad \theta \left[ \sum_{g\in G} \sum_{\substack{l\in L:\\ \mathrm{T}(l)\in T'}} \left( c_{g,\mathrm{T}(l)} \, \frac{\Delta W_{g,\mathrm{T}(l)} - <\Delta W_{g,\mathrm{T}(l)}>}{\overline{\Delta W_{g,\mathrm{T}(l)}}} \right) + <pK_i'> \right] \\
&= \theta <pK_i'> + \sum_{g\in G} \sum_{\substack{l\in L:\\ \mathrm{T}(l)\in T'}} \Bigg( \underbrace{\theta \; c_{g,\mathrm{T}(l)} \, \frac{\Delta W_{g,\mathrm{T}(l)} - <\Delta W_{g,\mathrm{T}(l)}>}{\overline{\Delta W_{g,\mathrm{T}(l)}}}}_{\mathrm{I}} + \underbrace{(1-\theta)\, c_{S,Paar}\, \Delta W_{g,\mathrm{T}(l)}}_{\mathrm{II}} \Bigg)
\end{aligned}
\qquad \textbf{Gl. 62}
$$

$\theta$ kann zwischen 0 und 1 variiert werden. Der erste Term (I) innerhalb der Klammern in der letzten Zeile von Gl. 62 steht dabei für den Beitrag aufgrund der spezifisch angepaßten, der zweite (II) für den Beitrag durch die ursprünglichen, allgemein abgeleiteten Felder. Im ersten Term sind die Zentrierung und Skalierung der einzelnen Spalten der Deskriptormatrix während der Anpassung mitberücksichtigt.

Ligandatome *l*, deren Typ T(l) nicht in *T‘* enthalten ist bzw. solche, die sich außerhalb der Grenzen des zur Anpassung verwendeten Gitters befinden (unabhängig von ihrem Typ), liefern einen Beitrag zur Bindungsaffinität ($pK_i$‘‘) entsprechend den ursprünglichen, nicht spezifisch angepaßten Präferenzen in Analogie zu Gl. 53.

Letztendlich ergibt sich die vorhergesagte Bindungsaffinität als Summe beider Einzelbeiträge aus Gl. 53 und 62 gemäß Gl. 54.

Ligandatome eines nicht in *T'* vertretenen Typs bzw. solche außerhalb des verwendeten Gitters gehen somit mit einer Gewichtung von 1 ($\theta = 0$) des auf die ursprünglichen, allgemein abgeleiteten Paarpräferenzen zurückgehenden Beitrags in die vorhergesagte Bindungsaffinität ein. Somit können auch Affinitäten für Liganden vorhergesagt werden, die aus Atomen aufgebaut sind, die (noch) nicht im Trainingsdatensatz vertreten waren sowie solche, die in Bereichen der Bindetasche liegen, die bislang von Verbindungen des Trainingsdatensatzes noch nicht eingenommen wurden.

Die Güte dieser Vorhersagen gemäß Gl. 54 läßt sich bestimmen, wenn für die verwendeten Komplexe ebenfalls experimentell bestimmte Bindungsaffinitäten bekannt sind.

$$r^2_{pred} = 1 - \frac{PRESS}{SSD} \qquad \textbf{Gl. 63}$$

*PRESS* ist hierbei wieder die Summe der quadrierten Differenzen zwischen vorhergesagten und experimentellen Bindungsaffinitäten, *SSD* die Summe der quadrierten Differenzen zwischen tatsächlichen $pK_i$-Werten und dem Mittelwert über alle Bindungsaffinitäten des Testdatensatzes. Die hier verwendete Definition von $r^2_{pred}$ weicht insofern von der von Cramer *et al.* (Cramer III *et al.*, 1988) gegebenen ab, da hier für die Ermittlung von *SSD* nicht der Mittelwert der Verbindungen des Trainings-, sondern der des Testdatensatzes herangezogen wird. Somit ist $r^2_{pred}$ aber auch mit den Ergebnissen vergleichbar, die bei direkter Anwendung der hier abgeleiteten Paarpotentiale erzielt wurden.

An dieser Stelle ergibt sich die Frage, *wieviel* Zusatzinformation in Form von Protein-Ligand-Komplexen mit bekannter Struktur und Bindungsaffinität benötigt wird, um eine gegenüber der Verwendung der ursprünglichen Paarpräferenzen gemäß Gl. 15 gesteigerte Verläßlichkeit bei der Vorhersage von Affinitäten *neuer* Verbindungen zu erlangen. Zu ihrer Beantwortung wurde oben beschriebener Prozeß zur Ableitung eines adaptierten Modells unter Verwendung einer (SAM)PLS-Analyse mit bzw. ohne Kreuzvalidierung jeweils 100 mal für aus dem Trainingsdatensatz gebildete Teilmengen mit einer festgelegten Anzahl von Verbindungen durchgeführt. Die Teilmengen umfaßten hierbei 5, 15, 30, 45 und 53 der 61 möglichen Trainingsfälle (bzgl. des Datensatzes s. Kap. 4.9.5), die für jeden der insgesamt 500 Läufe jeweils neu zufällig zusammengestellt wurden. Mit den für jede der Teilmengen erhaltenen 100 Modellen wurden nun Bindungsaffinitäten für einen Testdatensatz gemäß Gl. 54 berech-

net, indem zusätzlich der Parameter $\theta$ in Gl. 62 zwischen 0.1 und 1 mit einer Schrittweite von 0.1 variiert wurde ($\theta = 0$ entspricht einer Vorhersage unter Verwendung der ursprünglichen, am allgemeinen Datensatz abgeleiteten Paarpräferenzen). Die dabei erhaltenen statistischen Parameter $r^2_{pred}$ bzw. die Standardabweichung zwischen berechneten und experimentellen $pK_i$-Werten wurden sodann über alle 100 Läufe gemittelt.

## *4.9 Aufbereitung der Testdatensätze*

In diesem Kapitel werden die zur Validierung der entwickelten Bewertungsfunktion hinsichtlich ihrer Eignung zur Vorhersage von Struktur und Affinität verwendeten Datensätze zusammengefaßt und ihre Quellen sowie ihre Aufbereitung im einzelnen beschrieben. Sofern die Protein-Ligand-Komplexe in der PDB (Bernstein *et al.*, 1977) enthalten sind, werden sie hier nur mit dem dazugehörigen PDB-Code aufgeführt.

### 4.9.1 Testdatensätze zur Bestimmung nativ-ähnlicher Protein-Ligand-Konfigurationen

Die Eignung der Bewertungsfunktion zur Vorhersage nativ-ähnlicher Konfigurationen von Protein und Ligand wird an drei Datensätzen (DS) FlexX_DS1, FlexX_DS2 und DOCK_DS durchgeführt, die jeweils Untermengen des Datensatzes von 200 Rezeptor-Ligand-Komplexen sind, der zur Validierung des Docking-Programms FlexX verwendet wurde (Kramer *et al.*, 1999).

FlexX_DS1 enthält 91 Komplexe[‡], die so ausgewählt wurden, daß die enthaltenen Liganden nicht kovalent gebunden sind, bei visueller Inspektion keine augenscheinlichen Fehler in der Geometrie der Protein-gebundenen Moleküle sichtbar sind und sie einen großen Bereich molekularer Diversität abdecken. Die Anzahl darin auftretender drehbarer Bindungen beträgt 0 bis 27, und es sind 1 bis 17 Wasserstoffbrückendonatoren bzw. -akzeptoren sowie 2 bis 39 Kohlenstoffatome in den Molekülen enthalten. Das zweite Kriterium beinhaltet, daß nur in etwa der Hälfte der Fälle von FlexX eine Ligandkonfiguration mit *rmsd* < 2.0 Å bzgl. der Kristallstruktur auf dem ersten Bewertungsrang gefunden wird, wohingegen FlexX für die andere

[‡] 1abe 1abf 1atl 1azm 1bbp 1cbx 1cde 1cil 1com 1cps 1ctr 1did 1die 1dr1 1dwc 1dwd 1ela 1epb 1frp 1ghb 1hfc 1hgj 1hsl 1hyt 1icn 1imb 1ivc 1ivd 1ive 1ivf 1lah 1lcp 1lic 1lna 1lst 1mld 1mrg 1mrk 1nis 1nsc 1pbd 1phf 1poc 1pph 1ppl 1pso 1rbp 1rds 1rnt 1rob 1slt 1snc 1srj 1tlp 1tng 1tnh 1tni 1tpp 1ukz 1wap 1xid 1xie 2ada 2ak3 2cgr 2cht 2cmd 2cpp 2gbp 2mth 2pk4 2sim 2tmn 2xis 2ypi 3aah 3cpa 3hvt 4fbp 4hmg 4phv 4tim 4tln 4ts1 5abp 5p2p 6abp 6rnt 6tmn 7tim 8atc

Hälfte keine dieser geometrisch „guten“ Lösungen als bestbewertete identifizieren kann. FlexX_DS1 wurde für die Anpassung der Parameter der Bewertungsfunktion verwendet.

Der zweite Datensatz FlexX_DS2 enthält 68 Komplexe[§], die ebenfalls die für FlexX_DS1 angewendeten Kriterien erfüllen (0 – 35 drehbare Bindungen; 0 – 37 Wasserstoffbrückendonatoren und -akzeptoren; 0 – 43 Kohlenstoffatome). Dieser zweite Datensatz wurde nicht zur Parameteranpassung der Funktion verwendet und dient daher der Kreuzvalidierung.

Für jeden Protein-Ligand-Komplex wurden mit dem Programm FlexX bis zu 500 verschiedene Rezeptor-Ligand-Konfigurationen erzeugt. Dazu wurden Eingabedateien verwendet, wie sie von den Autoren der FlexX-Validierungsstudie (Kramer *et al.*, 1999) erstellt wurden. Unter Verwendung von SYBYL wurde der Ligand zunächst aus dem Protein-Ligand-Komplex extrahiert und anschließend Atom- und Bindungstypen nach der SYBYL-Notation (SYBYL) sowie Formalladungen zugewiesen. Die so erhaltene Molekülgeometrie diente als Referenzstruktur für die Berechnung von *rmsd*-Werten. Nach Addition von Wasserstoffatomen unter Verwendung von Standardgeometrien mit SYBYL ergibt eine abschließende Energieminimierung unter Verwendung des TRIPOS-Kraftfeldes (Clark *et al.*, 1989) die Eingabegeometrie des Liganden für FlexX mit standardisierten Bindungslängen und -winkeln. Alle Carbonsäure- und Phosphorsäuregruppen der Liganden wurden als deprotoniert, alle Amino-, Amidino- und Guanidinogruppen als protoniert angesehen. Für die Atome des Rezeptors wurden die Positionen verwendet, wie sie in der Kristallstruktur vorliegen. Wassermoleküle wurden generell entfernt bis auf Ausnahmen bei 1aaq, 4phv, 1lna und 1xie. Die ersten beiden Einträge sind HIV-Protease-Komplexe, bei denen das Wassermolekül HOH1 (sog. Strukturwasser in der „Flap-Region“) eine wichtige Rolle für die Ligandbindung spielt (Wlodawer, 1994). Bei den letzten beiden Komplexen (Val-Lys gebunden an Thermolysin sowie 1,5-Dianhydrosorbitol gebunden an D-Xylose-Isomerase) sind Wassermoleküle an Metallionen gebunden. Prosthetische Gruppen wurden bei 1coy (Cholesterol-Oxidase im Komplex mit Dehydroisoandrosteron) und 1dr1 (Dihydrofolat-Reduktase im Komplex mit Biopterin) als zum Protein zugehörig verwendet. Wasserstoffatome werden unter Verwendung von Standardgeometrien mit SYBYL an das Protein addiert und die Torsionswinkel der Hydroxylgruppen von Serin, Threonin und Tyrosin sowie die Position des Wasserstoffs in der Histidinseitenkette durch Betrachtung umliegender Wechselwirkungspartner festgelegt. Die Amino-

[§] 121p 1aaq 1acm 1aco 1aec 1aha 1ake 1apt 1avd 1bma 1byb 1cbs 1cdg 1coy 1dbb 1eap 1eed 1elb 1elc 1eld 1ele 1etr 1fen 1fkg 1glp 1glq 1hdc 1hef 1hvr 1ida 1igj 1ivb 1ldm 1lmo 1lpm 1mbi 1mdr 1mmq 1nco 1phd 1phg 1ppc 1ppi 1ppk 1ppm 1rne 1tnk 1tnl 1tph 1trk 2ctc 2er6 3cla 3gch 3ptb 4dfr 4fxn 4hvp 4phv 4tmn 5cts 5tim 5tmn 6cpa 6tim 7cpa 8gch 9hvp

gruppe von Lysin und die Guanidinogruppe von Arginin werden als protoniert, die Carbonsäuregruppen von Asparagin- und Glutaminsäure als deprotoniert angenommen. Als zu der Bindetasche dazugehörige Atome des Rezeptors werden jene definiert, die innerhalb eines Abstandes von 6.5 Å von den Ligandatomen der Referenzstruktur liegen.

Die Liganden werden mit FlexX jeweils automatisch flexibel in die Bindetaschen der zugehörigen Rezeptoren unter Verwendung eines inkremetellen Aufbaualgorithmus (Rarey *et al.*, 1997; Rarey *et al.*, 1996a; Rarey *et al.*, 1996b) gedockt (für eine Beschreibung s. a. Kap. 3.3). Um eine Vergleichbarkeit bezüglich der Validierungsstudie (Kramer *et al.*, 1999) zu gewährleisten, werden die vorgegebenen Parametereinstellungen verwendet. Eine Zusammenstellung der für den aus Basisfragmentauswahl, Basisfragmentplazierung und inkrementellem Aufbau bestehenden Algorithmus wichtigen Parameter ist in Tab. 3 gegeben.

Um die Abhängigkeit der Vorhersageergebnisse nativ-ähnlicher Konfigurationen von Protein und Ligand von der Generierungsmethode der relativen Anordnungen zu untersuchen, wurden für 100 kristallographisch bestimmte Komplexe[**] des Datensatzes DOCK_DS mit dem Dockingverfahren DOCK (Kuntz *et al.*, 1982) Rezeptor-Ligand-Geometrien erzeugt. DOCK_DS umfaßt zu mehr als 80 % Komplexe, die auch in FlexX_DS1 und FlexX_DS2 enthalten sind. Die Verteilungen molekularer Eigenschaften betragen: 0 – 27 drehbare Bindungen; 1 – 17 Wasserstoffbrückendonatoren und -akzeptoren; 2 – 39 Kohlenstoffatome. Die Aufbereitung der Eingabedaten folgte der Beschreibung von Ewing (Ewing, 1997).

[**] 1abe 1abf 1acj 1ack 1ase 1azm 1blh 1cbx 1cde 1cil 1cps 1ctr 1dbm 1did 1die 1dr1 1dwd 1ela 1frp 1ghb 1hfc 1hgi 1hgj 1hsl 1hti 1hyt 1icn 1imb 1ivc 1ivd 1ive 1ivf 1lah 1lcp 1lic 1lna 1lst 1mld 1mrg 1mrk 1mup 1nis 1nsc 1pbd 1phf 1poc 1pph 1ppl 1pso 1rds 1rnt 1rob 1snc 1srj 1tdb 1thy 1tlp 1tng 1tnh 1tni 1tpp 1ukz 1ulb 1wap 1xid 1xie 2ada 2ak3 2cgr 2cht 2cmd 2cpp 2gbp 2lgs 2mcp 2mth 2pk4 2r04 2r07 2sim 2tmn 2xis 2yhx 2ypi 3aah 3cpa 4cts 4est 4fab 4fbp 4phv 4tim 4tln 5abp 5p2p 6abp 6rnt 6tmn 7tim 8atc

**Tab. 3:** Für den inkrementellen Aufbaualgorithmus von FlexX wichtige Parameter, wie sie zur Erzeugung der Testdatensätze FlexX_DS1 und FlexX_DS2 verwendet wurden.

| | |
|---|---|
| **Behandlung von Überlappungen** | |
| max. zugelassenes Überlappungsvolumen zwischen einem Rezeptor- und einem Ligandatom | 2.5 $Å^3$ |
| max. mittleres zugelassenes Überlappungsvolumen zwischen Rezeptor- und Ligandatomen | 1.0 $Å^3$ |
| max. zugelassenes Überlappungsvolumen zwischen einem Wechselwirkungspunkt und einem Rezeptoratom | 2.5 $Å^3$ |
| **Selektion des Basisfragments** | |
| max. Anzahl von Fragmentierungen des Liganden für ein Basisfragment | 4 |
| **Basisfragmentplazierung** | |
| max. „Dreiecksfaktor“ | 2 |
| max. *rmsd* für die Zusammenfassung von plazierten Dreiecken von Ligandwechselwirkungszentren | 1.1 Å |
| max. *rmsd* für die Zusammenfassung von plazierten Paaren von Ligandwechselwirkungszentren | 0.4 Å |
| Schrittanzahl für die Diskretisierung des Plazierungswinkels | 2 |
| Schrittgröße für die Diskretisierung des Plazierungswinkels | 0.35 rad |
| **Inkrementeller Komplexaufbau** | |
| relative Energiegrenze, bis zu der Teillösungen in den nächsten Aufbauschritt gelangen | 209 kJ / mol |
| max. Anzahl von Teillösungen, die in den nächsten Aufbauschritt gelangen | 400 |
| zusätzliche Anzahl von Lösungen je Basisfragment, die in den nächsten Aufbauschritt gelangen | 100 |
| max. *rmsd* für die Zusammenfassung erzeugter Teillösungen | 0.7 Å |

Die Koordinaten der Ligandatome wurden aus den PDB-Dateien extrahiert und unter Verwendung der SYBYL-Notation (SYBYL) wurden Atom- und Bindungstypen gesetzt sowie Wasserstoffatome addiert. Atomladungen wurden jeweils nach dem Gasteiger-Marsili-Verfahren (Gasteiger & Marsili, 1980) berechnet. Abschließend erfolgte eine Energieminimierung unter Verwendung des TRIPOS-Kraftfeldes (Clark *et al.*, 1989) ohne Berücksichtigung elektrostatischer Energien. Dabei wurde darauf geachtet, daß sich die durch die Kristall-

struktur vorgegebenen Torsionswinkel nicht wesentlich verändern (der *rmsd*-Wert zwischen der Kristallstruktur und der minimierten Eingabestruktur beträgt dabei im Mittel 0.5 Å). Diese Ligandgeometrien wurden anschließend als Referenz zur Berechnung des *rmsd*-Wertes der gedockten Ligandanordnungen verwendet.

Die Daten für den Rezeptor wurden in mehreren Schritten aufbereitet.

- Aus der PDB-Datei wurden alle Nichtproteinatome (inklusive Wassermoleküle) sowie alle Wasserstoffatome entfernt, Aminosäuren auf ihre Vollständigkeit überprüft, die Hydroxylgruppe der C-terminalen Carboxylat-Gruppe mit der Kodierung OXT versehen und durch Disulfidbrücken verknüpfte Cysteine mit dem Monomertyp CYX benannt. Histidinseitenketten in der Nähe der Bindetasche wurden auf in Frage kommende benachbarte Wasserstoffbrückenbindungs-Partner untersucht und ihre Protonierung dementsprechend gesetzt. Als Bindetasche wurde ein Bereich von 7 Å um die ursprünglichen Ligandatompositionen gewählt.
- Unter Verwendung von AUTOMS aus der DOCK-Programmsammlung wurde mit einer Kugel von 1.4 Å Radius für den Bereich der Bindetasche die Lösemittel-zugängliche Oberfläche mit dem Programm MS von Connolly (Connolly, 1983) erzeugt. Mit SPHGEN wurde daraufhin der Bereich der Bindetasche mit Kugeln mit variablen Radien ausgefüllt, wobei die Kugeln mindestens zwei Kontaktpunkte mit der Connolly-Oberfläche der Tasche haben mußten. Die Größe des abschließend erhaltenen Clusters sollte dabei nicht über 100 Kugeln liegen; gegebenenfalls wurde die Lage einzelner Kugeln korrigiert.
- Um die während des Einpassens der Liganden in die Proteinbindetasche häufig erfolgende Berechnung der Wechselwirkungsenergie zwischen Ligand und Protein zu beschleunigen, wurden die nur vom Protein abhängigen Anteile für alle in DOCK möglichen Bewertungsfunktionen (Kontakt-Score, Kraftfeld-Score, „chemischer"-Score) (Ewing, 1997; Kuntz *et al.*, 1982; Meng *et al.*, 1992) für Punkte eines innerhalb der Bindetasche lokalisierten kubischen Gitters mit dem zu DOCK dazugehörigen Programm GRID vorherberechnet. Die Abstände der Gitterpunkte betragen hierzu 0.3 Å.

Obwohl DOCK bereits 1982 vorgestellt wurde (Kuntz *et al.*, 1982), existiert bislang keine Validierungsstudie, bei der eine große Anzahl verschiedener, kristallographisch bestimmter Rezeptor-Ligand-Komplexe unter Anwendung aller in DOCK möglichen Plazierungs- und Bewertungsalgorithmen gedockt wurde. Im Rahmen dieser Arbeit wurden daher die in DOCK4.0 zur Verfügung stehenden Möglichkeiten des Einpassens rigider Liganden (Ewing & Kuntz, 1997) sowie die des flexiblen Dockings von Liganden in die Proteinbindetasche

unter Verwendung eines inkrementellen Aufbaualgorithmus (Makino & Kuntz, 1997) kombiniert mit Bewertungsfunktionen basierend auf dem „Kontakt-Score“, dem „Kraftfeld-Score“ sowie dem „chemischen Score“ (Ewing, 1997; Kuntz *et al.*, 1982; Meng *et al.*, 1992) (s.a. Beschreibung in Kap. 3.3). Zusätzlich wurden die Bewertungen entweder direkt mit den von den Plazierungsalgorithmen erzeugten Protein-Ligand-Geometrien oder zusätzlich nach Anwendung einer auf einem Simplex-Algorithmus beruhenden Energieminimierung durchgeführt (Gschwend & Kuntz, 1996). Damit ergeben sich 12 verschiedene Dockingmöglichkeiten, die in Tab. 4 aufgeführt sind, und die auf die 100 Komplexe in DOCK_DS angewendet wurden.

In Tab. 5 werden jeweils für die Teilbereiche Ligandflexibilität, -plazierung, Bewertung und Minimierung wichtige Parameter aufgeführt, die gemäß den in Tab. 4 aufgeführten Kombinationsmöglichkeiten verwendet werden. Nichtaufgeführte Programmparameter folgen den empfohlenen Einstellungen (Ewing, 1997). Aus Rechenzeitgründen wird die maximale Anzahl an erhaltenen geometrischen Lösungen für rigides Docking auf 500 und für flexibles Docking auf 50 festgelegt.

**Tab. 4:** Für die Evaluierung von DOCK verwendete Kombinationen von Ligandplazierungsalgorithmen und Bewertungsverfahren für die erzeugten Rezeptor-Ligand-Geometrien.

| **Ligandplazierung** | **Bewertungsmethode** | **Minimierung[a)]** | **Abkürzung[b)]** |
|---|---|---|---|
| Rigides Docking | Kontakt-Score | nein | rig_cnt |
| | | ja | rig_cnt_min |
| | Kraftfeld-Score | nein | rig_nrg |
| | | ja | rig_nrg_min |
| | Chemischer Score | nein | rig_chm |
| | | ja | rig_chm_min |
| Flexibles Docking | Kontakt-Score | nein | flex_cnt |
| | | ja | flex_cnt_min |
| | Kraftfeld-Score | nein | flex_nrg |
| | | ja | flex_nrg_min |
| | Chemischer Score | nein | flex_chm |
| | | ja | flex_chm_min |

a) Verwendung einer auf einem Simplex-Algorithmus basierenden Energieminimierung der von den Plazierungsalgorithmen erzeugten Protein-Ligand-Geometrien. Die minimierte Energiefunktion entspricht dabei jeweils der Bewertungsmethode. b) Im folgenden benutzte Abkürzungen zur Kennzeichnung des jeweiligen Verfahrens.

**Tab. 5:** Für die Erzeugung von Rezeptor-Ligand-Komplexen mit DOCK verwendete wichtige Parameter.

| | |
|---|---|
| **Ligandflexibilität** | |
| Suche nach einem „Anker“ (Basisfragment) | ja |
| Verwendung mehrerer „Anker“ je Molekül | ja |
| min. Ankergröße | 10 Atome |
| Durchführung eines inkrementellen Molekülaufbaus | ja |
| Anzahl von Ligandkonfigurationen, die die nächste Stufe des inkrementellen Aufbaus erreichen | 50 |
| systematische Durchsuchung der Torsionswinkel | ja |
| **Ligandplazierung** | |
| Verwendung in der Bindetasche plazierter Kugeln zur Ligandorientierung | ja |
| Zufällige Ligandorientierung | nein |
| Generierung der gleichen Anzahl von Orientierungen für jeden Liganden | ja |
| Anzahl von Orientierungen, die für einen Liganden erzeugt werden | 500 |
| **Bewertung** | |
| Intramolekulare Energieberechnung (nur für flexible Liganden) | ja |
| Intermolekulare Energieberechnung | ja |
| Gitterbasierte Bewertung | ja |
| Verwendung eines Filters für Atom-Atom-Überlappungen | nein |
| Verwendung des Modells „vereinigter“ Atome | ja |
| **Minimierung** | |
| Relaxation im Torsionsraum (nur für flexible Liganden) | ja |
| Minimierung des Ankers (nur für flexible Liganden) | ja |
| Reminimierung des Ankers (nur für flexible Liganden) | ja |
| Reminimierung des Liganden (nur für flexible Liganden) | ja |
| Konvergenzkriterium bezogen auf jede Bewertungsfunktion | 0.1 |
| max. Anzahl von Minimierungszyklen | 2 |

### 4.9.2 Testdatensätze zur Vorhersage von Bindungsaffinitäten

Zur Validierung der Vorhersage von Bindungaffinitäten werden hier zum einen Rezeptor-Ligand-Komplexe mit kristallographisch bestimmter Geometrie und experimentell ermittelten $pK_i$-Werten verwendet. Allerdings beruhen Affinitätsvorhersagen im Rahmen von virtuellen Screening-Ansätzen auf von Hand modellierten oder mit Hilfe eines Dockingverfahrens erzeugten Protein-Ligand-Konfigurationen. In letzterem Fall muß die Bewertungsfunktion die wahrscheinlichste Geometrie zusätzlich in einer Menge alternativer Strukturen erkennen. Deshalb werden hier zusätzlich Datensätze benutzt, bei denen Ligandbindungsgeometrien mit dem Dockingprogramm FlexX erzeugt wurden.

Datensätze für in der PDB enthaltene Rezeptor-Ligand-Komplexe werden aus Arbeiten von Eldridge *et al.* (Eldridge *et al.*, 1997), Böhm (Böhm, 1994; Böhm, 1998) und Head *et al.* (Head *et al.*, 1996) entnommen, da sie bereits zur Validierung anderer Scoringfunktionen herangezogen wurden (Muegge & Martin, 1999) und so eine solide Vergleichsgrundlage bieten. Für die im folgenden aufgeführten Sätze wurden die Liganden und prosthetischen Gruppen aus den PDB-Dateien extrahiert und den Atomen SYBYL-Atomtypen zugewiesen. Die Berechnung der Bewertung nach Gl. 36 erfolgt unter Verwendung von Metallatomen und prosthetischen Gruppen als Teil des Proteins; Wassermoleküle werden – sofern nicht anders angegeben – nicht mit einbezogen.

Die folgenden Tab. 6 – Tab. 15 enthalten neben den Codes der aus der PDB verwendeten Protein-Ligand-Komplexe die experimentell bestimmten Bindungsaffinitäten in Form von $pK_i$- Werten. Tab. 6 umfaßt 16 Serinprotease-, Tab. 7: 15 Metalloprotease-, Tab. 8: 11 Endothiapepsin- und Tab. 9: 9 Komplexe des Arabinose-bindenden Proteins, die aus Eldridge *et al.* (Eldridge *et al.*, 1997) und Head *et al.* (Head *et al.*, 1996) entnommen wurden. Tab. 10 enthält einen von Muegge und Martin (Muegge & Martin, 1999) zusammengestellten Satz von 17 Komplexen aus der Arbeit von Böhm (Böhm, 1994), die alle nicht in Tab. 6 – Tab. 9 enthalten sind und eine Auflösung besser als 2.5 Å aufweisen. Dieser Datensatz wurde aus Vergleichsgründen mit der Arbeit von Muegge und Martin (Muegge & Martin, 1999) verwendet. Tab. 11 enthält 71 Komplexe aus einer Folgearbeit von Böhm (Böhm, 1998), die alle in der PDB abgelegt sind. Von Böhm modellierte Komplexe standen nicht zur Verfügung und konnten daher auch nicht verwendet werden. Für Referenzen bezüglich der angegebenen experimentellen Affinitäten sei auf die jeweiligen Originalarbeiten verwiesen.

**Tab. 6:** 16 Serinproteasekomplexe mit experimentell bestimmten Bindungsaffinitäten aus Tab. 2 und Tab. 6 der Arbeit von Eldridge *et al.* (Eldridge *et al.*, 1997).

| **PDB-Code** | $pK_i$ [a)] | **PDB-Code** | $pK_i$ [a)] |
|---|---|---|---|
| 1bra | 1.85 | 1tmt | 6.34 |
| 1dwb | 2.97 | 1tng | 2.98 |
| 1dwd | 8.62 | 1tnh | 3.42 |
| 1etr | 7.52 | 1tni | 1.72 |
| 1ets | 8.66 | 1tnj | 1.99 |
| 1ett | 6.29 | 1tnk | 1.51 |
| 1ppc | 6.56 | 1tnl | 1.90 |
| 1pph | 6.32 | 3ptb | 4.82 |

a) Für $T$ = 298 K aus den in Eldridge *et al.* (Eldridge *et al.*, 1997) angegebenen experimentellen Affinitäten berechnet.

**Tab. 7:** 15 Metalloproteasekomplexe mit experimentell bestimmten Bindungsaffinitäten aus Tab. 3 der Arbeit von Eldridge *et al.* (Eldridge *et al.*, 1997).

| **PDB-Code** | $pK_i$ [a)] | **PDB-Code** | $pK_i$ [a)] |
|---|---|---|---|
| 1cbx | 6.34 | 4tmn | 10.18 |
| 1mnc | 9.00 | 5tln | 6.36 |
| 1tlp | 7.55 | 5tmn | 8.04 |
| 1tmn | 7.30 | 6cpa | 11.52 |
| 2tmn | 5.88 | 6tmn | 5.04 |
| 3cpa | 3.88 | 7cpa | 13.96 |
| 3tmn | 5.90 | 8cpa | 9.14 |
| 4tln | 3.72 | - | - |

a) Für $T$ = 298 K aus den in Eldridge *et al.* (Eldridge *et al.*, 1997) angegebenen experimentellen Affinitäten berechnet.

**Tab. 8:** 11 Endothiapepsinkomplexe mit experimentell bestimmten Bindungsaffinitäten aus Tab. 6 der Arbeit von Eldridge *et al.* (Eldridge *et al.*, 1997) und Tab. 1 der Arbeit von Head *et al.* (Head *et al.*, 1996).

| **PDB-Code** | $pK_i$ [a)] | **PDB-Code** | $pK_i$ [a)] |
|---|---|---|---|
| 1eed | 4.79 | 2er9 | 7.79 |
| 1epo | 7.95 | 3er3 | 7.10 |
| 1epp | 7.16 | 4er1 | 6.61 |
| 2er0 | 6.40 | 4er4 | 6.79 |
| 2er6 | 7.22 | 5er2 | 6.56 |
| 2er7 | 9.00 | - | - |

a) Für $T$ = 298 K aus den in Eldridge *et al.* (Eldridge *et al.*, 1997) angegebenen experimentellen Affinitäten berechnet mit Ausnahme des Wertes von 2er0. Hier wurde der in Head *et al.* (Head *et al.*, 1996) direkt angegebene $pK_i$-Wert verwendet.

**Tab. 9:** 9 Arabinose-bindende Proteine enthaltende Komplexe mit experimentell bestimmten Bindungsaffinitäten aus Tab. 4 der Arbeit von Eldridge *et al.* (Eldridge *et al.*, 1997).

| PDB-Code | $pK_i$ [a)] | PDB-Code | $pK_i$ [a)] |
|---|---|---|---|
| 1abe | 7.01 | 6abp | 6.35 |
| 1abf | 5.41 | 7abp | 6.45 |
| 1apb | 5.82 | 8abp | 8.00 |
| 1bap | 6.85 | 9abp | 8.00 |
| 5abp | 6.63 | - | - |

a) Für $T$ = 298 K aus den in Eldridge *et al.* (Eldridge *et al.*, 1997) angegebenen experimentellen Affinitäten berechnet.

**Tab. 10:** 17 Komplexe aus dem kombinierten Trainings- und Testdatensatz der Arbeit von Böhm (Böhm, 1994), die nicht in Tab. 6 – Tab. 9 enthalten sind und eine Auflösung besser als 2.5 Å besitzen, mit ihren experimentell bestimmten Bindungaffinitäten.

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1fkf | 9.70 | 2tsc | 8.52 |
| 1mbi | 1.88 | 2xis | 5.82 |
| 1phf | 4.40 | 2ypi | 4.82 |
| 1phg | 8.66 | 4dfr | 9.70 |
| 1rbp | 6.72 | 4hvp | 6.15 |
| 1rne | 9.40 | 4phv | 9.15 |
| 2cpp | 6.07 | 5cna | 2.00 |
| 2gbp | 7.60 | 5cpp | 5.88 |
| 2ifb | 5.43 | - | - |

**Tab. 11:** 71 Komplexe aus dem kombinierten Trainings- und Testdatensatz der Arbeit von Böhm (Böhm, 1998), die auch in der PDB enthalten sind, mit ihren experimentell bestimmten Bindungaffinitäten.

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1acj | 7.30 | 2gbp | 7.60 |
| 1add | 6.74 | 2gpb | 2.77 |
| 1bzm | 6.03 | 2ifb | 5.43 |
| 1cbx | 6.30 | 2phh | 8.27 |
| 1cil | 9.43 | 2r04 | 6.38 |
| 1cps | 6.66 | 2tmn | 4.67 |
| 1ctt | 4.52 | 2tsc | 8.52 |
| 1dwb | 2.92 | 2xis | 5.82 |
| 1dwc | 7.40 | 2ypi | 4.82 |
| 1ela | 6.35 | 3cpa | 3.88 |
| 1elc | 7.15 | 3dfr | 10.30 |
| 1fkf | 9.70 | 3ptb | 4.74 |
| 1hpv | 9.22 | 3tpi | 4.30 |
| 1hvr | 9.51 | 4dfr | 9.70 |
| 1l83 | 3.75 | 4er4 | 6.79 |
| 1ldm | 5.40 | 4fab | 10.53 |

Fortsetzung von Tab. 11:

| | | | |
|---|---|---|---|
| 1mbi | 1.88 | 4gr1 | 1.70 |
| 1phe | 5.70 | 4hmg | 2.55 |
| 1phf | 4.40 | 4hvp | 6.15 |
| 1phg | 8.66 | 4phv[a)] | 9.15 |
| 1ppc | 6.46 | 4tln | 3.72 |
| 1pph | 6.22 | 4tmn | 10.19 |
| 1pso | 10.34 | 4ts1 | 5.60 |
| 1r09[a)] | 4.90 | 5cna | 2.00 |
| 1rbp | 6.72 | 5cpp | 5.88 |
| 1rne | 9.40 | 5tim | 2.30 |
| 1sbp | 6.92 | 5tln | 6.37 |
| 1sre | 4.00 | 5tmn | 8.04 |
| 1stp | 13.40 | 6acn | 3.00 |
| 1tlp | 7.55 | 6cpa | 11.52 |
| 1tmn | 7.30 | 6rsa | 5.00 |
| 1tnk | 1.49 | 7cpa | 14.00 |
| 1ulb | 5.30 | 7cpp | 3.80 |
| 2cpp | 6.07 | 9aat | 8.22 |
| 2ctc | 3.89 | 9hvp | 8.35 |
| 2er6 | 7.22 | - | - |

a) $IC_{50}$-Wert.

Für die Datensätze, bei denen Ligandbindungsgeometrien mit dem Dockingprogramm FlexX erzeugt wurden, wurden zum einen eine Serie von 2 x 32 Inhibitoren aus der Arbeit von Obst *et al.* (Obst, 1997; Obst *et al.*, 1997) verwendet, für die Inhibitionskonstanten gegenüber Thrombin bzw. Trypsin bestimmt worden waren (Tab. 12). Diese Inhibitoren wurden analog zu dem in Kap. 4.9.1 beschriebenen Vorgehen unter Verwendung der Parameter in Tab. 3 (S. 85) mit FlexX in die Rezeptorstrukturen von Obst *et al.* (Obst *et al.*, 1997) (für Thrombin) und 1pph (für Trypsin) gedockt. Ebenfalls mit FlexX wurden 61 bzw. 15 Thermolysininhibitoren mit bekannten Bindungsaffinitäten aus dem Trainings- bzw. Testdatensatz (beschrieben in Klebe *et al.* (Klebe *et al.*, 1994)) in die Proteinstruktur von 1tlp gedockt (Tab. 13, Tab. 14). Um neben diesen Datensätzen für Serien von Inhibitoren in bezug auf *ein* Protein auch die Vorhersage von Bindungsaffinitäten für in *unterschiedliche* Proteine gedockte Liganden zu untersuchen, wurden aus den Sätzen FlexX_DS1 und FlexX_DS2 53 Protein-Ligand-Komplexe extrahiert, für die in den Arbeiten von Eldrige *et al.* (Eldridge *et al.*, 1997), Head *et al.* (Head *et al.*, 1996) bzw. Böhm (Böhm, 1994; Böhm, 1998) eine experimentelle Bindungsaffinität berichtet wurde (Tab. 15). Für alle Datensätze wurden Bindungsaffinitäten nach Gl. 42 jeweils für die mit Gl. 36 bestbewertete Ligandgeometrie aus einer mit FlexX erzeugten Menge von Lösungen berechnet.

**Tab. 12:** 32 Inhibitoren aus der Arbeit von Obst *et al.* (Obst, 1997; Obst *et al.*, 1997) mit experimentell bestimmten Bindungsaffinitäten gegenüber Thrombin und Trypsin.

| **Bezeichnung**[a)] | $pK_i$ **(Thrombin)** | $pK_i$ **(Trypsin)** |
|---|---|---|
| UO54 | 6.17 | 5.31 |
| UO62a | 5.95 | 5.29 |
| UO62b | 3.77 | 3.37 |
| UO62c | 4.95 | 4.85 |
| UO62d | 5.53 | 4.65 |
| UO62e | 3.79 | 3.43 |
| UO62f | 5.82 | 5.29 |
| UO62g | 6.00 | 5.55 |
| UO62h | 6.65 | 5.58 |
| UO62i | 7.04 | 6.15 |
| UO62j | 6.45 | 5.27 |
| UO62k | 6.35 | 5.05 |
| UO62l | 5.00 | 5.22 |
| UO63i | 4.72 | 4.30 |
| UO63k | 5.82 | 5.49 |
| UO63l | 5.85 | 5.30 |
| UO67 | 6.56 | 5.42 |
| UO68 | 5.13 | 5.49 |
| UO71 | 5.29 | 4.49 |
| UO75 | 6.30 | 5.95 |
| UO89 | 5.20 | 4.50 |
| UO90 | 5.69 | 4.19 |
| UO95 | 5.03 | 4.38 |
| UO109 | 5.69 | 4.19 |
| UO110 | 7.02 | 5.13 |
| UO111 | 7.52 | 5.17 |
| UO112 | 6.06 | 4.48 |
| UO128 | 8.09 | 5.77 |
| UO129 | 8.00 | 5.63 |
| UO130 | 7.88 | 5.00 |
| UO131 | 5.77 | 5.35 |
| UO132 | 5.85 | 5.28 |

a) Die Nummern der Bezeichnungen sind identisch mit den in Obst (Obst, 1997) angegebenen.

**Tab. 13:** 61 Thermolysininhibitoren aus dem Trainingsdatensatz von Klebe *et al.* (Klebe *et al.*, 1994), für die experimentell bestimmte Bindungsaffinitäten bekannt sind.

| **Bezeichnung**[a)] | $pK_i$ | **Bezeichnung**[a)] | $pK_i$ |
|---|---|---|---|
| ACE_OHLEU_AGNH2 | 2.47 | PO3_FAGNH2 | 5.59 |
| BZSAG | 6.12 | PPHEOH | 4.14 |
| C6PCLTNME | 7.28 | R_THIORPHAN | 5.64 |
| C6PLTNME | 8.82 | S02P_FAGNH2 | 5.16 |
| C6POLTNME | 5.84 | S_THIORPHAN | 5.74 |
| CBZPHE | 3.29 | SO3_FAGNH2 | 2.37 |
| CH3COCH2CO_FAGNH2 | 2.51 | Z_D_APOLA | 4.62 |
| CH3O2S_FAGNH2 | 0.52 | Z_D_FPLA | 6.32 |
| CHO_OHLEU_AGNH2 | 2.47 | Z_D_FPOLA | 4.52 |
| CLTZNCRYS | 7.47 | Z_D_LPOLA | 4.38 |
| DAH50 | 7.96 | Z_NH_GLNH2 | 3.42 |
| DAH51 | 6.22 | Z_NH_GLNHOH | 5.57 |
| DAH52 | 5.55 | ZALA | 6.07 |
| DAH53 | 6.66 | ZAPOLA | 5.74 |
| DAH54 | 5.77 | ZFPLAZNCRYS | 10.17 |
| DAH55 | 2.42 | ZFPOLA | 7.35 |
| HOCH2CO_FAGNH2 | 2.54 | ZG_D_LNHOH | 4.32 |
| NHOHBZMAGNA | 6.37 | ZGG_D_LNHOH | 3.60 |
| NHOHBZMAGNH2 | 6.18 | ZGGLNHOH | 4.41 |
| NHOHBZMAGOH | 6.18 | ZGGNHOH | 3.03 |
| NHOHBZMOET | 4.70 | ZGLNH2 | 1.68 |
| NHOHIBMAGNH2 | 6.32 | ZGLNHOH | 4.89 |
| NHOHLEU | 3.72 | ZGLNMEOH | 2.65 |
| NHOHMALAGNH2 | 2.96 | ZGLY | 6.39 |
| OHBZMAGNH2 | 3.38 | ZGPCLLZNCRYS | 6.74 |
| P_ILE_AOH | 6.44 | ZGPLA | 7.78 |
| P_OPHE_OME_LEUNH2 | 0.52 | ZGPLLZNCRYS | 8.04 |
| PAAOH | 4.06 | ZGPOLA | 4.89 |
| PHOSPHORAMIDON | 7.55 | ZGPOLLZNCRYS | 5.05 |
| PLEUNH2 | 4.10 | ZLPOLA | 6.17 |
| PNHET | 0.52 | - | - |

a) Die Bezeichnungen folgen den in Klebe *et al.* (Klebe *et al.*, 1994) angegebenen.

**Tab. 14:** 15 Thermolysininhibitoren aus dem Testdatensatz von Klebe *et al.* (Klebe *et al.*, 1994), für die experimentell bestimmte Bindungsaffinitäten bekannt sind.

| **Bezeichnung**[a)] | $pK_i$ | **Bezeichnung**[a)] | $pK_i$ |
|---|---|---|---|
| PLFOH | 7.72 | ZGPLG | 6.57 |
| PPPHE | 2.79 | ZGPLNH2 | 6.12 |
| ZFGNH2 | 3.46 | ZGPOLF | 4.27 |
| ZGPCLA | 7.73 | ZGPOLG | 3.64 |
| ZGPCLF | 7.18 | ZGPOLNH2 | 3.18 |
| ZGPCLG | 6.52 | ZLGNH2 | 2.51 |
| ZGPCLNH2 | 5.85 | ZYGNH2 | 3.66 |
| ZGPLF | 7.12 | - | - |

a) Die Bezeichnungen folgen den in Klebe *et al.* (Klebe *et al.*, 1994) angegebenen.

**Tab. 15:** 53 Protein-Ligand-Komplexe, die aus den Datensätzen FlexX_DS1 und FlexX_DS2 extrahiert wurden und für die eine Bindungsaffinität in Eldrige *et al.* (Eldridge *et al.*, 1997), Head *et al.* (Head *et al.*, 1996) bzw. Böhm (Böhm, 1994; Böhm, 1998) berichtet wurde.

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1aaq | 8.40 | 1tni | 1.70 |
| 1abe | 7.02 | 1tnk | 1.49 |
| 1abf | 5.42 | 1tnl | 1.88 |
| 1apt | 9.40 | 2cgr | 7.28 |
| 1cbx | 6.30 | 2cpp | 6.07 |
| 1cps | 6.66 | 2ctc | 3.89 |
| 1dwd | 8.48 | 2er6 | 7.22 |
| 1eed | 4.79 | 2gbp | 7.60 |
| 1ela | 6.35 | 2tmn | 5.89 |
| 1elc | 7.15 | 2xis | 5.82 |
| 1etr | 7.40 | 2ypi | 4.82 |
| 1hsl | 7.30 | 3cpa | 3.89 |
| 1hvr | 9.51 | 3ptb | 4.74 |
| 1ldm | 5.40 | 4dfr | 9.70 |
| 1mbi | 1.88 | 4hmg | 2.55 |
| 1nsc | 5.44 | 4hvp | 6.11 |
| 1phf | 4.40 | 4phv | 9.15 |
| 1phg | 8.66 | 4tln | 3.72 |
| 1ppc | 6.46 | 4tmn | 10.19 |
| 1pph | 6.22 | 4ts1 | 5.60 |
| 1ppk | 7.66 | 5abp | 6.64 |
| 1pso | 10.34 | 5tmn | 8.04 |
| 1rbp | 6.72 | 6abp | 6.36 |
| 1rne | 9.40 | 6cpa | 11.52 |
| 1tlp | 7.55 | 6tmn | 5.05 |
| 1tng | 2.93 | 7cpa | 14.00 |
| 1tnh | 3.37 | - | - |

### 4.9.3 Testdatensätze für virtuelles Screening

Für die Untersuchung der Güte der entwickelten Bewertungsfunktion für Anwendungen im Rahmen des virtuellen Screenings wurden zwei Datensätze aufgebaut.

Im ersten Fall wurden Strukturen von 31 Verbindungen aus der Arbeit von Murray *et al.* mit bekannten Inhibitionskonstanten ($pK_i$-Werte im Bereich zwischen 2 und 7) für Thrombin und Trypsin (Abb. 1 und Tab. 1 in (Murray *et al.*, 1998)) mit Hilfe von SYBYL (SYBYL) erzeugt (s.a. Kap. 4.9.1). Diese Inhibitoren bilden den Satz der „Aktiven“. Unter Verwendung von UNITY (UNITY) wurden in der Moleküldatenbank ACD (***A**vailable **C**hemicals **D**irectory*) (MDL) der Version 98.2 mit 183564 Einträgen zunächst alle Moleküle mit nichtzyklischen Amidino- und Amidinium-Funktionen gesucht. Dabei wurden 2398 bzw. 105 Einträge erhalten. Eine Auswahl aller Moleküle mit einem Molekulargewicht < 500 Da lieferte

daraus 977 Einträge. Mit CORINA (Gasteiger *et al.*, 1990) wurden anschließend 3D-Geometrien erzeugt, wobei die Optionen *wh* (Ausgabe mit Wasserstoffatomen), *rs* (Entfernen kleiner Fragmente, z.B. Gegenionen) und *neu* (Neutralisieren geladener funktioneller Gruppen) verwendet wurden. Nach dem Entfernen sich nur durch den Einbau verschiedener Isotope unterscheidender, ansonsten doppelt vorhandener Moleküle verblieben 824 Einträge. Bei ihnen wurden alle Carboxyl-, Phosphorsäure und Sulfonsäure-Gruppen deprotoniert, wohingegen aliphatische Amine sowie nicht-zyklische Amidino- und Guanidinogruppen protoniert wurden. Hierzu sowie zur Entfernung doppelt vorhandener Moleküle wurden eigene Programme beruhend auf Subgraphen-Algorithmen verwendet. Diese Moleküle bilden den Satz der „Inaktiven". Beide Sätze wurden mit FlexX (Rarey *et al.*, 1996a) - wie in Kap. 4.9.1 beschrieben - in die Rezeptorstrukturen der PDB-Einträge 1dwd (Thrombin) und 1pph (Trypsin) gedockt. Die Bewertung der jeweils bis zu 500 erhaltenen Dockinglösungen eines Liganden bzgl. des Proteins erfolgte mit Gl. 36.

### 4.9.4 Testdatensätze zur Untersuchung der impliziten Berücksichtigung von Direktionalität in Paar-Potentialen

Zur Untersuchung der impliziten Berücksichtigung von Direktionalität wurden die 159 Protein-Ligand-Kristallstrukturen aus den Testdatensätzen zur Bestimmung nativ-ähnlicher Protein-Ligand-Konfigurationen FlexX_DS1 und FlexX_DS2 vereinigt. Da hier bereits Proteine und Liganden aus den jeweiligen PDB-Dateien extrahiert wurden und die Liganden mit SYBYL-Atomtypnotation vorlagen (s.a. Kap. 4.9.1), konnten die Daten direkt für die Berechnungen eingesetzt werden.

### 4.9.5 Trainings- und Testdatensatz für die proteinspezifische Adaptierung der Bewertungsfunktion durch Einbeziehung von Zusatzinformation

Die zur Untersuchung der proteinspezifischen Adaptierung der Bewertungsfunktion herangezogenen Trainings- und Testdatensätze umfassten die 61 bzw. 15 in Tab. 13 (S. 95) und Tab. 14 (S. 95) zusammen mit ihren experimentell bestimmten $pK_i$-Werten aus der Arbeit von Klebe *et al.* (Klebe *et al.*, 1994) aufgeführten Verbindungen. Ihre Auswahl erfolgte aus zwei Gründen. Zum einen wurden alle 61 Trainingsverbindungen sowie mindestens 11 der 15 Testverbindungen in Arbeiten von Waller und Marshall (Waller & Marshall, 1993), DePriest *et al.* (De Priest *et al.*, 1993) sowie Klebe *et al.* (Klebe & Abraham, 1999) verwendet, um den

Einfluß verwendeter Überlagerungstechniken und unterschiedlicher Deskriptorfelder auf die Güte des erhaltenen Modells zu untersuchen. Somit können die Ergebnisse der hier vorgestellten Methode mit denen in den erwähnten Arbeiten verglichen werden. Zum anderen sind für 8 der 61 Verbindungen Kristallstrukturdaten in der PDB verfügbar (PDB-Codes: 1tlp, 1tmn, 2tmn, 4tln, 4tmn, 5tln, 5tmn, 6tmn).

Von den Kristallstrukturen wurden daher 1tlp, 1tmn, 4tmn und 5tmn als Template für den strukturellen Aufbau der Liganden verwendet; die für die Berechnung der Felder (s.a. Kap. 4.8.1) zugrundegelegte Proteinstruktur stammt aus 1tlp. Unter der generellen Annahme deprotoniert vorliegender Carbonsäure- und Phosphorsäuregruppen (ansonsten folgt die Wahl der Atomtypen der Liganden der in Tab. 2 (S. 65) beschriebenen) wurden die Liganden der vier oben erwähnten Kristallstrukturen zunächst in der starr gehaltenen Bindetasche von 1tlp mit dem Kraftfeld MAB (Gerber, 1998; Gerber & Müller, 1995) des Modellierungsprogramms MOLOC minimiert, wobei kontrolliert wurde, daß sich die durch die Kristallgeometrien vorgegebenen Torsionswinkel nicht wesentlich ändern. Hierdurch sollten insbesondere Bindungslängen und -winkel auf die im MAB-Kraftfeld vorgegebenen Parameter angepaßt werden. Die weiteren Verbindungen des Trainings- und Testdatensatzes wurden nun sukzessive mit MOLOC aufgebaut, wobei zu den Templatmolekülen unveränderte Grundgerüste sowie Wechselwirkungszentren der Liganden mit dem Protein direkt überlagert wurden; sterische Überlappungen mit dem Protein wurden verhindert. Abschließend wurden die Liganden mit dem MAB-Kraftfeld in der Bindetasche minimiert.

# 5 Ergebnisse und Diskussion

In dem folgenden Kapitel 5.1 werden zunächst die Eigenschaften der abgeleiteten distanzabhängigen Paarpotentiale beschrieben. Die von der Lösemittel-zugänglichen Oberfläche abhängigen Einteilchenpotentiale werden anschließend in Kapitel 5.2 aufgeführt. Kapitel 5.3 gibt eine kritische Betrachtung des gewählten wissensbasierten Ansatzes. Kapitel 5.4 stellt die Ergebnisse bei der Bewertung nativ-ähnlicher Komplexgeometrien dar. Zu Beginn wird dabei eine Validierung des Docking-Programms DOCK (Kuntz *et al.*, 1982; Makino & Kuntz, 1997; Meng *et al.*, 1992) vorgestellt. In Kapitel 5.5 wird die entwickelte Bewertungsfunktion zur Priorisierung von Liganden verwendet und ein Beispiel zum virtuellen Screening von Substanzbibliotheken vorgestellt. Die Untersuchung der impliziten Berücksichtigung gerichteter Wechselwirkungen in den Paarpotentialen und ihre potentielle Anwendbarkeit zur Ligandoptimierung wird in Kapitel 5.6 behandelt. Kapitel 5.7 stellt abschließend eine Methode vor, wie die abgeleiteten Paarpotentiale unter Einbeziehung zusätzlicher Informationen proteinspezifisch adaptiert werden können.

## *5.1 Eigenschaften distanzabhängiger Paarpotentiale*

Die in diesem Abschnitt beschriebenen Paarpotentiale entsprechen denjenigen, die nach Gl. 15 bzw. Gl. 36 zur Vorhersage von Struktur und Bindungsaffinität von Protein-Ligand-Komplexen im weiteren Verlauf der Arbeit verwendet werden. Die für sie getroffene Wahl des Referenzzustandes gemäß Gl. 24, der Intervallparameter und Parameter der Glättungsfunktion (s.a. Kap. 4.2.3) sowie die Art der Behandlung von Verteilungen geringer Datenanzahl gemäß Gl. 26 mit $\chi = 10^{-4}$ hat sich unter Verwendung des Kalibrierungsdatensatzes FlexX_DS1 als optimal erwiesen (s.a. Kap. 5.4.5). Eine Korrektur des zur Verfügung stehenden Kugelschalenvolumens nach Gl. 30 wird hierbei nicht angewendet. Ergebnisse unter Verwendung der in Kap. 4.2 vorgestellten Parameter- und Berechnungsalternativen werden ebenfalls in Kap. 5.4.5 dargestellt. In diesem Kapitel wird allerdings der Einfluß der Qualität, Größe und Zusammensetzung des zur Ableitung der Paarpotentiale verwendeten Datensatzes auf ihren Verlauf diskutiert. Für alle im folgenden auftretenden Fälle bezieht sich hierbei der Index des ersten Atomtyps auf ein Ligandatom, der Index des zweiten Atomtyps auf ein Proteinatom. Da bei der Verwendung von 17 Atomtypen insgesamt 289 mögliche Paarkombinationen auftreten, wird außerdem nur eine Auswahl der erhaltenen Potentiale vorgestellt.

### 5.1.1 Auftrittshäufigkeiten von Paarwechselwirkungen

Paarverteilungsfunktionen für Protein-Ligand-Atompaare wurden unter Verwendung von Gl. 13 durch Auszählen der Auftrittshäufigkeit der jeweiligen Paare nach Gl. 14 bestimmt. Basierend auf 1376 kristallographisch bestimmten Protein-Ligand-Komplexen und über den gesamten Intervallbereich von 1 bis 6 Å integriert, ist die Auftrittshäufigkeit aller Paarwechselwirkungen in Tab. 16 aufgeführt.

Die für die hier vorgestellten Beispiele (Abb. 10, S. 107 und Abb. 11, S. 110) geringste Zahl von Auftrittshäufigkeiten ergibt sich für die O.3-O.3-Wechselwirkung mit 7106, die höchste Anzahl wird für die C.3-C.3-Wechselwirkung mit 118848 gefunden. O.co2-N.pl3, O.3-O.co2 und C.ar-C.ar liegen mit 7971, 11382 und 26806 Einträgen dazwischen. Im O.3-O.3-Fall bedeutet dieses, daß im Durchschnitt mehr als 140 Einträge je Intervall der Weite 0.1 Å gefunden werden. Obwohl diese mittlere Anzahl einen Eindruck von der statistischen Signifikanz der Verteilung gibt, ist zu bemerken, daß v.a. für Distanzen < 2.4 Å die Anzahl der Einträge deutlich geringer wird. Die Verteilung in diesem Distanzbereich ist daher wesentlich weniger signifikant.

Betrachtet man alle Paarverteilungen, so sind in 172 Fällen (entsprechend 60 %) weniger als 500 Einträge insgesamt vorhanden (grau unterlegte Felder in Tab. 16), d.h. weniger als 10 Einträge je Intervall im Mittel. In 156 Fällen davon (d.h. in 54 % aller Fälle) hat das Ligand- oder das Proteinatom den Typ S.3, P.3, C.cat, Metall, F, Cl oder Br. Das seltene Auftreten von S.3, F, Cl und Br im zur Ableitung der Potentiale verwendeten Datensatz wird dabei auch in den verwendeten Testdatensätzen reflektiert. Solange Wechselwirkungen zwischen S.3/F/Cl/Br-X-Kontakten (X sei ein beliebiger Typ) die gesamten Wechselwirkungen zwischen Protein und Ligand nicht dominieren, kann daher von der Annahme ausgegangen werden, daß die Bewertung der Gesamtwechselwirkungen auf Grundlage der häufiger populierten Verteilungen verläßlich durchführbar ist.

Für den Fall der P.3-Atome gilt insbesondere, daß sie sehr häufig tetraedrisch von Sauerstoff-, Kohlenstoff und Stickstoffatomen unter Bildung von Phosphat-, Phosphonat- und Phosphinat-Derivaten umgeben sind. Derart vergraben sollte der größte Anteil der Wechselwirkungen dieser funktionellen Gruppen durch die häufig populierten Verteilungen der umgebenden Atome und weniger durch P.3 an sich bedingt werden. Wenn auch abgeschwächt, kann dieses ebenfalls für C.cat-Atome angenommen werden, die das Kohlenstoffatom in Amidino- und Guanidinofunktionen beschreiben. Zumindest in der Ebene dieser funktionellen Gruppen ist das Kohlenstoffatom dann von Stickstoffatomen abgeschirmt.

Die in Tab. 16 auf der Proteinatomseite aufgeführten P.3-, F-, Cl-, Br- und Metall-Typen resultieren aus der Einbeziehung von prosthetischen Gruppen bzw. Kofaktoren als Teile der Proteine während der Ableitung der Potentiale. Besonders die Halogen-Typen sind dabei nur schwach populiert; für X-Br-Verteilungen mit X als beliebigem Typ werden überhaupt keine Kontakte gefunden. Allerdings ermöglicht die Anwendung der Dreiecksfunktion zur Glättung der Rohdaten und die Behandlung von Verteilungen mit einer geringen Anzahl von Beobachtungen gemäß Gl. 26 auch die Verwendung der seltenen Informationen im Falle der F- und Cl-Typen. Auswirkungen durch das Ausschließen dieser Atomtypen auf die Vorhersagekraft der Potentiale werden in Kap. 5.4.5 diskutiert. Atome des Typs Metall auf der Ligandseite resultieren hingegen aus der Betrachtung von Metallatomen als Liganden während der Potentialerstellung. Eine mögliche Anwendbarkeit für Potentiale mit diesem Ligandatomtyp besteht in der Identifikation von Metallbindungszentren in Proteinen.

### 5.1.2 Referenzzustand der Paarwechselwirkungen

Bei der Interpretation einzelner Merkmale in den Verteilungsfunktionen sowie in den daraus abgeleiteten Potentialen ist zu berücksichtigen, daß die Paarverteilungsfunktionen Mittelungen über die gesamte betrachtete Protein-Ligand-Datenbank sind. Sie enthalten daher implizit sowohl Informationen über in kondensierten Systemen auftretende Vielkörperwechselwirkungen als auch Informationen über Wechselwirkungen, die von verschiedenem physikalischem Ursprung sind (etwa elektrostatische und sterische Wechselwirkungen sowie Lösemitteleffekte). Aus diesem Grund ist eine Erklärung einzelner Merkmale auf Grundlage der Wechselwirkungen zwischen Atomen eines isoliert betrachteten Paares und somit eine Separation in Zweikörperwechselwirkungen nur bedingt möglich (Moult, 1997).

Der in Abb. 10 (S. 107) jeweils mit angegebene Referenzzustand wird durch Mittelung über alle normierten radialen Paarverteilungsfunktionen gemäß Gl. 24 berechnet. Dementsprechend kann er als Repräsentation von Wechselwirkungen zwischen Atomen eines *allgemeinen* Typs angesehen werden, d.h. er enthält hauptsächlich unspezifische Informationen bedingt durch generelle Packungseffekte (Sippl, 1990; Sippl, 1993).

Eine dennoch zu bemerkende Strukturierung im Verlauf des Referenzzustandes kann für kleine Abstände um 2 Å auf das Auftreten von Kontakten zu Metallatomen zurückgeführt werden. Eine Schulter bei 2.7 Å tritt durch in diesem Bereich besonders häufige polare bzw. geladene Wechselwirkungen auf. Die erhöhte Wahrscheinlichkeit um 4 Å geht vermutlich auf häufig auftretende aromatische Wechselwirkungen sowie Wechselwirkungen in sekundär-

strukturartigen Faltblattanordnungen peptidischer Liganden mit umgebenden Proteinketten zurück. Eine Strukturierung bei noch größeren Abständen läßt sich durch das Auftreten von Wechselwirkungsmustern in der zweiten Koordinationssphäre um ein betrachtetes Zentralatom erklären.

**Tab. 16:** Auftrittshäufigkeiten von Atom-Atom-Kontakten in einem Abstand von 1 bis 6 Å, ermittelt aus 1376 kristallographisch bestimmten Protein-Ligand-Komplexen. Paarverteilungen mit weniger als 500 Kontakten sind grau unterlegt dargestellt.

| Ligand- | Proteinatome | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| atome | C.3 | C.2 | C.ar | C.cat | N.3 | N.ar | N.am | N.pl3 | O.3 | O.2 | O.co2 | S.3 | P.3 | F | Cl | Br | Met |
| **C.3** | 118848 | 54466 | 55377 | 3029 | 3213 | 3157 | 35597 | 7719 | 14044 | 28911 | 17783 | 2177 | 73 | 41 | 40 | 0 | 715 |
| **C.2** | 54215 | 24267 | 22480 | 1432 | 974 | 2174 | 14709 | 3888 | 5120 | 13347 | 4999 | 1313 | 37 | 3 | 12 | 0 | 276 |
| **C.ar** | 65316 | 21657 | 26806 | 1522 | 463 | 1263 | 13922 | 3872 | 5851 | 14175 | 3747 | 1550 | 31 | 6 | 35 | 0 | 462 |
| **C.cat** | 545 | 351 | 80 | 9 | 0 | 1 | 272 | 18 | 65 | 222 | 127 | 44 | 0 | 0 | 0 | 0 | 0 |
| **N.3** | 3147 | 1895 | 1507 | 70 | 55 | 66 | 1244 | 238 | 497 | 1026 | 795 | 55 | 6 | 0 | 4 | 0 | 20 |
| **N.ar** | 19110 | 7223 | 6448 | 395 | 245 | 430 | 4932 | 1089 | 1509 | 4627 | 1638 | 363 | 23 | 0 | 0 | 0 | 60 |
| **N.am** | 19961 | 8687 | 6940 | 404 | 245 | 481 | 6818 | 1110 | 1559 | 5171 | 2070 | 319 | 15 | 0 | 0 | 0 | 80 |
| **N.pl3** | 1190 | 735 | 221 | 20 | 1 | 1 | 568 | 61 | 131 | 532 | 307 | 91 | 0 | 0 | 0 | 0 | 0 |
| **O.3** | 58375 | 31537 | 25289 | 1973 | 2416 | 1878 | 21717 | 5062 | 7106 | 16596 | 11382 | 936 | 73 | 59 | 5 | 0 | 447 |
| **O.2** | 17478 | 8971 | 8367 | 392 | 266 | 603 | 6044 | 983 | 1501 | 5901 | 1984 | 340 | 16 | 2 | 4 | 0 | 110 |
| **O.co2** | 47930 | 22976 | 11097 | 3153 | 2465 | 1401 | 17417 | 7971 | 6780 | 13739 | 5250 | 586 | 128 | 127 | 0 | 0 | 317 |
| **S.3** | 5363 | 1659 | 1507 | 78 | 27 | 137 | 1325 | 244 | 604 | 959 | 238 | 340 | 2 | 0 | 5 | 0 | 23 |
| **P.3** | 11440 | 5756 | 1516 | 694 | 711 | 292 | 4736 | 1724 | 1905 | 3286 | 1288 | 108 | 37 | 53 | 0 | 0 | 67 |
| **F** | 1462 | 533 | 477 | 30 | 15 | 40 | 412 | 74 | 171 | 301 | 75 | 45 | 0 | 10 | 0 | 0 | 30 |
| **Cl** | 317 | 96 | 104 | 20 | 4 | 9 | 72 | 52 | 42 | 52 | 13 | 2 | 0 | 0 | 14 | 0 | 5 |
| **Br** | 76 | 11 | 20 | 0 | 1 | 0 | 6 | 0 | 3 | 10 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Met** | 1584 | 644 | 469 | 43 | 0 | 62 | 597 | 110 | 128 | 238 | 95 | 332 | 0 | 0 | 0 | 0 | 0 |

### 5.1.3 Individuelle Paarverteilungen und daraus abgeleitete statistische Präferenzen für Paarwechselwirkungen

Um nativ-ähnliche Protein-Ligand-Geometrien in einer Menge von erzeugten Komplexstrukturen erkennen sowie verschiedene Liganden bzgl. eines oder mehrerer Proteine korrekt priorisieren zu können, müssen die erhaltenen typspezifischen Paarverteilungen sowie die daraus nach Gl. 11 berechneten Paarpotentiale ausreichend voneinander verschieden sein.

Für die Beurteilung der Ähnlichkeit jeweils zweier Verteilungen $\rho_A$ bzw. $\rho_B$ kann das Skalarprodukt analoger Verteilungswerte verwendet werden, wobei die Summation über alle Intervalle $I$ der diskreten Verteilungen läuft:

$$sim'(\rho_A, \rho_B) = \frac{\sum_{i \in I} \rho_A(i)\ \rho_B(i)}{\left(\sum_{i \in I} \rho_A(i)^2\right)^{1/2} \left(\sum_{i \in I} \rho_B(i)^2\right)^{1/2}} \qquad \textbf{Gl. 64}$$

Bedingt durch die Normalisierung im Nenner ergibt sich für identische Verteilungen ein Wert von 1, für sich bei jedem Intervall genau im Vorzeichen des Wertes unterscheidende Verteilungen ein Wert von –1. Ein Wert von 1 wird allerdings auch für zwei Verteilungen erhalten, deren Werte sich nur um einen konstanten (positiven) Faktor voneinander unterscheiden. Das auf Gl. 64 basierende Maß beschreibt also die Ähnlichkeit der *Form* der Verteilungen, nicht jedoch die Ähnlichkeit der Größenordnungen ihrer einzelnen Werte. Im Rahmen der Ähnlichkeitsbeurteilung von Molekülen wurde von Hodgkin und Richards (Hodgkin & Richards, 1987) ein Maß vorgeschlagen, das diesen Nachteil umgeht:

$$sim(\rho_A, \rho_B) = \frac{2 \sum_{i \in I} \rho_A(i)\ \rho_B(i)}{\left(\sum_{i \in I} \rho_A(i)^2\right) + \left(\sum_{i \in I} \rho_B(i)^2\right)} \qquad \textbf{Gl. 65}$$

Auch hier ergibt sich wieder ein Wert von 1 für identische Verteilungen sowie von –1 für solche, die sich nur im Vorzeichen unterscheiden. Allerdings erhält man nun für zwei Verteilungen, die sich nur um einen konstanten Faktor $n$ unterscheiden, ein Ähnlichkeitsmaß von $2\,n\,/\,(1 + n^2)$. Somit wird also Form *und* Größenordnung der Verteilungen gleichermaßen beurteilt. Für die $\rho_A$ bzw. $\rho_B$ in Gl. 64 bzw. Gl. 65 können formal jeweils die nach Gl. 13 erhaltenen Paarverteilungsfunktionen wie auch die nach Gl. 11 berechneten Paarpotentiale eingesetzt werden.

Gl. 64 wurde nun angewendet, um die Ähnlichkeit zweier Potentiale zu untersuchen, bei denen der Typ von Ligand- und Proteinatom jeweils vertauscht ist. Das Ergebnis ist in Abb. 9

in farbcodierter Form für alle möglichen Atom-Atom-Paare gezeigt; rote Farben stehen dabei für (nahezu) identische Paarpotentiale, (hell-)blaue für unähnliche. Im Falle dunkelblauer Flächen konnte kein Wert ermittelt werden, da mindestens eines der Potentiale aufgrund fehlender Kontakte in der zur Ableitung verwendeten Datenbank nicht ermittelt werden konnte. Bis auf den oberen linken Bereich der Abb. 9 mit Ähnlichkeitswerten für Kohlenstoff-Kohlenstoff-Paarpotentiale um 1 überwiegen in den meisten anderen Fällen jedoch deutlich davon abweichende Werte (für 66 % aller berechenbaren Fälle liegt das Ähnlichkeitsmaß unter 0.9). Dieses Ergebnis wird auch durch visuellen Vergleich jeweils zweier Potentiale mit ausgetauschten Ligand- und Proteinatomtypen erhalten. Aus diesem Grund wird darauf verzichtet, die Paarverteilungsfunktionen für Atompaare mit den Typen $t_1$ - $t_2$ mit solchen zu vereinigen, die die Typen $t_2$ - $t_1$ besitzen. Eine Erklärung für diese Unterschiede resultiert als Eigenheit der verwendeten Daten: nur im Fall der Betrachtung eines isolierten Atompaares oder eines Atompaares, das in eine homogene, nicht-strukturierte (molekulare) Umgebung eingebettet ist, ergeben sich bzgl. der Atomtypen des Paares symmetrische Verhältnisse. Bei dem hier verfolgten Weg der Ermittlung der Verteilungsfunktionen aus strukturierten, dicht gepackten Protein-Ligand-Komplexen zeigen sich dagegen klare Unterschiede zwischen $t_1$ - $t_2$- und $t_2$ - $t_1$-Verteilungen, die auf eine Einbettung der jeweiligen Atome in eine unterschiedliche molekulare Umgebung zurückzuführen sind. So kommen z. B. Atome des Typs O.2 im Protein nur in Peptidbindungen und in den Amidgruppen von Asn und Gln vor, in Ligandmolekülen dagegen in einer Vielzahl funktioneller Gruppen. Insbesondere für Atomtypen S.3, P.3, F, Cl, Br und Metall fallen deutliche Unähnlichkeiten zwischen den verglichenen $t_1$ - $t_2$ sowie $t_2$ - $t_1$ Potentialen auf, was sicherlich auch mit der geringen Anzahl von Beobachtungen für Verteilungen mit diesen Typen und der daraus resultierenden geringen statistischen Signifikanz erklärt werden kann. Daß dieses aber nicht der alleinige Grund ist, zeigt sich z. B. für die Fälle O.co2 – O.2 / O.2 – O.co2 (13739 bzw. 1984 Beobachtungen) oder O.3 – C.2 / C.2 – O.3 (31537 bzw. 5120 Beobachtungen) mit jeweils ausreichender Anzahl von Atom-Atom-Kontakten und dennoch stark unterschiedlichen Potentialen. So beträgt der nach Gl. 65 berechnete Ähnlichkeitswert im ersteren Fall 0.42, im letzteren 0.45.

Analog wurde für die Paarpotentiale der Einfluß des in Gl. 13 benötigten Kugelschalenvolumens untersucht, indem anstelle des jeweiligen Gesamtvolumens nur noch der durch benachbarte Ligandatome nicht besetzte Teil gemäß Gl. 30 verwendet wurde. Hierbei beträgt der Volumeninhalt *eines* Kugelschalenelementes maximal 2 % des Volumens der gesamten Kugelschale (für $\theta = \pi / 2$ entsprechend $\sin\theta = 1$). In mehr als 99% der verglichenen Fälle beträgt das nach Gl. 65 berechnete Ähnlichkeitsmaß allerdings mehr als 0.9, so daß die Vo-

lumenkorrektur hier keinen Einfluß auf die Form der Potentiale zeigt. Dies steht auf den ersten Blick im Gegensatz zu Ergebnissen von Muegge und Martin (Muegge & Martin, 1999), die die Notwendigkeit einer Volumenkorrektur betonen. Sie verwenden allerdings im Gegensatz zu dem hier vorgestellten Ansatz keinerlei Referenzzustand sowie einen maximalen Abstand zwischen zwei Atomen bei der Potentialableitung von 12 Å, wohingegen die hier erhaltenen Potentiale nur für Abstände bis 6 Å ermittelt wurden.

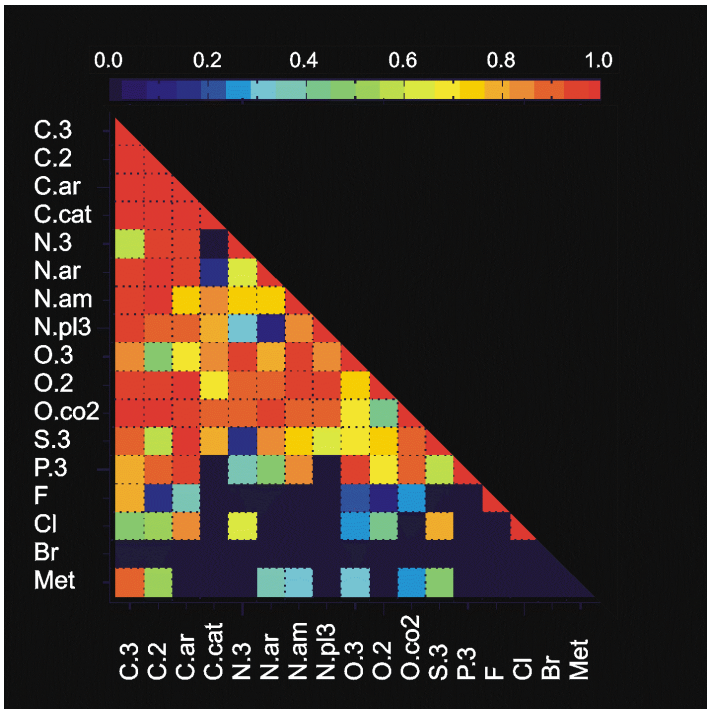
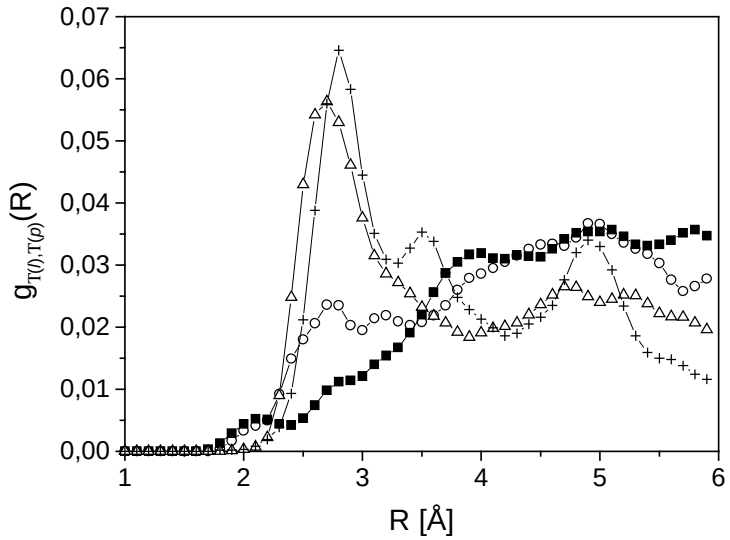
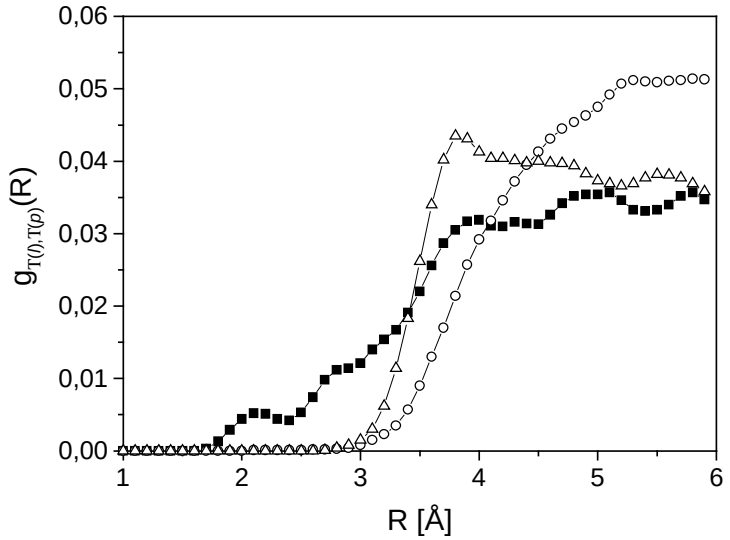



**Abb. 9:** Nach Gl. 65 berechnete Ähnlichkeiten von Paarpotentialen, bei denen der Atomtyp von Ligand und Protein jeweils vertauscht ist. Das Ähnlichkeitsmaß ist farbcodiert angegeben; rote Farbtöne entsprechen Ähnlichkeitswerten um 1, d.h. identischen Paarpotentialen, blaue Farbtöne deuten auf unähnliche Potentiale hin (s. a. Farbskala am oberen Rand). Die dunkelblauen Quadrate stehen für Fälle, in denen mindestens eines der verglichenen Potentiale aufgrund mangelnder Informationen im Datensatz nicht definiert ist.

Die erhaltenen Paarverteilungsfunktionen können in zwei Hauptklassen unterteilt werden: die erste umfaßt Wechselwirkungen zwischen polaren und geladenen Atomen und zeigt deutliche Maxima zwischen 2.5 Å und 3.0 Å. Sie sind auf Wasserstoff- und Salzbrückenbindungen zwischen den jeweiligen Atompaaren zurückzuführen (Abb. 10 a). Die zweite Klasse beinhaltet nichtpolare Wechselwirkungen und zeigt i.a. breitere Verteilungsfunktionen mit weniger stark ausgeprägten Maxima. Für sie ergeben sich höhere Wahrscheinlichkeiten verglichen mit dem Referenzzustand für Atom-Atom-Abstände > 3.5 Å (Abb. 10 b).




**a)**




**b)**

**Abb. 10:** Paarverteilungsfunktionen polarer und geladener (a) (O.co2-N.pl3 (+), O.3-O.co2 (Δ), O.3-O.3 (○)) bzw. nichtpolarer (b) (C.ar-C.ar (Δ), C.3-C.3 (○)) Atom-Atom-Wechselwirkungen, wie sie unter Verwendung von Gl. 13 bzw. Gl. 14 aus 1376 experimentell bestimmten Kristallstrukturen von Protein-Ligand-Komplexen ermittelt wurden. Die erste Atomtypbezeichnung bezieht sich dabei auf das betrachtete Ligandatom, die zweite auf das Proteinatom. Der nach Gl. 24 als gemittelte Verteilung über alle möglichen Paare berechnete Referenzzustand ist in beiden Fällen zusätzlich angegeben (■).

Zusätzlich zu dieser ersten Unterteilung lassen sich jedoch für Verteilungen innerhalb einer Gruppe weitere Unterschiede feststellen. Für die in Abb. 10 a gezeigten Paarverteilungsfunk-

tionen nimmt die Breite der auf die nächsten Nachbarn zurückgehenden Maxima in der Reihe O.co2 - N.pl3, O.3 - O.co2 und O.3 - O.3 zu, wohingegen ihre Höhe in umgekehrter Richtung zunimmt. Die aufgeführten Kombinationen von Atomtypen sind dabei charakteristische Vertreter für Salzbrücken, ladungsunterstützte Wechselwirkungen und „normale“ Wasserstoffbrückenbindungen (Davis & Teague, 1999). Zusätzlich zu diesen ersten Maxima ergibt sich für die O.co2 – N.pl3-Verteilung ein zweites Maximum bei 3.6 Å, das auf sekundäre Wechselwirkungen zwischen den nicht unmittelbar gegenüber stehenden Atomen zurückgeht (s.a. Abb. 3 a, S. 13). In allen drei Verteilungen schließlich ist ein weiteres Maximum zwischen 4.8 und 5.0 Å zu beobachten, das die zweite Koordinationssphäre des jeweiligen Wechselwirkungspartners um das betrachtete Zentralatom widerspiegelt. Das Auftreten dieser Art von Maxima ist charakteristisch für Systeme im kondensierten Zustand (Ben-Naim, 1992). Für die in Abb. 10 b gezeigten Paarverteilungsfunktionen nichtpolarer Atome tritt für den C.ar – C.ar-Fall ein deutliches Maximum bei 3.7 Å auf, das die π-π-Wechselwirkungen aromatischer Systeme widerspiegelt (Hunter & Sanders, 1990; Hunter *et al.*, 1991). Die C.3 – C.3-Verteilungsfunktion dagegen weist einen über einen weiten Distanzbereich verlaufenden Anstieg auf, ohne ein charakteristisches Maximum zu zeigen.

Die nach Gl. 11 aus diesen Paarverteilungsfunktionen und dem angegebenen Referenzzustand berechneten statistischen Präferenzen sind in Abb. 11 gezeigt. Beachtet man die logarithmischen Einheiten auf der Ordinate, so ergibt sich beim Vergleich der Potentialminima von O.3 – O.3 bzw. O.3 – O.co2 in Abb. 11 a bei einem Abstand um 2.5 Å, daß erstere Wechselwirkung um das Zweieinhalbfache weniger günstig sind als letztere. Auf den ersten Blick überraschend ist dagegen, daß der Unterschied zwischen O.3 – O.co2 und O.co2 – N.pl3 deutlich geringer ausfällt und letztere Wechselwirkung sogar weniger günstig ist als erstere. Dies scheint zunächst der gängigen Meinung zu widersprechen, daß Salzbrücken zwischen zwei geladenen Partnern mehr zur Stabilisierung von Protein-Ligand-Komplexen beitragen als lediglich ladungsunterstützte Wasserstoffbrücken (Davis & Teague, 1999; Hossain & Schneider, 1999). Allerdings treten gemäß der Definition der Atomtypen in Tab. 2 (S. 65) N.pl3-Typen nur in Amidino- und Guanidinogruppen auf. In einer idealen bidentaten geometrischen Anordnung einer Carboxylat- und einer Amidino- (oder Guanidino-) Gruppe (Abb. 3, S. 13) ergeben sich daher immer *zwei* O.co2 – N.pl3-Wechselwirkungen im Gegensatz zu *einer* O.3 – O.co2-Wechselwirkung in einer ideal orientierten ladungsunterstützten Wasserstoffbrücke. Beschränkt man die Betrachtung von Wechselwirkungen nur auf die jeweils nächsten Nachbarn, so trägt eine bidentate Salzbrücke daher etwa zweimal soviel zur Stabilisierung eines Protein-Ligand-Komplexes bei wie eine Wechselwirkung zwischen einem pola-

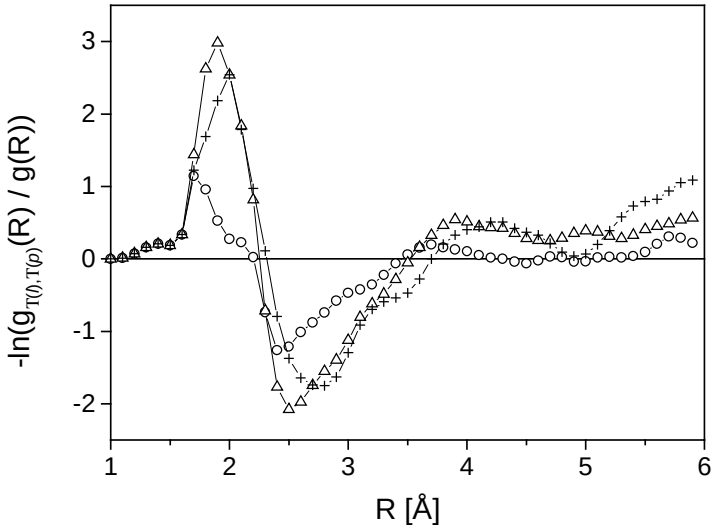
ren und einem geladenen Atom. Zusätzlich sei auf die Schulter in der statistischen Präferenz von O.co2 – N.pl3 bei 3.5 Å hingewiesen, die auf die oben erwähnten sekundären Wechselwirkungen zwischen nicht unmittelbar gegenüberstehenden Atomen in diesen bidentaten Salzbrücken zurückzuführen ist.

Bei den in Abb. 11 b gezeigten statistischen Präferenzen für ausgewählte nichtpolare Wechselwirkungen ergibt sich für C.ar – C.ar-Kontakte eine erhöhte Präferenz um 3.7 Å in Übereinstimmung mit beschriebenen aromatischen-aromatischen Wechselwirkungen (Hunter & Sanders, 1990; Hunter *et al.*, 1991). Die C.3 – C.3-Präferenz ist dagegen über den gesamten Bereich ab 4 Å bis zum Abstandsmaximum bei 6 Å zwar kleiner als Null (und die Wechselwirkung damit günstig), dabei aber wenig strukturiert. Dies ist in Übereinstimmung mit der Tatsache, daß diese Art von Wechselwirkungen nahezu keinen geometrischen Beschränkungen unterliegt.

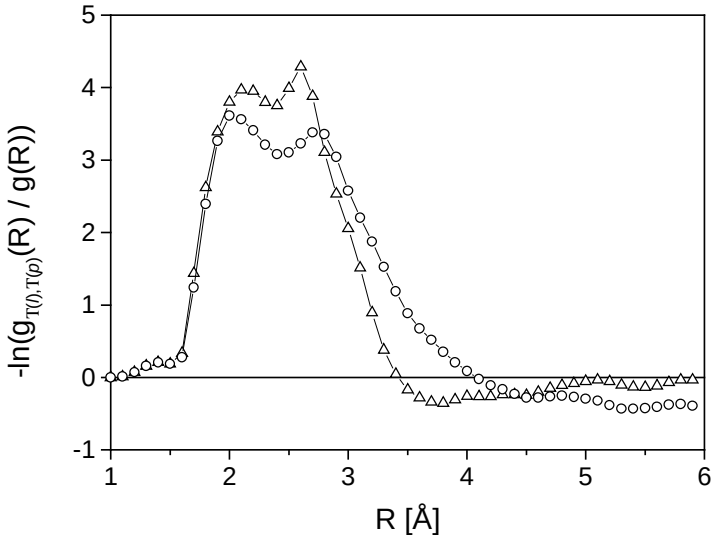
Beim Vergleich der Tiefe der jeweiligen Potentialminima muß zudem beachtet werden, daß bei kurzen Distanzen (< 3 Å) im Mittel nur ein direkter Wechselwirkungspartner involviert ist. Mit zunehmendem Abstand werden jedoch zusätzliche (nächste) Nachbarn einbezogen; ihre Anzahl ist dabei proportional zur dritten Potenz des Abstandes. Bei der Beurteilung des Beitrages eines spezifischen Kontaktes zur Stabilität eines Protein-Ligand-Komplexes muß daher sowohl dessen individuelle Stärke als auch die Häufigkeit seines Auftretens berücksichtigt werden.

Verglichen mit van der Waals-Wechselwirkungen zwischen zwei Atomen eines isolierten Paares zeigen die erhaltenen statistischen Präferenzen zwei charakteristische Unterschiede. Zum einen resultieren die zusätzlichen Minima bei größeren Abständen aus (dichten) Packungsanordnungen der Atome in der betrachteten molekularen Umgebung und spiegeln deren Organisation in Form von Koordinationsschalen höherer Ordnung wider. Zum anderen fallen die Werte der erhaltenen Präferenzen für Abstände < 2.0 Å wieder ab und zeigen bei 1 Å jeweils ein Minimum. Der Grund hierfür liegt zum einen darin, daß bei diesen Distanzen keine Atom-Atom-Kontakte mehr in den experimentell bestimmten Strukturen beobachtet werden können. Zum anderen kommt hier die nach Gl. 26 angewendete Dämpfung „lokaler Unsicherheit“ zum Tragen: in den Abstandsbereichen, in denen kaum oder keine Informationen für die spezifischen Paarverteilungen *und* den Referenzzustand vorliegen, bewirkt sie Potentialwerte um Null. Für die Bewertung von Protein-Ligand-Komplexen spielt dieser letzte Unterschied keine Rolle, denn sowohl in Kristallstrukturen als auch in von Dockingprogrammen wie FlexX, DOCK oder GOLD erzeugten Molekülanordnungen treten keine Fälle auf, bei

denen zwei Atome sich derart überlappen. Zudem ist es möglich, einen künstlichen Abtoßungsterm für diesen Abstandsbereich einzuführen (s.a.Kap. 4.8.1).




**a)**




**b)**

**Abb. 11:** Statistische Präferenzen für polare und geladene (a) (O.co2-N.pl3 (+), O.3-O.co2 (Δ), O.3-O.3 (○)) bzw. nichtpolare (b) (C.ar-C.ar (Δ), C.3-C.3 (○)) Paarwechselwirkungen als Funktion des Abstandes zwischen den betrachteten Atomen, berechnet nach Gl. 11. Die erste Atomtypbezeichnung bezieht sich dabei auf das betrachtete Ligandatom, die zweite auf das Proteinatom.

### 5.1.4 Abhängigkeit der Paarpotentiale von Qualität, Umfang und Zusammensetzung des zu ihrer Ableitung verwendeten Datensatzes

Die im vorherigen Kapitel beschriebenen statistischen Paarpräferenzen wurden aus 1376 kristallographisch bestimmten Protein-Ligand-Komplexen abgeleitet, die alle eine Auflösung ≤ 2.5 Å aufwiesen. Entfernt wurden aus diesem Datensatz diejenigen Komplexe, die für die Validierung der erhaltenen Bewertungsfunktion bei der Erkennung nativ-ähnlicher Protein-Ligand-Anordnungen verwendet wurden. Hier nun werden die so erhaltenen Paarpotentiale unter Anwendung des in Gl. 65 beschriebenen Ähnlichkeitsmaßes mit solchen verglichen, die auf in ihrer Zusammensetzung veränderten Datensätzen beruhen. Damit soll der Einfluß von Qualität und Umfang der Originaldaten auf die Paarpotentiale untersucht werden.

Zunächst wurden für die Ableitung der Präferenzen nur noch 700 Protein-Ligand-Komplexe mit einer Auflösung ≤ 2.0 Å verwendet. Für die damit erhaltenen Potentiale wurde unter Verwendung von Gl. 65 die Ähnlichkeit in bezug auf die ursprünglichen, unter Verwendung von 1376 Komplexen bis zu 2.5 Å Auflösung berechneten Potentiale ermittelt und in Abb. 12 farbcodiert dargestellt. Rote Farbtöne deuten dabei auf ein Ähnlichkeitsmaß von 1 hin, d.h. in diesem Fall sind die verglichenen Paarpräferenzen hinsichtlich Form und Größenordnung (nahezu) identisch. Die dunkelblauen Flächen stehen für Fälle, in denen mindestens eines der verglichenen Potentiale aufgrund mangelnder Information in den Ursprungsdaten nicht definiert ist. Insbesondere für die Paarpotentiale, die gemäß Tab. 16 auf mehr als 500 Atom-Atom-Kontakten beruhen, überwiegen die roten Farbtöne, d.h. für ausreichend populierte Verteilungen und davon abgeleiteten Präferenzen ist kein wesentlicher Einfluß der Auflösung der zugrundeliegenden Daten festzustellen. Unterschiede in den Potentialen sind dagegen v.a. bei solchen zu bemerken, bei denen mindestens ein Atomtyp S.3, F, Cl, Br oder Metall ist. Im Zusammenhang mit Tab. 16 ergibt sich jedoch, daß diese Unterschiede hauptsächlich durch das Wegfallen einzelner Kontakte in nicht statistisch signifikanten Bereichen der zugrundeliegenden Verteilungen (d.h. *quantitativ*) bedingt sind und weniger durch *qualitativ* andere Informationen. Dies wird auch in Abb. 13 deutlich: beim Vergleich der O.3-O.co2-Paarpotentiale, die unter Verwendung von Protein-Ligand-Komplexen bis zu 2.0 Å bzw. bis zu 2.5 Å Auflösung abgeleitet wurden, kommt es zu keinen merklichen Veränderungen der Lage der einzelnen Potentialminima in bezug auf den Abstand zwischen beiden Atomen. Allerdings ist zu bemerken, daß beim Übergang zu besser aufgelösten Kristallstrukturen die Extremalpunkte stärker ausgeprägt werden, d.h., daß der Betrag der Funktionswerte an diesen Positionen größer wird und daß die Breite der Extrema abnimmt. Dieses gilt insbesondere für das erste Minimum bei etwa 2.5 Å.

Ein ähnliches Auftreten von „konformativen Attraktoren“ (Walther & Cohen, 1999) wurde auch bei der Untersuchung von Verteilungsfunktionen der Proteinhauptketten-Torsionswinkel $\varphi$ und $\psi$ gefunden. Auch hier wird in den Ramachandran-Auftragungen lediglich eine „Konvergenz“ der Datenpunkte zu einigen wenigen ausgezeichneten $\varphi$/$\psi$-Kombinationen festgestellt, qualitativ andere Ergebnisse werden allerdings nicht erhalten. Aus diesem Grund ist die Einbeziehung von Komplexen mit geringerer Auflösung durchaus gerechtfertigt, liefern sie doch wertvolle Zusatzinformationen für schlecht populierte Verteilungen und ermöglichen so die Verwendung einer erhöhten Anzahl von Atomtypen, die ein breiteres und differenzierteres Spektrum von Wechselwirkungseigenschaften beschreiben können. Allerdings zeigen die Veränderungen in den Potentialverläufen in Abhängigkeit von der Qualität der zugrundegelegten Protein-Ligand-Komplexe, daß durchaus eine Abhängigkeit von der Auflösung – wenn auch nicht qualitativ, so doch quantitativ – bei den hier verwendeten Grenzen besteht.

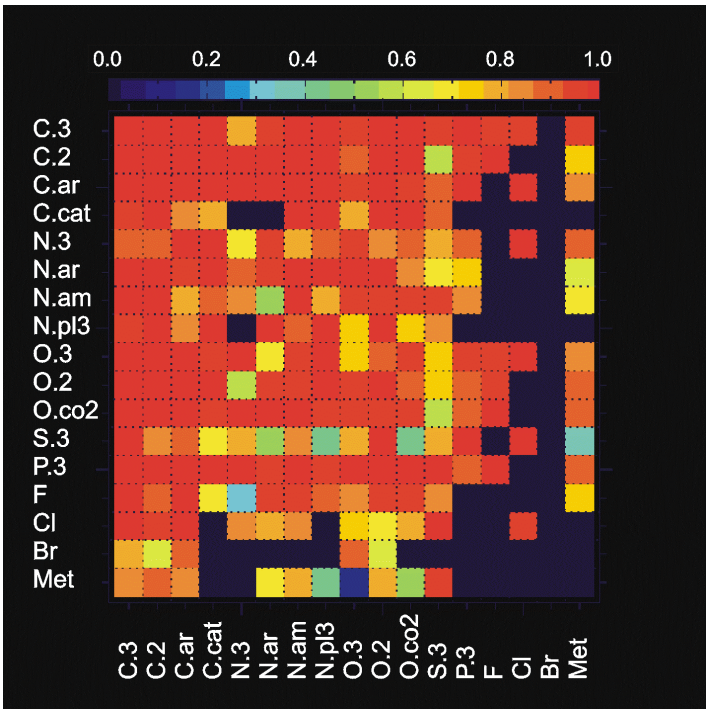



**Abb. 12:** Nach Gl. 65 berechnete Ähnlichkeiten von Paarpotentialen, die von 1376 Protein-Ligand-Komplexen mit einer Auflösung ≤ 2.5 Å bzw. von 700 Komplexen mit einer Auflösung ≤ 2.0 Å abgeleitet wurden. Typen für Ligandatome stehen in Zeilen, solche für Proteinatome in Spalten. Das Ähnlichkeitsmaß ist farbcodiert angegeben; rote Farbtöne entsprechen Ähnlichkeitswerten um 1, d.h. identischen Paarpotentialen, blaue Farbtöne deuten auf unähnliche Potentiale hin (s. a. Farbskala am oberen Rand). Die dunkelblauen Quadrate stehen für Fälle, in denen mindestens eines der verglichenen Potentiale aufgrund mangelnder Informationen im Datensatz nicht definiert ist.

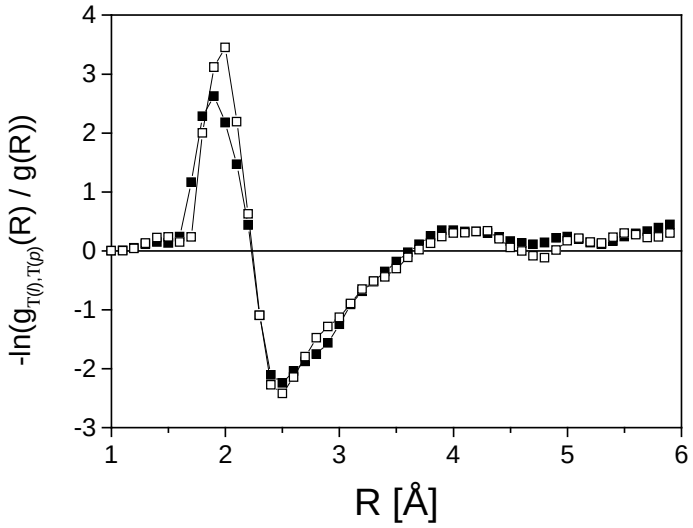
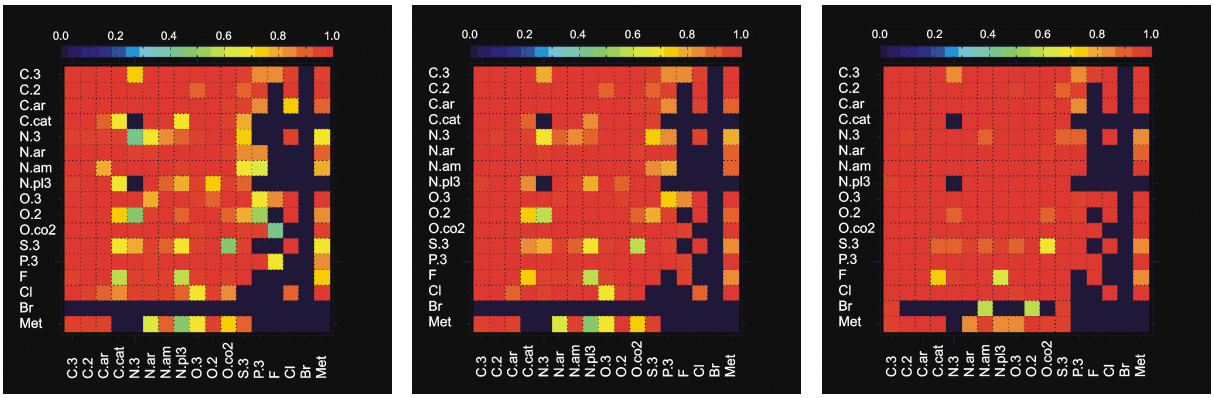



**Abb. 13:** Vergleich statistischer Präferenzen für O.3-O.co2-Paarwechselwirkungen als Funktion des Abstandes zwischen den betrachteten Atomen (berechnet nach Gl. 11), für die Protein-Ligand-Komplexe bis zu einer Auflösung von 2.0 Å (□) bzw. bis zu 2.5 Å (■) verwendet wurden.

In einer weiteren Untersuchung wurde aus den ursprünglich 1376 verwendeten Protein-Ligand-Komplexen zufällig ein Drittel, die Hälfte bzw. zwei Drittel ausgewählt und zur Ableitung der Potentiale verwendet. Die so erhaltenen Potentiale wurden wiederum gemäß Gl. 65 mit den unter Verwendung aller Komplexe abgeleiteten Präferenzen verglichen und die berechneten Ähnlichkeiten in Abb. 14 farbcodiert dargestellt. Die Bedeutung der Farben entspricht der im vorherigen Absatz beschriebenen. Wie aus Abb. 14 a ersichtlich ist, besteht schon bei Verwendung von nur einem Drittel der insgesamt bereitstehenden kristallographisch bestimmten Komplexe zur Potentialableitung nahezu Identität mit den in Kap. 5.1.3 beschriebenen Potentialen für solche Fälle, die gemäß Tab. 16 ausreichend häufig (> 500 Kontakte) populiert sind. Die in Abb. 14 a noch auftretenden Fälle mit einem Ähnlichkeitsmaß < 0.7 verringern sich beträchtlich schon bei der Verwendung von nur der Hälfte aller möglichen Kristallstrukturen (Abb. 14 b). Sie verschwinden nahezu vollständig (weniger als 9 % der verglichenen Fälle haben ein Ähnlichkeitsmaß < 0.9), wenn zwei Drittel aller Komplexe zur Potentialableitung verwendet werden (Abb. 14 c). Für die in Tab. 16 stark populierten Verteilungen und die daraus abgeleiteten Paarpotentiale ist also zu erwarten, daß auch durch Hinzunahme weiterer Protein-Ligand-Komplexe keine Veränderungen hinsichtlich ihrer Form mehr auftreten. Allerdings sind zusätzliche Daten für Atom-Atom-Kontakte in den schwach populierten Verteilungen notwendig, damit die dort *beobachteten* Atom-Atom-Verteilungen zu den *tatsächlichen* konvergieren (s.a. Kap. 6.2). Im Fall von Aminosäure-

Aminosäure-Kontaktpotentialen zur Bewertung von erzeugten Proteinstrukturen wurde von Miyazawa und Jernigan ähnliches berichtet: unter Verwendung des jeweils gleichen Formalismus wurden von den Autoren 1985 (Miyazawa & Jernigan, 1985) sowie 1996 (Miyazawa & Jernigan, 1996) entsprechende Kontaktpotentiale abgeleitet. Obwohl im letzteren Fall mehr als die 6-fache Anzahl an Kontakten insgesamt gegenüber der früheren Arbeit zur Verfügung stand, traten deutliche Veränderungen bei den Werten der Kontaktpotentiale nur für die Aminosäuren Trp, Met und Leu auf. Die Unterschiede im Fall der ersten beiden Aminosäuren lassen sich dabei auf zu geringe Auftrittshäufigkeiten zurückführen. Für den Leu-Fall konnte dagegen von den Autoren keine Begründung gefunden werden (Miyazawa & Jernigan, 1996).

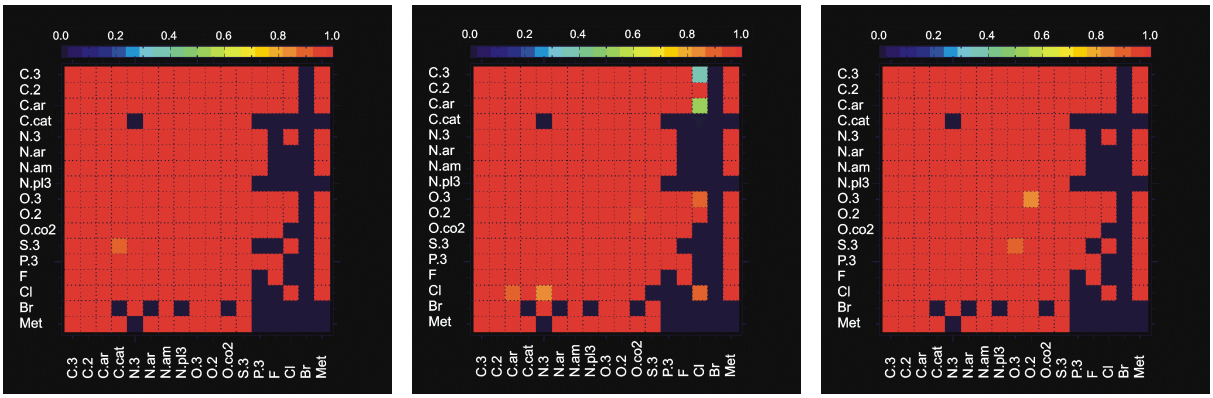



**a)** **b)** **c)**

**Abb. 14:** Nach Gl. 65 berechnete Ähnlichkeiten von Paarpotentialen, die von einem Drittel (a), der Hälfte (b) bzw. zwei Drittel (c) zufällig ausgewählten Protein-Ligand-Komplexen abgeleitet wurden, zu solchen, die unter Verwendung der insgesamt als Datenbasis dienenden 1376 Komplexe erhalten wurden. Typen für Ligandatome stehen in Zeilen, solche für Proteinatome in Spalten. Für die Bedeutung der Farbcodierung siehe Abb. 12, S. 112.

Um den Einfluß der Zusammensetzung des zur Ableitung verwendeten Datensatzes auf die Form der Paarpotentiale zu untersuchen, wurden in drei Fällen jeweils alle Komplexe mit mehr als 30 % Sequenzhomologie des Proteins zu einer gegebenen Proteinsequenz von diesem Datensatz ausgeschlossen und die Präferenzen erneut abgeleitet. Die Templatproteine waren Ribonuclease-1 (PDB-Code: 1rga), HIV-1-Protease (PDB-Code: 4phv) und Lysozym (PDB-Code: 129l); die Sequenzhomologie wurde mit FASTA (Pearson, 2000) bestimmt. Die Anzahlen so ausgeschlossener Protein-Ligand-Komplexe betrugen 30, 69 bzw. 137. Die erhaltenen Potentiale wurden wiederum gemäß Gl. 65 mit den unter Verwendung aller Komplexe abgeleiteten Präferenzen verglichen und die berechneten Ähnlichkeiten in Abb. 15 farbcodiert dargestellt. Die Bedeutung der Farben entspricht der im ersten Absatz beschriebenen. In allen drei Fällen sind nahezu alle Paarpräferenzen, die von den reduzierten Datensätzen ab-

geleitet wurden, mit den ursprünglichen, auf dem vollen Datensatz beruhenden, identisch. Dieses kann als ein Hinweis auf die Allgemeingültigkeit der Potentiale gewertet werden. So bestimmen nicht etwa spezielle, in einzelnen Proteinfamilien auftretende Wechselwirkungsmuster die Form der statistischen Präferenzen; ein allein auf einem „Lerneffekt“ beruhender Erfolg bei der Vorhersage von Struktur und Affinität von Protein-Ligand-Komplexen kann daher ausgeschlossen werden. Bei der Untersuchung des Einflusses der Datenbankzusammensetzung auf die Vorhersagefähigkeit von distanzabhängigen Paarpotentialen bezüglich Proteinstrukturen fanden Furuichi und Koehl (Furuichi & Koehl, 1998) dagegen, daß Potentiale, die aus einer Datenbank von ausschließlich α-Proteinen abgeleitet wurden, die beste Eignung zur Erkennung von α-Proteinen besitzen. Analoges gilt für Potentiale aus β-Proteinen. Dies verwundert indes nicht, bedingt doch die strikte Beschränkung auf einen Proteintyp, daß für jeweils charakteristische Wechselwirkungen des anderen jegliche Informationen in den erhaltenen Potentialen fehlen. Bei der hier vorgestellten Untersuchung verbleiben sogar nach Ausschluß von zu Lysozym sequenzhomologen Proteinen immer noch mehr als 1200 Protein-Ligand-Komplexe zur Ableitung der Paarpotentiale.




**a)** **b)** **c)**

**Abb. 15:** Nach Gl. 65 berechnete Ähnlichkeiten von Paarpotentialen, die unter Ausschluß von Komplexen mit > 30 % Sequenzhomologie zu Ribonuclease-1 (1rga) (a), HIV-1-Protease (4phv) (b) bzw. Lysozym (129l) (c) abgeleitet wurden, zu solchen, die unter Verwendung der insgesamt als Datenbasis dienenden 1376 Komplexe erhalten wurden. Typen für Ligandatome stehen in Zeilen, solche für Proteinatome in Spalten. Für die Bedeutung der Farbcodierung siehe Abb. 12, S. 112.

## *5.2 Von der Lösemittel-zugänglichen Oberfläche abhängige Einteilchen-Potentiale*

Für die im vorherigen Abschnitt beschriebenen distanzabhängigen Paarpotentiale ist zunächst von Sippl (Sippl, 1993) und später auch von Miyazawa und Jernigan (Miyazawa &

Jernigan, 1999) angemerkt worden, daß sie Lösemitteleffekte nicht ausreichend berücksichtigen. Um diese für Protein-Ligand-Wechselwirkungen bedeutsamen Effekte (s.a. Kap. 2.2.2) jedoch in die Bewertungsfunktion miteinzubeziehen, wurden zusätzlich von der Lösemittelzugänglichen Oberfläche abhängige Einteilchenpotentiale gemäß Gl. 31 aus kristallographisch bestimmten Rezeptor-Ligand-Komplexen ermittelt. Sie ergeben sich als negative Logarithmen des Verhältnisses zweier SAS-abhängiger, normierter Verteilungsfunktionen, die die Wahrscheinlichkeit angeben, mit der ein Ligand- oder Proteinatom eines gegebenen Typs im komplexierten bzw. freien Zustand mit einer gegebenen SAS angetroffen wird. Abweichend von den Paarpotentialen gibt es in diesem Fall also keinen gemeinsamen Referenzzustand, die erhaltenen statistischen Präferenzen sind demnach nicht voneinander abhängig.

Die in diesem Abschnitt beschriebenen Potentiale entsprechen denjenigen, die in Gl. 36 zur Vorhersage von Struktur und Bindungsaffinität von Protein-Ligand-Komplexen im weiteren Verlauf der Arbeit verwendet werden. Unter Verwendung des Kalibrierungsdatensatzes FlexX_DS1 hat sich die für sie getroffene Wahl der Intervallparameter und Parameter der Glättungsfunktion (s.a. Kap. 4.3.3) sowie die Art der Behandlung von Verteilungen geringer Datenanzahl gemäß Gl. 26 (unter Verwendung der Verteilungsfunktionen aus Gl. 33) mit $\chi = 0.25$ als optimal erwiesen. Der $\chi$-Parameter mußte hierbei wesentlich größer gewählt werden als der für die Paarpotentiale verwendete. Auch hier wird nur eine Auswahl der erhaltenen Potentiale vorgestellt.

Tab. 17 enthält die jeweils für die Ligand- bzw. Proteinatome gefundenen Beobachtungshäufigkeiten, mit denen ein Atom des angegeben Typs während der Ligandbindung an das Protein im Bindungsepitop *vergraben* wird (d.h. $SAS_x^{Kpl} < SAS_x^{Frei}$ unter Verwendung der Symbole aus Kap. 4.3.1). Die P.3, F, Cl, Br und Metall-Einträge auf der Proteinseite gehen dabei wieder auf die Verwendung von Kofaktoren als Teile des Proteins zurück. 22 von 34 Verteilungen besitzen insgesamt weniger als 500 Beobachtungen. Die insgesamt deutlich geringeren Anzahlen auf der Seite der Proteinatome verglichen mit denen der Ligandatome sind auf die unterschiedlichen Krümmungen der jeweiligen Molekülоberflächen zurückzuführen. Während eine Ligandoberfläche typischerweise konvex ist, sind Binde*taschen*bereiche eher konkav geformt. Zusammen mit der in Kap. 4.3.2 gegebenen Oberflächendefinition führt dies im Proteinfall zu einer Aufteilung der zur Verfügung stehenden Fläche auf weniger Atome als auf der Ligandseite.

**Tab. 17:** Für die Ableitung der von der Lösemittel-zugänglichen Oberfläche abhängigen Einteilchenpotentiale verwendete Anzahl der Beobachtungen, ermittelt aus 1376 kristallographisch bestimmten Protein-Ligand-Komplexen. Verteilungen mit weniger als 500 Kontakten sind grau unterlegt dargestellt.

| Atomtyp | Ligandatome | Proteinatome |
|---|---|---|
| C_3 | 11791 | 5582 |
| C_2 | 4836 | 2435 |
| C_ar | 6405 | 385 |
| N_3 | 149 | 55 |
| N_ar | 1740 | 39 |
| N_am | 1153 | 589 |
| N_pl3 | 32 | 29 |
| O_3 | 5217 | 124 |
| O_2 | 1429 | 1628 |
| O_co2 | 2953 | 192 |
| S_3 | 463 | 25 |
| P_3 | 4 | 7 |
| C_cat | 55 | 55 |
| F | 110 | 0 |
| Cl | 30 | 0 |
| Br | 1 | 0 |
| Met | 120 | 4 |

Abb. 16 zeigt jeweils für Ligandatome der Typen C.3 bzw. O.co2 die SAS-abhängigen Verteilungsfunktionen für den freien bzw. komplexgebundenen Zustand, wie sie unter Verwendung von Gl. 33 sowie des in Kap. 4.3.2 beschriebenen Verfahrens zur Berechnung der Lösemittel-zugänglichen Oberfläche aus 1376 Protein-Ligand-Komplexen bestimmt wurden. Während Ligandatome des Typs C.3 auch im freien, d.h. vollständig vom Protein getrennten Zustand eine erhöhte Wahrscheinlichkeit für fast vollständige Vergrabung aufweisen, tritt der Zustand der völligen Vergrabenheit (SAS = 0) im proteingebundenen Fall nahezu zehnmal so häufig auf. Für Atome des Typs O.co2 ist dagegen der Zustand weitestgehender Lösemittelexposition im nicht an das Protein gebundenen Fall am häufigsten. Im komplexgebundenen Zustand dagegen werden SAS-Werte um 15 bis 20 am häufigsten gefunden. Dieses Ergebnis ergibt sich als Konsequenz mehrerer Faktoren. Auf der einen Seite wird durch die Vergrabung

eines O.co2-Atoms dessen SAS reduziert. Andererseits werden in dem hier verfolgten Ansatz diejenigen Oberflächenanteile als *nicht* vergraben betrachtet, die sich an der Kontaktfläche zweier *polarer* (d.h. Sauerstoff- oder Stickstoff-) Atome befinden. Somit resultiert eine „teilweise“ Vergrabung als bester Kompromiß im proteingebundenen Zustand.

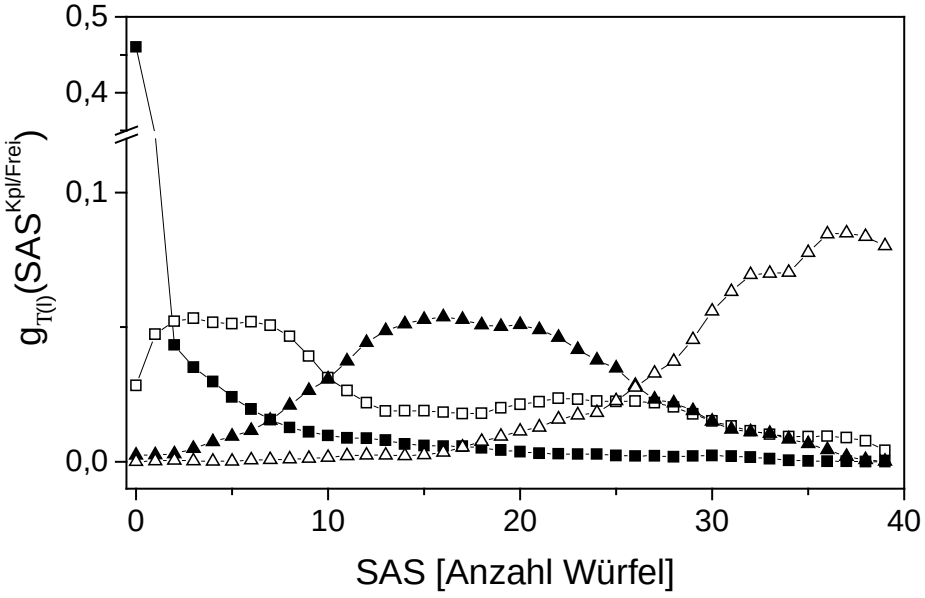



**Abb. 16:** SAS-abhängige Verteilungsfunktionen für Ligandatome des Typs C.3 (■, □) bzw. O.co2 (▲, Δ) jeweils im freien (□, Δ) und komplexgebundenen (■, ▲) Zustand, wie sie unter Verwendung von Gl. 33 und Alg. 1 aus 1376 experimentell bestimmten Kristallstrukturen von Protein-Ligand-Komplexen ermittelt wurden.

Abb. 17 zeigt die unter Anwendung von Gl. 31 erhaltenen statistischen Präferenzen für die Ligandatome des Typs C.3 und O.co2. Mit Ausnahme sehr kleiner nichtvergrabener Oberflächenanteile ist eine vollständige Vergrabung von C.3-Atomen für die Bildung von Protein-Ligand-Komplexen demnach stark günstig. Das Potential für O.co2 spiegelt die schon für die Verteilungsfunktionen in Abb. 16 diskutierten Einflüsse wider, die zu einer partiellen Vergrabung während der Proteinbindung führen. So zeigt sich, daß die teilweise Vergrabung (bzw. die Vergrabung in einer polaren Umgebung) günstiger ist als eine vollständige Vergrabung. Eine Exposition zum Lösemittel ist dagegen im proteingebundenen Zustand ungünstig. Letzteres kann damit erklärt werden, daß Wechselwirkungen von Carboxylatgruppen mit dem Protein für die Ligandbindung günstig sind. Somit treten im proteingebundenen Zustand vollständig solvatisierte Carboxylatgruppen weniger häufig auf als im Zustand vollständiger Separation eines Liganden von seinem Rezeptor.

Bemerkenswerterweise zeigt sich bei den auf diese Weise erhaltenen, SAS-abhängigen Einteilchenpotentialen für ein einzelnes Atom keine lineare Beziehung zwischen dem jeweiligen Beitrag zur Bindungsaffinität und der während der Komplexbildung vergrabenen Oberfläche, im Gegensatz zu den auf die Arbeiten von Chothia (Chothia, 1974), Chothia und Janin (Chothia & Janin, 1975), Eisenberg und McLachlan (Eisenberg & McLachlan, 1986) und Ooi *et al*. (Ooi *et al.*, 1987) zurückgehenden Methoden zur Berechnung von freien Solvatationsenthalpien basierend auf „atomaren Solvatationsparametern" (s. a. Kap. 2.2.2). Eine Abweichung von der Linearität insbesondere in den Bereichen vollständiger Vergrabung bzw. Solvatation wird dagegen auch bei den wissensbasierten Ansätzen von Jones *et al.* (Jones *et al.*, 1992) (Aminosäure-basierte „Solvatationspotentiale") sowie Delarue und Koehl (Delarue & Koehl, 1995) gefunden.

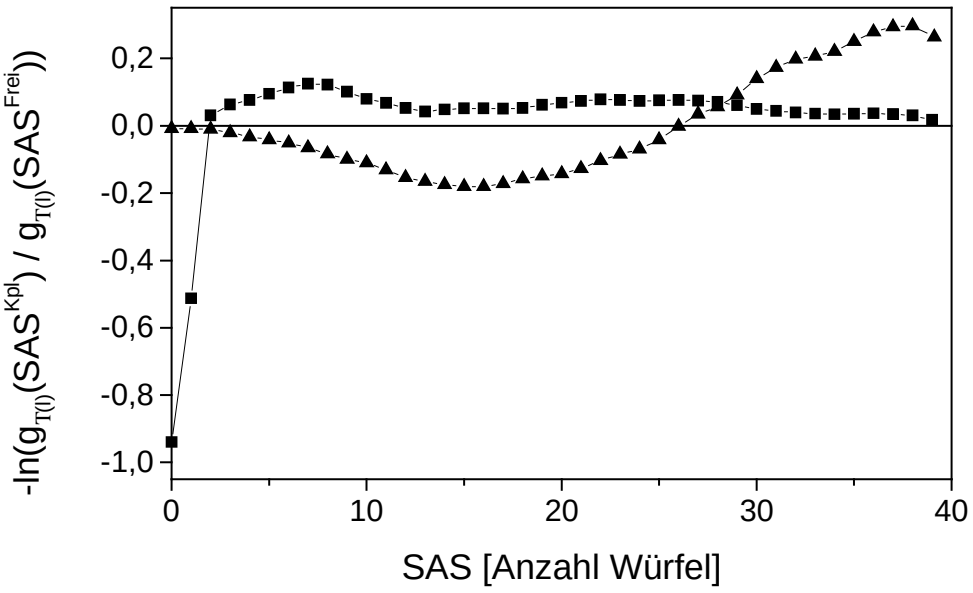



**Abb. 17:** SAS-abhängige Einteilchenpotentiale für die Ligandatome des Typs C.3 (■) und O.co2 (▲), wie sie nach Gl. 31 aus 1376 kristallographisch bestimmten Protein-Ligand-Komplexen abgeleitet wurden.

### *5.3 Kritische Betrachtung des gewählten Ansatzes*

Der hier gewählte wissensbasierte Ansatz wird im Vergleich zu Kraftfeld- und Regressions-basierten Ansätzen (s.a. Kap. 3.2.2 und 3.2.3) als umfassender betrachtet (Jernigan & Bahar, 1996; Moult, 1997), schließt er – bedingt durch die Ableitung von statistischen Präferenzen aus einer als Wissensbasis gewählten Datenbank von experimentell bestimmten Pro-

tein-Ligand-Komplexen – doch *implizit* auch solche Effekte mit ein, die noch nicht vollständig verstanden sind. Dies gilt insbesondere für auf Kooperativität beruhenden Effekten wie auch solchen, die hauptsächlich entropisch bedingt sind (Jones & Thornton, 1996; Sippl, 1995). Zusätzlich ergeben in der Datenbank weniger häufig als der Referenzzustand auftretende Ereignisse Präferenzen mit einem Wert größer Null, werden also als „ungünstig" erkannt und können so als „Strafterme" für nicht-native Protein-Ligand-Wechselwirkungen angesehen werden. Allerdings bedingt dieser *mean-field*-Charakter der erhaltenen Präferenzen auch, daß der Zusammenhang zwischen beobachteten Häufigkeiten struktureller Eigenschaften, den daraus abgeleiteten Präferenzen und den zugrundeliegenden physikalischen Ursachen keineswegs eindeutig ist (Vajda *et al.*, 1997) (s.a. Kap. 5.1.2).

Die häufig im Sinne des „inversen" Boltzmannschen Gesetzes erfolgende Rechtfertigung der wissensbasierten Ansätze und die Identifikation der damit erhaltenen Ergebnisse als „freie Energien" ist – obwohl auch in jüngerer Zeit noch kontrovers diskutiert (Sippl *et al*., 1996; Thomas & Dill, 1996) – nicht haltbar (Ben-Naim, 1997; Bürgi & Dunitz, 1988). Eine rein *statistische* Betrachtungsweise (s.a. Kap. 4.2.1 und (Samudrala & Moult, 1998)) umgeht dieses und hat darüber hinaus den Vorteil, daß prinzipiell jeder denkbare Referenzzustand (in anderen Worten: jede denkbare *a priori* Verteilung) in Gl. 11 bzw. Gl. 19 verwendet werden kann. Bei einer *thermodynamischen* Betrachtungsweise müßte ein gewählter Referenzzustand jedoch dahingehend gerechtfertigt werden, daß er alle möglichen Zustände des Systems umfaßt.

Die Anwendbarkeit der distanzabhängigen Paarpotentiale beruht auf den zu Beginn von Kap. 4.2.1 formulierten drei Annahmen. Eine visuelle Betrachtung der in Abb. 10 (S. 107) dargestellten Paarverteilungsfunktionen bzw. der in Abb. 11 (S. 110) daraus abgeleiteten Paarpräferenzen zeigt, daß diese für verschiedene Atom-Atom-Paarkombinationen deutlich verschieden voneinander sind, wodurch die dritte Annahme bestätigt ist. Inwieweit die damit berechenbaren Bewertungen für Protein-Ligand-Anordnungen tatsächlich auch diskriminierend sind, wird in den folgenden Kapiteln untersucht. Die erste Annahme, nach der Distanzen zwischen Ligand- und Proteinatomen paarweise unabhängig voneinander sind, gilt jedoch nur stark eingeschränkt. So ist z.B. bei der Wechselwirkung eines Ligandatoms des Typs O.2 einer Carbonylgruppe mit dem N.am-Atom eines Proteins allein aus geometrischen Gründen zu erwarten, daß sich das C.2-Atom der Carbonylgruppe in der Nähe des Protein-N.am-Atoms befindet. Demgemäß findet eine gewisse Mehrfachzählung von Wechselwirkungen statt. Auch die zweite Annahme, nach der die Verteilungen interatomarer Distanzen für verschie-

dene molekulare Umgebungen des betrachteten Atompaares ähnlich sind, ist sicher nur eingeschränkt gültig. So hängt z.B. die Stärke einer elektrostatischen Wechselwirkung (und analog der Abstand der betreffenden Atome) auch von dem Abschirmungseffekt des umgebenden Mediums ab (s.a. Kap. 2.2.1). Prinzipiell wäre eine Ableitung von Paarpotentialen in Abhängigkeit der jeweiligen molekularen Umgebung denkbar; in Anbetracht der dafür benötigten Anzahl experimenteller Beobachtungen für statistisch signifikante Verteilungsfunktionen wurde hierauf jedoch in der vorliegenden Arbeit verzichtet. Bei der Anwendung der Paarpotentiale ist zudem zu berücksichtigen, daß außer der gerade betrachteten Paarwechselwirkung zusätzliche Wechselwirkungen durch umliegende Atome ausgebildet werden. Inwieweit diese „zusammengesetzte“ Repräsentation den tatsächlichen Wechselwirkungen im molekularen Umfeld entspricht, wird in Kap. 4.7 untersucht.
Die Abhängigkeit der von den Paarpotentialen berücksichtigten Beiträge zu Protein-Ligand-Wechselwirkungen von der Wahl des Referenzzustandes sowie weiteren Parametern ist bereits in Kap. 4.2 erwähnt worden. Die hier verwendete Kombination eines kurzreichweitigen, distanzabhängigen Paarpotentials sowie eines SAS-abhängigen Einteilchenpotentials wurde im Rahmen der Erkennung korrekt gefalteter Proteinstrukturen bereits von Jones *et al.* (Jones *et al.*, 1992) beschrieben; allerdings benutzen die Autoren für ihr Einteilchenpotential einen auf die Arbeiten von Sippl (Sippl, 1993) zurückgehenden, über alle Atomtypen gemittelten Referenzzustand im Gegensatz zu dem hier gemäß Gl. 33 verwendeten sowie ein relatives Oberflächenmaß. Die hier verfolgte Zweiteilung der Beschreibung von Protein-Ligand-Wechselwirkungen ist auch in theoretischer Hinsicht von Bedeutung. So wird von Ben-Naim betont (Ben-Naim, 1997), daß die Annahme der Additivität von Beiträgen aus Paarpotentialen dann umso gerechtfertigter ist, wenn Lösemittel-induzierte Beiträge in diesen eher untergeordnet enthalten sind. Eine zusätzliche Beschreibung von Lösemittelbeiträgen durch Oberflächen-abhängige Terme (Eisenberg & McLachlan, 1986; Ooi *et al.*, 1987) ergänzt diese fehlende Information.

Die Verwendung von PDB-Daten bedingt, daß aufgrund der Datenlage statistische Präferenzen nur für Nichtwasserstoffatome (s.a. Kap. 4.4) sowie eine begrenzte Zahl von Atomtypen abgeleitet werden können. Kap. 5.1.1, 5.1.4 und 5.2 zeigen außerdem, daß für viele Atomtypen nur unzureichende Anzahlen von Beobachtungen zur Verfügung stehen und daß insbesondere für nur gering populierte Verteilungen die daraus abgeleiteten *beobachteten* mit den *tatsächlichen* Präferenzen noch nicht übereinstimmen. Eine denkbare Möglichkeit zur Umgehung dieser Probleme besteht in der Verwendung von aus der Kleinmoleküldatenbank

CSD (Allen *et al.*, 1991) gewonnenen Informationen. Zwei grundlegende Unterschiede sind hierbei allerdings zu bedenken: zum einen ergeben sich die so erhaltenen Verteilungen aus intermolekularen Wechselwirkungen in der Kristallpackung *eines* Moleküls, im Gegensatz zu den Wechselwirkungen zwischen Ligand und Protein für aus der PDB erhaltene Informationen. Zum anderen werden Protein-Ligand-Komplexe normalerweise aus wäßrigen Lösungen kristallisiert, wohingegen Kristalle niedermolekularer organischer Moleküle auch aus einer ganzen Reihe organischer Lösemittel erhalten werden. In letzterem Fall kann daher angenommen werden, daß der Einfluß des hydrophoben Effektes (s.a. Kap. 2.2.2) auf die aus der CSD gewonnenen Daten geringer ist als der auf die PDB-Daten. Dieser Unterschied wurde von Verdonk *et al.* (Verdonk *et al.*, 1999) während der Entwicklung des Programms SuperStar beschrieben. Die dabei aus der CSD gewonnenen Verteilungen mußten durch eine adäquate Skalierung an die aus der PDB angepaßt werden.

## *5.4 Bewertung nativ-ähnlicher Ligandenkonfigurationen*

In diesem Abschnitt wird die Eignung der mit Gl. 36 beschriebenen Bewertungsfunktion zur Auswahl nativ-ähnlicher Protein-Ligand-Anordnungen aus einer Menge gegebener Alternativen untersucht. Die Rezeptor-Ligand-Geometrien werden dafür durch die Docking-Programme FlexX (Rarey *et al.*, 1995; Rarey *et al.*, 1996a) und DOCK (Kuntz *et al.*, 1982; Makino & Kuntz, 1997; Meng *et al.*, 1992) erzeugt. Da für DOCK bislang keine Validierungsstudie für einen umfassenden Datensatz gemischter Protein-Ligand-Komplexe vorliegt, zudem gemäß Tab. 4 (S. 88) mehrere Alternativen zur Plazierung und Bewertung der Liganden existieren, wird im folgenden eine Evaluation dieses Programmes basierend auf 100 Protein-Ligand-Komplexen mit bekannter Kristallstruktur vorgestellt.

### 5.4.1 Evaluation des Programmes DOCK und Vergleich mit FlexX und GOLD

Tab. 18 enthält eine zusammenfassende Statistik für alle untersuchten Kombinationen von Ligandplazierungsalgorithmen und Bewertungsfunktionen für die Protein-Ligand-Anordnungen, wie sie in Tab. 4 (S. 88) aufgeführt sind. Als Maß für die Genauigkeit wird hierbei der *rmsd*-Wert (Gl. 37) der Positionsabweichungen der Nichtwasserstoffatome der gedockten Ligandgeometrien von denen des (minimierten) Liganden aus der Kristallstruktur verwendet, im Gegensatz zu den von Jones *et al.* (Jones *et al.*, 1997) eingeführten subjektiven Kategorien (*good, close, errors, wrong*). Zu beachten ist allerdings, daß der *rmsd*-Wert eine

gemittelte Größe ist, d.h., daß sowohl mittlere Abweichungen im gesamten Molekül als auch eine nahezu perfekte Geometrie in einem Teil verbunden mit großen Abweichungen in einem anderen Teil des Moleküls zu demselben Zahlenwert führen können. Außerdem werden in diesem Fall sämtliche Atome eines Liganden gleich behandelt, unabhängig davon, ob sie z.B. zu einem ins Lösemittel ragenden Teil des Moleküls gehören und daher als eher flexibel angenommen werden sollten. Weiterhin ist zu beachten, daß bei der Berechnung des *rmsd* Symmetrien in Moleküluntereinheiten nicht beachtet werden. Gemäß den in Kap. 4.6 angegebenen Kriterien werden solche Docking-Lösungen als „gut" betrachtet, deren *rmsd* kleiner als 2.0 Å ist. Dies ist in Übereinstimmung mit den Arbeiten von Kramer *et al.* (Kramer *et al.*, 1999) und Jones *et al.* (Jones *et al.*, 1997).

Als Kriterium für die Güte der Ligandplazierung wird der kleinste *rmsd*-Wert herangezogen, der ohne Beachtung einer Bewertung dieser Ligandplazierung innerhalb der erzeugten Lösungsmenge gefunden wird. Allerdings ist zu beachten, daß im Falle des flexiblen Dokkings unter Verwendung des inkrementellen Aufbaualgorithmus' von den auf einer Stufe erzeugten partiellen Lösungen nur ein vorher festgelegter Anteil in die nächste Stufe gelangt und so eine Einflußnahme der Bewertungsfunktion auf die resultierende Lösungsmenge vorliegt. Durch Verwendung des *uniform sampling*, d.h. der Erzeugung der gleichen Anzahl geometrischer Lösungen, die in diesem Fall von der verwendeten Bewertungsfunktion unabhängig sind, ist im Fall des rigiden Dockings *ohne* anschließende Minimierung diese Einflußnahme jedoch ausgeschlossen. Ein Eindruck von der Eignung der jeweiligen Bewertungsfunktion ergibt sich, wenn nur noch der *rmsd*-Wert für die bestbewertete Ligandanordnung herangezogen wird. Dies folgt auch der Notwendigkeit, sich im Rahmen von virtuellen Screening-Ansätzen auf die erfolgte Bewertung verlassen zu müssen, ohne eine durch die anfallende Datenmenge nicht durchführbare zusätzliche visuelle Kontrolle der Ergebnisse.

Die Verfahren zum rigiden Docking der Kristallstruktur des Liganden sind aus zwei Gründen mit untersucht worden. Zum einen kann das Docking flexibler Liganden auch erfolgen, indem vorher berechnete Ligandkonformationen anschließend *rigide* in die Bindetasche eingepaßt werden (Charifson *et al.*, 1999). Zum anderen liefern die dabei erhaltenen Ergebnisse eine Aussage über die Qualität des Ligandplazierungsalgorithmus' alleine, ohne Einflußnahme z.B. der für die Modellierung der Ligandenflexibilität verwendeten Parameter. Schließlich ist hiermit auch eine Aussage über die Qualität der Strukturaufbereitungen für das Docking-Verfahren möglich: sollte es nicht einmal gelingen, die Kristallkonformation des Liganden rigide in die Bindetasche zu docken, wären die verwendeten Eingabedaten zu überprüfen.

Tab. 18 zeigt, daß im Fall der Plazierung rigider Liganden ein mittlerer kleinster *rmsd*-Wert zwischen 0.9 und 1.3 Å erreicht wird; für das flexible Docking ist er dagegen mit 2.3 bis 3.1 Å etwa doppelt so groß. In allen Fällen werden jeweils minimale Abweichungen von den Kristallstrukturen gefunden, die im Bereich der experimentellen Ungenauigkeiten (bis 0.4 Å) liegen. Die für die Verfahren rig_cnt, rig_nrg und rig_chm (rigides Docking unter Verwendung der „Kontakt"-Funktion, der „Energie-Funktion" bzw. der „Chemischen Bewertungsfunktion") aufgeführten Ergebnisse sind identisch; dies resultiert aus dem oben angesprochenen *uniform sampling*. Werden die mit dem Plazierungsalgorithmus erhaltenen Ligandgeometrien mit der jeweiligen Bewertungsfunktion optimiert, ergeben sich für die rigiden Verfahren Verbesserungen in der Ligandplazierung (d.h. kleinere *rmsd*-Werte). Im Fall des flexiblen Dockings bleibt der mittlere kleinste *rmsd*-Wert dagegen gleich (flex_nrg / flex_nrg_min) oder nimmt sogar zu. Da hierbei nicht nur der vollständig in die Bindetasche eingepaßte Ligand, sondern auch das Ankerfragment bzw. nur partielle aufgebaute Liganden nach einem heuristischen Verfahren flexibel optimiert werden (Gschwend & Kuntz, 1996), ist eine monokausale Begründung für diese Beobachtung nicht möglich.

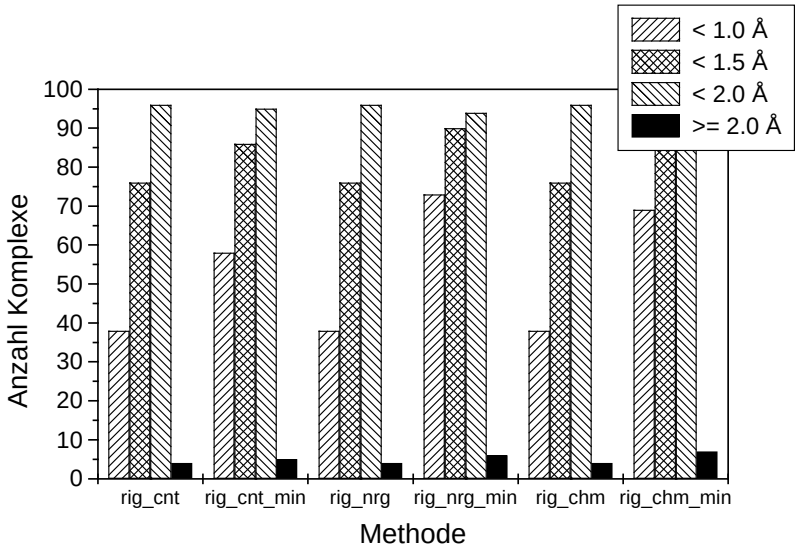
Wird der Mittelwert des *rmsd* derjenigen Protein-Ligand-Konfigurationen gebildet, die von den jeweiligen Bewertungsfunktionen als am günstigsten eingestuft werden, so betragen die Werte zwischen 4.0 und 6.5 Å. Im Zusammenhang mit dem jeweils kleinsten *rmsd*, der überhaupt erzeugt wurde, weist dieses ganz eindeutig auf die ungenügende Eignung der Bewertungsfunktionen hin, aus einem gegebenen Satz von Protein-Ligand-Geometrien die nativähnlichen herauszufinden. Wie im Falle des kleinsten *rmsd*-Wertes, erfolgt auch hier eine Verbesserung der Ligandplazierung durch nachfolgende Optimierung bei allen rigiden Verfahren, wohingegen für die flexiblen Methoden die mit Energieminimierung erhaltenen Geometrien nur bei der Verwendung der „Energie-Bewertung" (flex_nrg_min) besser werden. Sowohl im Falle des rigiden als auch flexiblen Dockings mit DOCK erweist sich der Einsatz der auf das AMBER-Kraftfeld (Weiner *et al.*, 1984) zurückgehenden „Energie-Bewertungsfunktion" in Kombination mit einer Geometrieoptimierung der Liganden in der Proteinbindetasche als beste Wahl, wird hiermit doch ein mittlerer *rmsd* für bestbewertete Ligandgeometrien von 4.0 bzw. 4.1 Å erreicht. Zusammen mit dem mittleren kleinsten *rmsd* von 2.3 Å für das flex_nrg_min-Verfahren werden somit zu FlexX vergleichbare Ergebnisse erzielt (2.2 Å bzw. 4.0 Å gemäß (Kramer *et al.*, 1999)).

**Tab. 18:** *rmsd*-Statistik der mit DOCK durchgeführten Docking-Experimente für einen Datensatz von 100 Protein-Ligand-Komplexen.

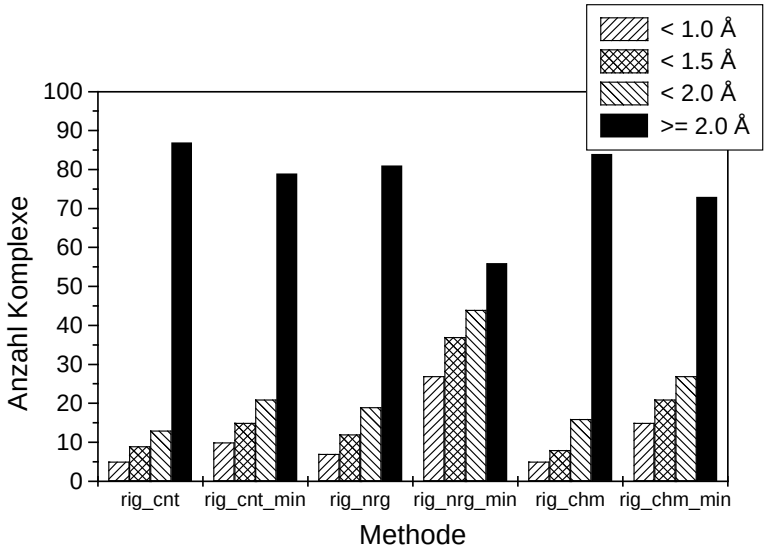
| **Methode**[a] | **Kleinster *rmsd*[b]** | | | | ***rmsd*[b] auf Rang 1** | | | |
|---|---|---|---|---|---|---|---|---|
| | **E**[c] | **σ**[d] | **Min.**[e] | **Max.**[f] | **E**[c] | **σ**[d] | **Min.**[e] | **Max.**[f] |
| **rig_cnt** | 1.3 | 0.8 | 0.3 | 5.9 | 6.1 | 3.9 | 0.4 | 18.4 |
| **rig_cnt_min** | 1.0 | 0.8 | 0.1 | 6.0 | 4.6 | 3.5 | 0.3 | 14.0 |
| **rig_nrg** | 1.3 | 0.8 | 0.3 | 5.9 | 5.6 | 4.1 | 0.3 | 21.8 |
| **rig_nrg_min** | 0.9 | 0.9 | 0.1 | 7.3 | 4.1 | 4.1 | 0.2 | 18.9 |
| **rig_chm** | 1.3 | 0.8 | 0.3 | 5.9 | 6.5 | 4.4 | 0.3 | 21.9 |
| **rig_chm_min** | 1.0 | 1.0 | 0.2 | 6.8 | 5.9 | 4.6 | 0.4 | 19.5 |
| **flex_cnt** | 2.3 | 1.6 | 0.4 | 9.2 | 5.7 | 3.5 | 0.6 | 21.8 |
| **flex_cnt_min** | 2.7 | 1.8 | 0.2 | 10.5 | 5.9 | 2.8 | 0.7 | 11.8 |
| **flex_nrg** | 2.3 | 1.7 | 0.3 | 9.2 | 4.8 | 3.3 | 0.5 | 21.7 |
| **flex_nrg_min** | 2.3 | 2.1 | 0.2 | 12.2 | 4.0 | 3.2 | 0.4 | 14.5 |
| **flex_chm** | 2.5 | 1.7 | 0.3 | 10.1 | 6.0 | 3.9 | 0.6 | 18.9 |
| **flex_chm_min** | 3.1 | 2.5 | 0.2 | 12.1 | 5.9 | 4.0 | 0.3 | 16.7 |

a) Die angegebenen Abkürzungen entsprechen den in Tab. 4 (S. 88) eingeführten. b) In Å. c) Mittelwert. d) Standardabweichung. e) Minimum. f) Maximum.

Abb. 18 und Abb. 19 stellen jeweils die Anzahl der Komplexe dar, für die bei Anwendung rigiden bzw. flexiblen Dockings ein *rmsd* < 1.0, < 1.5, < 2.0 sowie ≥ 2.0 Å als bester Wert überhaupt bzw. als Wert für die auf Rang 1 bewertete Protein-Ligand-Konfiguration gefunden wurde. Gemäß obigen Kriteriums werden danach bei rigidem Docking für alle Verfahren in etwa 95% der Fälle „gute“ Ligandanordnungen (*rmsd* < 2.0 Å) gefunden (Abb. 18 a). Allerdings wird nur in maximal 44 Fällen (Methode rig_nrg_min) eine „gute“ Lösung nach der Bewertung auch auf Rang 1 erhalten (Abb. 18 b).




**a)**

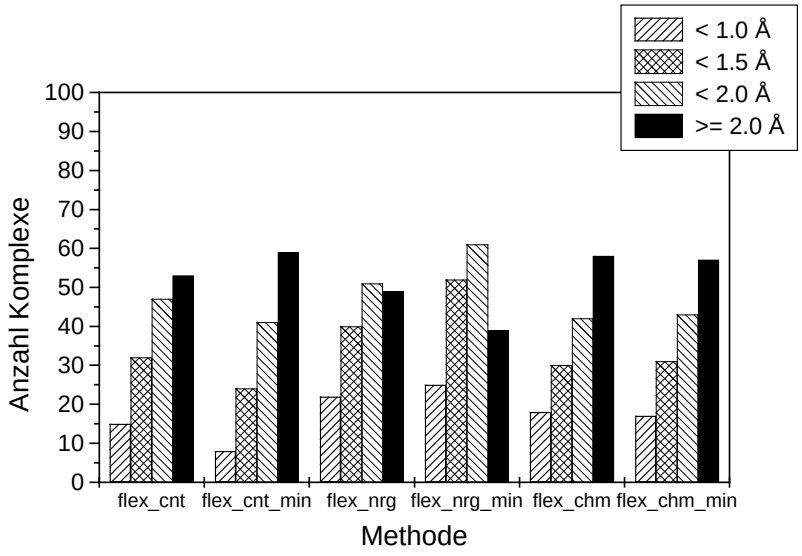



**b)**

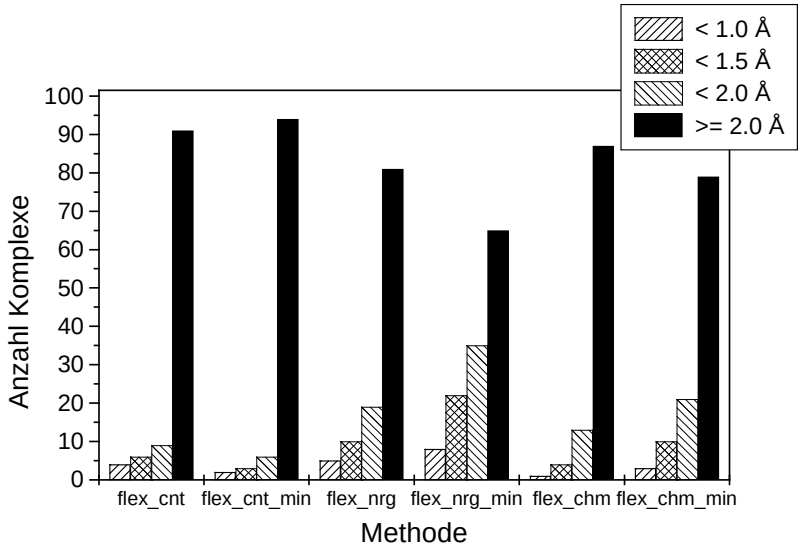
**Abb. 18:** Anzahl der Komplexe, für die bei Anwendung *rigiden* Dockings ein *rmsd* < 1.0, < 1.5, < 2.0 bzw. ≥ 2.0 Å als bester Wert überhaupt (a) bzw. als Wert für die auf Rang 1 bewertete Protein-Ligand-Konfiguration (b) gefunden wurde. Die Bezeichnungen der Methoden folgen den in Tab. 4 (S. 88) eingeführten.

Im Fall des flexiblen Dockings ergibt sich nur für die Methode flex_nrg_min, daß für 61 Fälle eine „gute“ Lösung unabhängig von ihrer Bewertung überhaupt erzeugt werden kann; für alle anderen Verfahren liegt die Anzahl zwischen 40 und 50 (Abb. 19 a). In 35 Fällen wird bei flex_nrg_min auch eine Lösung mit *rmsd* < 2.0 Å auf Rang 1 gefunden, d.h. in 57 % derjenigen Fälle, in denen überhaupt eine solche Lösung erzeugt wurde. Das zweitbeste Verfahren

zum flexiblen Docking ist in dieser Hinsicht flex_chm_min, bei der die „Chemische Bewertungsfunktion“ als Abwandlung der „Energie-Funktion“ verwendet wird. Hier ergibt sich in 21 Fällen eine „gute“ Lösung auf Rang 1, d.h. in 48% der 43 Fälle, in denen überhaupt eine „gute“ Lösung erzeugt wurde.




**a)**




**b)**

**Abb. 19:** Anzahl der Komplexe, für die bei Anwendung *flexiblen* Dockings ein *rmsd* < 1.0, < 1.5, < 2.0 bzw. ≥ 2.0 Å als bester Wert überhaupt (a) bzw. als Wert für die auf Rang 1 bewertete Protein-Ligand-Konfiguration (b) gefunden wurde. Die Bezeichnungen der Methoden folgen den in Tab. 4 (S. 88) eingeführten.

Abb. 20 zeigt die für alle 100 Protein-Ligand-Komplexe unter Verwendung des Verfahrens flex_nrg_min erhaltenen kleinsten *rmsd*-Werte bzw. diejenigen, die den bestbewerteten Protein-Ligand-Anordnungen entsprechen, als Funktion der Anzahl drehbarer Bindungen der jeweiligen Liganden. Bindungen zu terminalen Gruppen (etwa Methyl- oder Hydroxylgruppen) sowie Bindungen in Ringen werden dabei als nichtdrehbar angesehen. Die Anzahl drehbarer Bindungen gibt einen Hinweis auf die Komplexität des Docking-Problems, denn an ihnen werden die Liganden zunächst in Untereinheiten zerlegt, um im anschließenden inkrementellen Aufbauprozeß in der Bindetasche unter Berücksichtigung der konformativen Flexibilität wieder zusammengesetzt zu werden. Während für 86 Liganden mit ≤ 10 drehbaren Bindungen in 61 Fällen (71 %) eine Ligandkonfiguration mit einem *rmsd* < 2.0 Å erzeugt werden kann, gelingt dieses in keinem Fall für Liganden mit mehr als 10 drehbaren Bindungen. Für 66 Liganden mit ≤ 5 drehbaren Bindungen wird sogar in 79 % der Fälle eine „gute“ Ligandgeometrie generiert. Allerdings wird nur in 44 bzw. 41 % der Fälle von Liganden mit ≤ 5 bzw. ≤ 10 drehbaren Bindungen auch eine „gute“ Lösung als bestbewertete gefunden. Ein Zusammenhang zwischen der Qualität der erhaltenen Lösungen und der Anzahl von Wasserstoffbrückendonatoren bzw. -akzeptoren sowie der Anzahl nichtpolarer Atome in den Liganden konnte allerdings nicht festgestellt werden.

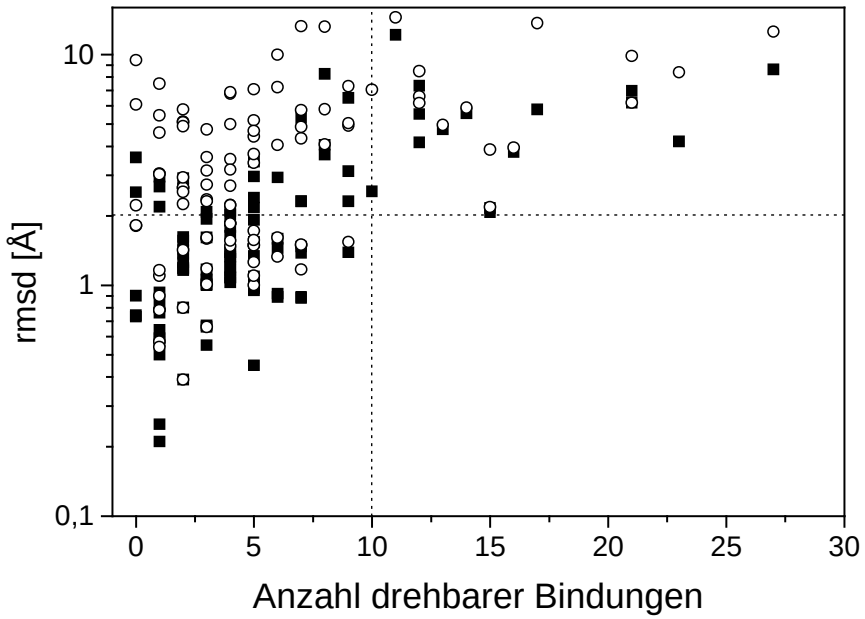



**Abb. 20:** Für die 100 mit dem Verfahren flex_nrg_min gedockten Protein-Ligand-Komplexe erhaltene kleinste *rmsd*-Werte (■) sowie *rmsd*-Werte der Ligandkonfiguration auf Rang 1 (○) als Funktion der Anzahl drehbarer Bindungen in dem jeweiligen Ligandmolekül.

In der folgenden Tab. 19 werden die in dieser Arbeit beim flexiblen Docking von 100 Protein-Ligand-Komplexen mit DOCK unter Verwendung der „Energiefunktion“ sowie der „Chemischen Bewertungsfunktion“ erhaltenen Ergebnisse verglichen mit Validierungsstudien für die Programme FlexX (Kramer *et al.*, 1999) und GOLD (Jones *et al.*, 1997). Die Qualität der Ergebnisse von FlexX wurde dabei an 200 Protein-Ligand-Komplexen ermittelt, im Fall von GOLD wurden 100 Komplexe verwendet. Da der GOLD-Datensatz eine Untermenge des FlexX-Datensatzes ist und die hier verwendeten Komplexe aus letzterem ausgewählt wurden, besteht für alle drei Fälle eine deutliche Überschneidung der eingesetzten Testdatensätze. Als Kriterium sollen hier erneut die Anteile von Komplexen herangezogen werden, für die ein *rmsd* < 1.0, < 1.5, < 2.0 bzw. ≥ 2.0 Å überhaupt bzw. auf dem ersten Rang gefunden werden konnte. Zusätzlich wird eine Studie von Knegtel *et al.* (Knegtel *et al.*, 1999) verwendet, bei der 32 Thrombininhibitoren mit DOCK ebenfalls flexibel unter Verwendung oben angeführter Bewertungsfunktionen in Thrombin gedockt wurden.

**Tab. 19:** Vergleich der Ergebnisse von Validierungsstudien für flexibles Docking mit den Programmen DOCK, FlexX und GOLD.

| **Methode** | **Anteil[a] Komplexe mit kleinstem *rmsd*** | | | | **Anteil[a] Komplexe mit *rmsd* auf Rang 1** | | | |
|---|---|---|---|---|---|---|---|---|
| | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** |
| **flex_nrg_min[b]** | 25 | 52 | 61 | 39 | 8 | 22 | 35 | 65 |
| **flex_chm_min[b]** | 17 | 31 | 43 | 57 | 3 | 10 | 21 | 79 |
| **DOCK energy[c]** | -[d] | -[d] | 22 | 78 | -[d] | -[d] | -[d] | -[d] |
| **DOCK chemical[c]** | -[d] | -[d] | 34 | 66 | -[d] | -[d] | -[d] | -[d] |
| **FlexX[c]** | -[d] | -[d] | 13 | 87 | -[d] | -[d] | -[d] | -[d] |
| **FlexX[e]** | 51 | 64 | 69 | 31 | 17 | 32 | 47 | 53 |
| **GOLD[f]** | -[d] | -[d] | -[d] | -[d] | 35 | 56 | 67 | 33 |

a) In Prozent. b) Diese Arbeit. Die Abkürzungen entsprechen den in Tab. 4 (S. 88) eingeführten. c) Die verwendeten Bezeichnungen für die Methode entsprechen den in (Knegtel *et al.*, 1999) gewählten. Die Zahlenangaben wurden dort Abb. 5 A und Tab. 4 entnommen. d) Keine Angaben vorhanden. e) Die Angaben wurden Tab. 4 in (Kramer *et al.*, 1999) entnommen. f) Die Angaben wurden Tab. 3 in (Jones *et al.*, 1997) entnommen.

Bei der Generierung von „guten“ Ligandgeometrien (*rmsd* < 2.0 Å) unabhängig von ihrer Bewertung werden von allen drei hier verglichenen Programmen ähnliche Ergebnisse im Be-

reich von 60 bis 70 % für Datensätze mit verschiedenen Proteinen erhalten, wenn für DOCK die „Energiefunktion“ während des Ligandaufbaus verwendet wird. Im Fall der Verwendung der „Chemischen Bewertungsfunktion“ sinkt der Anteil dagegen auf 43 %. Von den Autoren der GOLD-Studie (Jones *et al.*, 1997) wurde hierzu angemerkt, daß *zusätzlich* zu den im rechten Teil der Tabelle aufgeführten Zahlenwerten bei 7 Komplexen zwar eine „akzeptable Lösung“ erzeugt, aber nicht am besten bewertet wurde. Allerdings wurden keine weiteren Angaben zum *rmsd*-Wert dieser Lösungen gegeben. Zu beachten ist auch, daß die Auswahl der Protein-Ligand-Konfigurationen ohne Beachtung ihres Ranges bei FlexX aus > 400 erzeugten Lösungen erfolgt, bei DOCK mit den hier gewählten Parametern (Tab. 5, S. 89) aus 50 Lösungen sowie bei GOLD aus 20 Lösungen. Interessanterweise erhalten Knegtel *et al.* (Knegtel *et al.*, 1999) beim flexiblen Docking der 32 Thrombininhibitoren mit DOCK bessere Ergebnisse bei der Verwendung der „Chemischen Bewertungsfunktion“ im Vergleich zur „Energiefunktion“. In ihrem Fall schneidet FlexX bei der Ligandplazierung am schlechtesten ab. Betrachtet man nur noch Lösungen auf dem ersten Bewertungsrang, so ergeben sich für GOLD in 67 % aller Fälle „gute“ Ligandgeometrien (*rmsd* < 2.0 Å) und für FlexX in 47 % aller Fälle, dagegen für DOCK nur in 35 % („Energiefunktion“) bzw. 21 % („Chemische Bewertungsfunktion“) aller Fälle.

### 5.4.2 Korrelation der berechneten Bewertung für Protein-Ligand-Anordnungen mit ihrem *rmsd*-Wert bezüglich der Kristallstruktur

Die Eignung der entwickelten Bewertungsfunktion zur Erkennung nativ-ähnlicher Protein-Ligand-Geometrien in einer Menge von Alternativen wird zunächst an vier Beispielen gezeigt. Hierzu werden für mit FlexX generierte Docking-Lösungen der kristallographisch bekannten Komplexe 1bbp (Bilin-bindendes Protein im Komplex mit Biliverdin IX), 1lst (Lysin- / Arginin- / Ornithin-bindendes Protein im Komplex mit Lysin), 1rbp (Retinol-bindendes Protein im Komplex mit Retinol) und 2ada (Adenosin-Desaminase im Komplex mit 6-Hydroxy-1,6-dihydropurin-ribonucleosid) mit Gl. 36 Bewertungen berechnet und gegen den *rmsd*-Wert der jeweiligen Protein-Ligand-Anordnung in Bezug auf die Kristallstruktur aufgetragen (Abb. 21). Die hierzu verwendeten Parameter der Bewertungsfunktion entsprechen den in Kap. 5.1 und 5.2 beschriebenen; der $\gamma$-Wert für die Gewichtung von Paar- und Einteilchenpotentialen nach Gl. 36 wurde auf 0.5 gesetzt. Zusätzlich wird auch die Bewertung der im Kristall vorliegenden Ligandkonfiguration vorgenommen. Die erhaltenen Bewertungen werden aus Vergleichsgründen auf Zahlenwerte zwischen 0 und 100 normiert; kleine Bewertun-

gen weisen auf günstige Protein-Ligand-Konfigurationen hin. Die Komplexe wurden so ausgewählt, daß sie möglichst verschiedene Testfälle repräsentieren. So beträgt der Anteil nichtpolarer Atome an der Gesamtatomzahl zwischen 52 % (2ada) und 95 % (1rbp), die Anzahl drehbarer Bindungen liegt zwischen 1 (1rbp, 2ada) und 6 (1bbp) und die Anzahl von FlexX jeweils generierter Lösungen zwischen 99 (1rbp) und 289 (1lst).

Für die so erhaltenen Auftragungen ist zu beachten, daß sie eine Projektion der multidimensionalen Bewertungshyperfläche auf den *rmsd*-Wert als einzigen geometrischen Deskriptor darstellen. Somit ist eine genaue Vorhersage des zu erwartenden Zusammenhangs zwischen beiden Größen *a priori* nicht möglich; es sollte jedoch erwartet werden, daß die Kristallstruktur sowie nur geringfügig davon abweichende Ligandgeometrien besser bewertet werden als Konfigurationen mit großen *rmsd*-Werten. Der untere, zur Abszisse gewandte Rand der Auftragungen kann außerdem mehrere lokale Minima aufweisen in Abhängigkeit von der Topographie der betrachteten Bewertungshyperfläche. Zusätzlich ist zu berücksichtigen, daß ähnliche *rmsd*-Werte nicht notwendigerweise auch auf ähnliche Protein-Ligand-Anordnungen hindeuten, was insbesondere im Bereich großer *rmsd*-Werte gilt. Aus diesem Grund können auch für Ligandkonfigurationen mit ähnlichen *rmsd*-Werten unterschiedliche Bewertungen mit Gl. 36 erhalten werden. Allerdings sollten geometrisch ähnliche Protein-Ligand-Anordnungen auch eine ähnliche Bewertung erhalten, vorausgesetzt, die zugrundeliegenden Potentiale sind ausreichend „weich“, so daß geringfügige geometrische Abweichungen toleriert werden.

Für alle in Abb. 21 gezeigten Beispiele wird für die kristallographisch bestimmte Protein-Ligand-Anordnung (*rmsd* = 0) die beste Bewertung erhalten. Für 1bbp, 1lst und 2ada ist die native Geomtrie darüber hinaus energetisch deutlich von allen durch FlexX erzeugten Ligandkonfigurationen abgesetzt. Zusätzlich sind in allen vier Fällen stark von der Kristallstruktur abweichende Ligandkonfigurationen von den nativ-ähnlichen Geometrien merklich separiert. Für letztere Lösungen kann zudem eine Zunahme der Bewertung mit zunehmendem *rmsd*-Wert beobachtet werden.

Zusammenfassend vermag es die in Gl. 36 definierte Bewertungsfunktion in allen vier vorgestellten Fällen, die experimentelle bzw. „gut“ gedockte Ligandkonfigurationen in einer Menge von bis zu 289 durch FlexX erzeugten Alternativen zu identifizieren.

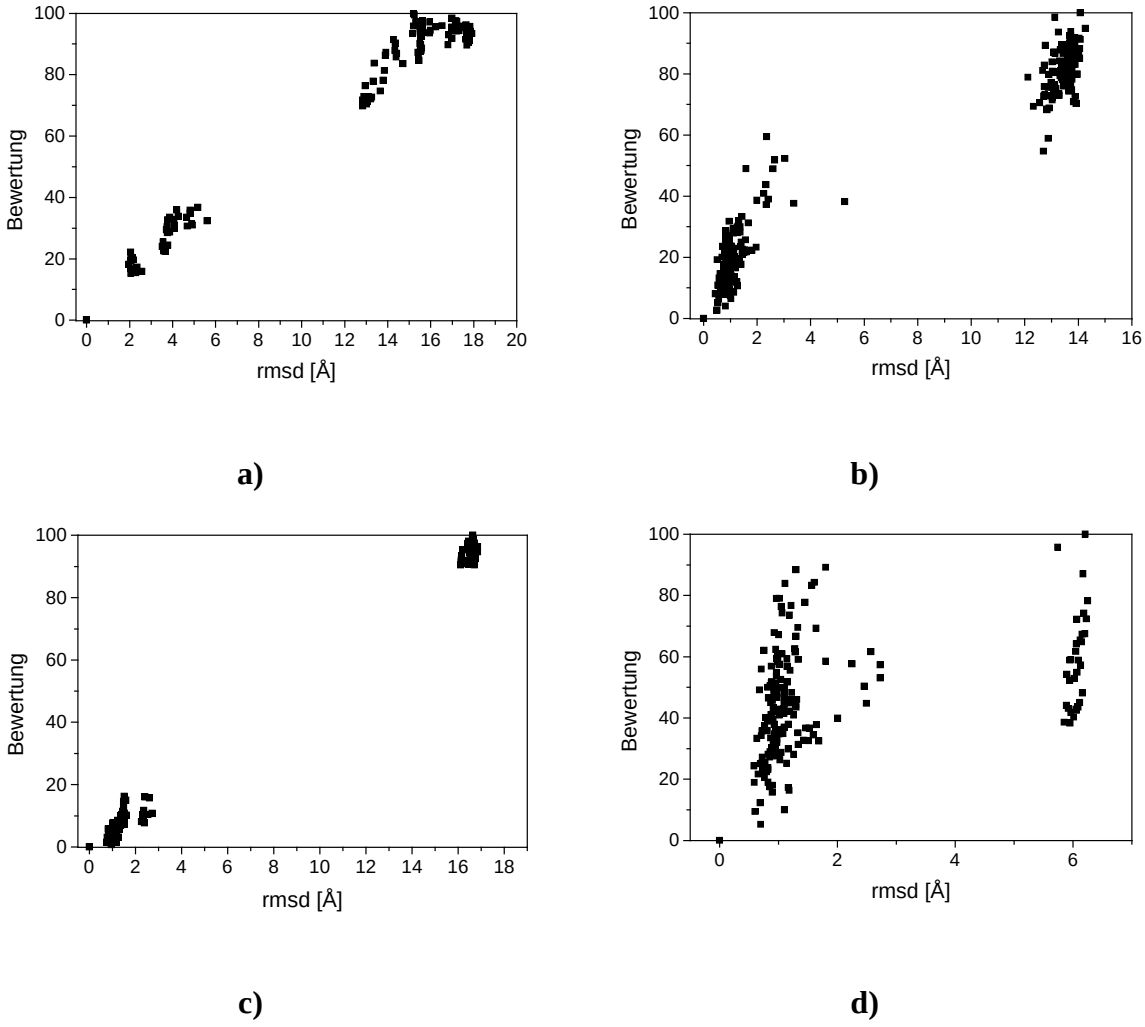



**Abb. 21:** Auftragungen von nach Gl. 36 erhaltenen und auf Werte zwischen 0 und 100 normierten Bewertungen gegen den *rmsd*-Wert bzgl. der Kristallstruktur von mit FlexX erzeugten Ligandgeometrien für die Komplexe 1bbp (a), 1lst (b), 1rbp (c) und 2ada (d). Kleine Ordinatenwerte stehen dabei für günstige Protein-Ligand-Anordnungen. Die Kristallstruktur (*rmsd* = 0) ist in allen Fällen hinzugefügt.

### 5.4.3 Erkennung von nativ-ähnlichen Protein-Ligand-Konfigurationen

Während die in dem vorherigen Kapitel vorgestellten Ergebnisse einen ersten Eindruck von der Eignung der entwickelten Bewertungsfunktion für die Erkennung nativ-ähnlicher Ligandkonfigurationen vermitteln, erfolgt hier eine statistische Evaluierung dieser Fähigkeit basierend auf den Datensätzen FlexX_DS1 (91 kristallographisch bestimmte Protein-Ligand-Komplexe), FlexX_DS2 (68 Komplexe) und DOCK_DS (100 Komplexe) (s. a. Kap. 4.9.1). FlexX_DS1 wurde als Kalibrierungsdatensatz für die in Kap. 4.2 vorgestellten Parameter- und Berechnungsalternativen verwendet. FlexX_DS2 wurde dagegen nicht zur Parameteranpas-

sung genutzt und dient daher der Kreuzvalidierung. Um zu untersuchen, inwieweit die verwendete Methode zur Generierung der Protein-Ligand-Anordnungen einen Einfluß auf die erhaltenen Ergebnisse ausübt, wurde zusätzlich der Datensatz DOCK_DS verwendet.

FlexX generiert für FlexX_DS1 in 84 % aller Fälle überhaupt eine „gute“ (*rmsd* < 2.0 Å) Ligandkonfiguration (d.h. ohne Berücksichtigung ihrer Bewertung), in 54 % davon wird wiederum eine „gute“ Lösung auch als beste von FlexX bewertet. Im Hinblick auf den letzten Zahlenwert und im Vergleich mit Tab. 19 stellt FlexX_DS1 bezüglich den von FlexX damit erzielten Ergebnissen eine repräsentative Auswahl von Komplexen dar. Für die Komplexe des Datensatzes werden im Mittel jeweils 263 (Standardabweichung: 118) mögliche Protein-Ligand-Konfigurationen von FlexX erzeugt.

In Abb. 22 ist die akkumulierte Anzahl von Komplexen aus FlexX_DS1 als Funktion des *rmsd*-Wertes bezüglich der jeweiligen Kristallstruktur für die Protein-Ligand-Anordnungen dargestellt, die von FlexX bzw. der Bewertungsfunktion nach Gl. 36 jeweils am besten innerhalb einer Menge von generierten Alternativen bewertet wurden. Ebenfalls mit dargestellt ist die akkumulierte Anzahl von Komplexen als Funktion des kleinsten *rmsd*-Wertes, der jeweils von FlexX ohne Beachtung der Bewertung generiert wurde. Diese Kurve gibt einen Eindruck des Limits, das eine ideale Bewertungsfunktion erreichen könnte.

Wie aus dem erhaltenen Diagramm zu ersehen ist, wird durch die hier entwickelte Bewertungsfunktion für deutlich mehr Protein-Ligand-Komplexe eine „gute“ Ligandkonfiguration am besten bewertet als durch die in FlexX implementierte Funktion, die auf die Arbeit von Böhm (Böhm, 1994) zurückgeht. Der Anteil der Testfälle in FlexX_DS1, für den eine Lösung mit einem *rmsd* < 1.0, < 1.5, < 2.0 sowie ≥ 2.0 Å auf dem ersten Bewertungsrang gefunden wird, ist in Tab. 20 aufgeführt. In Bezug auf die Erkennung von Ligandkonfigurationen mit einem *rmsd* < 2.0 Å wird der Anteil von ursprünglich 54 % durch die Bewertung mit der hier entwickelten Funktion auf 73 % gesteigert, was einer Verbesserung um 35 % Prozent entspricht.

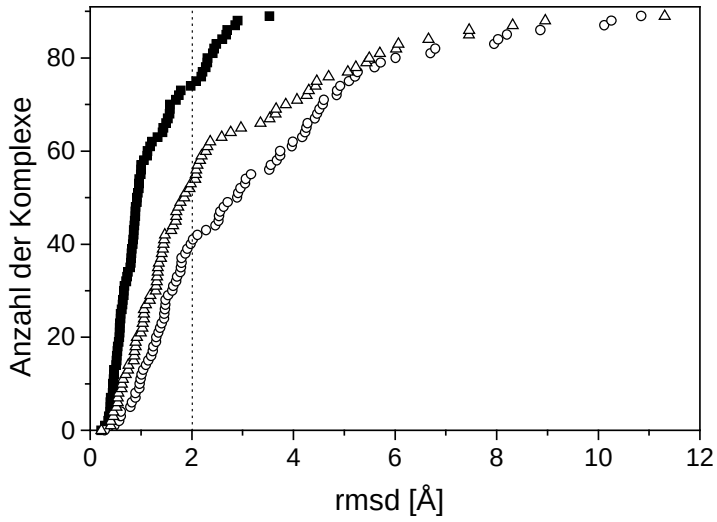



**Abb. 22:** Akkumulierte Anzahl von Protein-Ligand-Komplexen aus FlexX_DS1 als Funktion des *rmsd*-Wertes bezüglich der jeweiligen Kristallstruktur für die Protein-Ligand-Anordnungen, die von FlexX (o) bzw. der hier entwickelten Bewertungsfunktion (Gl. 36) (Δ) jeweils am besten innerhalb einer Menge von generierten Alternativen bewertet wurden. Ebenfalls angegeben ist die akkumulierte Anzahl von Komplexen, für die eine Geometrie mit kleinstem *rmsd*-Wert unabhängig von einer Bewertung überhaupt generiert wurde (■). Diese Kurve gibt einen Eindruck des Limits, das eine ideale Bewertungsfunktion erreichen könnte. Alle Kurven wurden unabhängig voneinander nach dem *rmsd*-Wert sortiert.

**Tab. 20:** Ergebnisse für von FlexX generierte Docking-Lösungen für 91 Komplexe des Datensatzes FlexX_DS1, die von der in FlexX implementierten Funktion sowie der hier entwickelten (Gl. 36) bewertet wurden.

| | | **Anteil[a] Komplexe mit *rmsd*** | | | |
|---|---|---|---|---|---|
| | | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** |
| **Alle Bewertungsränge[b]** | | 65 | 76 | 84 | 16 |
| **1. Rang[c]** | **FlexX** | 20 | 37 | 54 | 46 |
| | **Gl. 36** | 39 | 66 | 73 | 27 |
| **Verbesserung[d]** | | 95 | 78 | 35 | -41 |

a) In Prozent. b) Alle mit FlexX erzeugten Protein-Ligand-Anordnungen der 91 Komplexe in FlexX_DS1 werden betrachtet. Der Zahlenwert gibt somit den Anteil der Komplexe an, für den eine Lösung mit jeweiligem *rmsd*-Wert überhaupt (d.h. ohne Beachtung ihrer Bewertung) generiert wurde. c) Nur diejenigen Ligandkonfigurationen werden betrachtet, die jeweils am besten bewertet werden. Die Zahlenwerte sind dabei auf die in der ersten Zeile bezogen. d) Die Verbesserung wird gemäß (%(Gl. 36) - %FlexX) / %FlexX berechnet.

Abb. 23 zeigt, um wieviel sich jeweils der *rmsd*-Wert für Protein-Ligand-Geometrien auf Rang 1 nach Bewertung mit FlexX bzw. der hier entwickelten Funktion für einzelne Komplexe unterscheidet. Hierbei wird in 10 Fällen von der Funktion nach Gl. 36 eine Lösung als am besten bewertet, die um mehr als 1 Å *weiter* von der Kristallgeometrie abweicht als die von FlexX als am besten bewertete. Demgegenüber werden in 28 Fällen durch die hier entwickelte Funktion Ligandanordnungen als beste ausgewählt, die um mehr als 1 Å *näher* an der Kristallstruktur liegen. Gemittelt über alle 91 Komplexe ergibt sich so eine Reduzierung der *rmsd*-Werte für Ligandkonfigurationen auf Rang 1 um 0.7 Å.

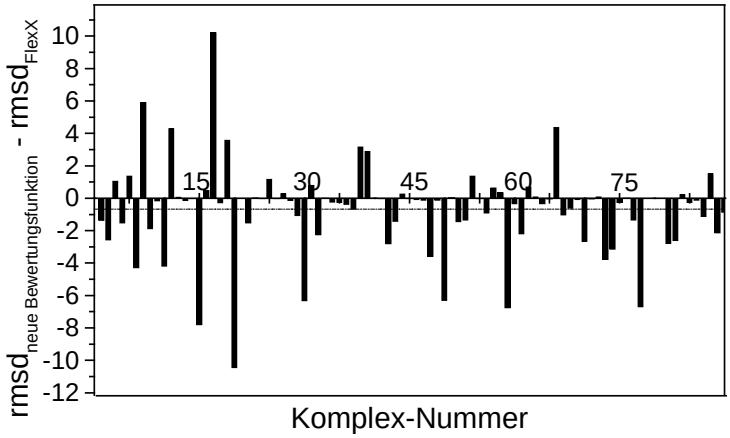



**Abb. 23:** Die Differenz der *rmsd*-Werte für Lösungen, die von der hier entwickelten Funktion (Gl. 36) bzw. von FlexX am besten bewertet wurden, sind für alle 91 Komplexe dargestellt. Die gestrichelt gezeichnete Linie zeigt den Mittelwert aller Differenzen bei –0.7 Å.

Für den zweiten Datensatz FlexX_DS2 mit 68 Komplexen wird von FlexX nur in 28 Fällen eine Lösung mit einem *rmsd* < 2.0 Å auf dem ersten Bewertungsrang gefunden (Tab. 21). Für 38 Protein-Ligand-Komplexe erzeugt FlexX dagegen überhaupt keine Ligandkonfigurationen mit *rmsd* < 2.0 Å. Somit findet FlexX hier in 93 % der möglichen Fälle eine „gut“ gedockte Lösung auf dem ersten Rang. Mit der hier entwickelten Bewertungsfunktion ergibt sich eine nahezu identische Erfolgsrate von 90 %.

**Tab. 21:** Ergebnisse für von FlexX generierte Docking-Lösungen für 68 Komplexe des Datensatzes FlexX_DS2, die von der in FlexX implementierten Funktion sowie der hier entwickelten (Gl. 36) bewertet wurden.

| | | **Anteil[a] Komplexe mit *rmsd*** | | | |
|---|---|---|---|---|---|
| | | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** |
| **Alle Bewertungsränge[b]** | | 35 | 44 | 44 | 56 |
| **1. Rang[c]** | **FlexX** | 54 | 66 | 93 | 7 |
| | **Gl. 36** | 38 | 60 | 90 | 10 |
| **Verbesserung[d]** | | -30 | -9 | -3 | 43 |

a) In Prozent. b) Alle mit FlexX erzeugten Protein-Ligand-Anordnungen der 68 Komplexe in FlexX_DS2 werden betrachtet. Der Zahlenwert gibt somit den Anteil der Komplexe an, für den eine Lösung mit jeweiligem *rmsd*-Wert überhaupt (d.h. ohne Beachtung ihrer Bewertung) generiert wurde. c) Nur diejenigen Ligandkonfigurationen werden betrachtet, die jeweils am besten bewertet werden. Die Zahlenwerte sind dabei auf die in der ersten Zeile bezogen. d) Die Verbesserung wird gemäß (%(Gl. 36) - %FlexX) / %FlexX berechnet.

Für den DOCK_DS-Datensatz werden unter Verwendung der Methoden flex_nrg_min („Energiefunktion“) und flex_chm_min („Chemische Bewertungsfunktion“) von DOCK in 57 % bzw. 48 % der möglichen Fälle eine Ligandkonfiguration mit *rmsd* < 2.0 Å am besten bewertet (Tab. 22 und Tab. 23; siehe dazu auch Tab. 19 (S. 129)). In Bezug auf diese „gut“ gedockten Lösungen wird durch die hier entwickelte Bewertungsfunktion dieser Anteil auf 70 % (das entspricht einer Verbesserung um 46 %) gesteigert, wenn die nach flex_chm_min erzeugten Lösungen betrachtet werden (Tab. 23 und Abb. 24). Für die Protein-Ligand-Anordnungen, die mit der „Energie-Bewertung“ von DOCK erzeugt wurden, ergibt sich dafür mit der hier entwickelten Funktion ein Anteil von 54 % (Tab. 22).

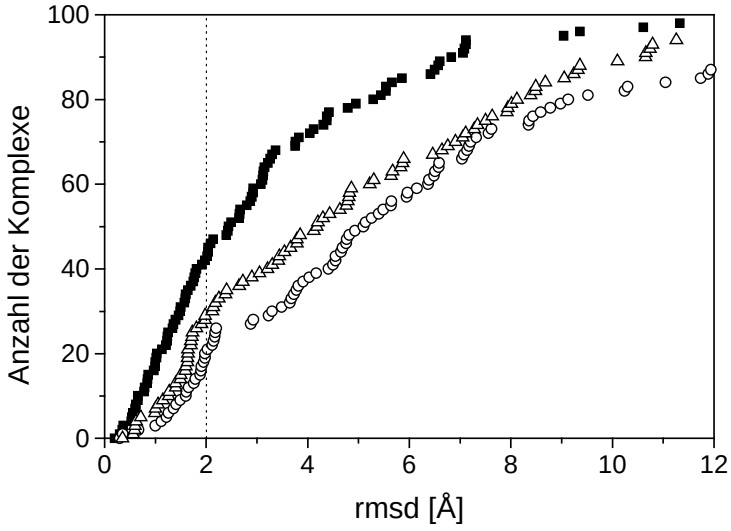



**Abb. 24:** Akkumulierte Anzahl von Protein-Ligand-Komplexen aus DOCK_DS als Funktion des *rmsd*-Wertes bezüglich der jeweiligen Kristallstruktur für die Protein-Ligand-Anordnungen, die von der „Chemischen Bewertungsfunktion“ in DOCK (o) bzw. der hier entwickelten Bewertungsfunktion (Gl. 36) (Δ) jeweils am besten innerhalb einer Menge von generierten Alternativen bewertet wurden. Ebenfalls angegeben ist die akkumulierte Anzahl von Komplexen, für die eine Geometrie mit kleinstem *rmsd*-Wert unabhängig von einer Bewertung überhaupt generiert wurde (■). Diese Kurve gibt einen Eindruck des Limits, das eine ideale Bewertungsfunktion erreichen könnte. Alle Kurven wurden unabhängig voneinander nach dem *rmsd*-Wert sortiert.

**Tab. 22:** Ergebnisse für von DOCK generierte Docking-Lösungen für 100 Komplexe des Datensatzes DOCK_DS, die von der in DOCK implementierten „Energiefunktion“ sowie der hier entwickelten (Gl. 36) bewertet wurden.

| | | **Anteil[a] Komplexe mit *rmsd*** | | | |
|---|---|---|---|---|---|
| | | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** |
| **Alle Bewertungsränge[b]** | | 25 | 52 | 61 | 39 |
| **1. Rang[c]** | **DOCK** | 32 | 42 | 57 | 43 |
| | **Gl. 36** | 40 | 42 | 54 | 46 |
| **Verbesserung[d]** | | 25 | 0 | -5 | 7 |

a) In Prozent. b) Alle mit DOCK erzeugten Protein-Ligand-Anordnungen der 100 Komplexe in DOCK_DS werden betrachtet. Der Zahlenwert gibt somit den Anteil der Komplexe an, für den eine Lösung mit jeweiligem *rmsd*-Wert überhaupt (d.h. ohne Beachtung ihrer Bewertung) generiert wurde (s. a. Tab. 19, S. 129). c) Nur diejenigen Ligandkonfigurationen werden betrachtet, die jeweils am besten bewertet werden. Im Gegensatz zu Tab. 19 sind die Zahlenwerte dabei auf die in der ersten Zeile bezogen. d) Die Verbesserung wird gemäß (%(Gl. 36) - %DOCK) / %DOCK berechnet.

**Tab. 23:** Ergebnisse für von DOCK generierte Docking-Lösungen für 100 Komplexe des Datensatzes DOCK_DS, die von der in DOCK implementierten „Chemischen Bewertungsfunktion“ sowie der hier entwickelten (Gl. 36) bewertet wurden.

| | | **Anteil[a] Komplexe mit *rmsd*** | | | |
|---|---|---|---|---|---|
| | | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** |
| **Alle Bewertungsränge[b]** | | 17 | 31 | 43 | 57 |
| **1. Rang[c]** | **DOCK** | 18 | 32 | 48 | 51 |
| | **Gl. 36** | 41 | 48 | 70 | 30 |
| **Verbesserung[d]** | | 128 | 50 | 46 | -41 |

a) In Prozent. b) Alle mit DOCK erzeugten Protein-Ligand-Anordnungen der 100 Komplexe in DOCK_DS werden betrachtet. Der Zahlenwert gibt somit den Anteil der Komplexe an, für den eine Lösung mit jeweiligem *rmsd*-Wert überhaupt (d.h. ohne Beachtung ihrer Bewertung) generiert wurde (s. a. Tab. 19, S. 129). c) Nur diejenigen Ligandkonfigurationen werden betrachtet, die jeweils am besten bewertet werden. Im Gegensatz zu Tab. 19 sind die Zahlenwerte dabei auf die in der ersten Zeile bezogen. d) Die Verbesserung wird gemäß (%(Gl. 36) - %DOCK) / %DOCK berechnet.

### 5.4.4 Erkennung von kristallographisch bestimmten Protein-Ligand-Anordnungen

Unter der Annahme, die kristallographisch bestimmte Protein-Ligand-Anordnung sei „optimal“, ist für sie die jeweils *beste* Bewertung mit Gl. 36 unter allen von Docking-Programmen generierten und davon geometrisch abweichenden Lösungen zu erwarten (s. a. Kriterium III in Kap. 4.6.1). Gemäß Abb. 25 ist dieses auch für 54 % aller 91 Protein-Ligand-Komplexe in FlexX_DS1 der Fall. Wird ein gelockertes Kriterium definiert, nach dem die Bewertungsfunktion auch dann als erfolgreich bezeichnet wird, wenn entweder die Kristallstruktur oder „gute“ (*rmsd* < 2.0 Å) gedockte Ligandkonfigurationen am besten bewertet werden, so ist dieses sogar in 71 % aller Fälle für FlexX_DS1 erfüllt.

Wird die entwickelte Bewertungsfunktion auf den Datensatz FlexX_DS2 angewendet, so werden in 65 % aller Fälle die Kristallgeometrien der Liganden auf dem ersten Rang gefunden. In sogar 90 % aller 68 Protein-Ligand-Komplexe werden nach dem gelockerten Kriterium entweder die Kristallstruktur oder eine „gut“ gedockte Ligandkonfiguration am besten bewertet.

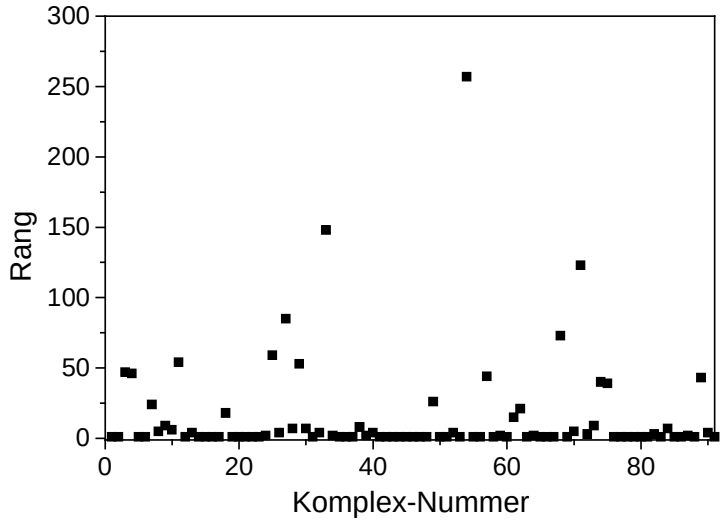



**Abb. 25:** Nach Gl. 36 berechneter Rang der Kristallstruktur unter allen von FlexX erzeugten Alternativen für jeweils alle 91 Protein-Ligand-Komplexe des Datensatzes FlexX_DS1.

Um zu untersuchen, inwieweit die Qualität der durch ein Docking-Programm zusätzlich generierten Lösungen das Auffinden der Kristallstruktur beeinflußt, wurden wiederum die mit den Methoden flex_nrg_min und flex_chm_min (s. Tab. 4, S. 88) durch DOCK erzeugten Protein-Ligand-Anordnungen des Datensatzes DOCK_DS verwendet. In 76 % (57 %) aller 100 betrachteten Protein-Ligand-Komplexe wird für die mit der „Chemischen Bewertungsfunktion“ („Energiefunktion“) erzeugten Lösungen die Kristallgeometrie des Liganden auf Rang 1 gefunden. Unter Verwendung des gelockerten Kriteriums werden in 83 % (62 %) entweder eine Kristallstruktur oder aber eine „gut“ gedockte (*rmsd* = 2.0 Å) Ligandkonfiguration am besten bewertet.

Faßt man die Ergebnisse für die Datensätze FlexX_DS1 und FlexX_DS2 zusammen, so wird in nahezu 80 % aller 159 untersuchten Fälle eine Kristallstruktur oder eine dazu „ähnliche“ (*rmsd* < 2.0 Å) Geometrie auf dem *ersten* Bewertungsrang erhalten. Für die mit der „Chemischen Bewertungsfunktion“ in DOCK generierten Protein-Ligand-Konfigurationen des Datensatzes DOCK_DS gleichen die erhaltenen Ergebnisse denen mit den FlexX_DS-Datensätzen. Im Falle der mit der auf dem AMBER-Kraftfeld (Weiner *et al.*, 1984) beruhenden „Energiefunktion“ erzeugten Ligandanordnungen wird dagegen ein geringerer Anteil richtig erkannt, wobei auch dieses Ergebnis noch über den von den Standardverfahren in FlexX und DOCK erzielten liegt (Tab. 19, S. 129). Insgesamt betrachtet ist die hier erhaltene Erkennungsrate von nativ(-ähnlichen) Protein-Ligand-Anordnungen bislang noch von keinem

anderen Bewertungsverfahren in der Literatur beschrieben worden (s.a. Zusammenfassung der Validierungsstudien von Dockingprogrammen in Tab. 19, S. 129). So wird die Eignung zur Erkennung nativ-ähnlicher Bindungsmoden des im Verlauf dieser Arbeit veröffentlichten Ansatzes von Mitchell *et al.* (Mitchell *et al.*, 1999a; Mitchell *et al.*, 1999b) nur am Beispiel der Heparinbindung an bFGF (PDB-Code: 1bfc) getestet. Während hierbei die Kristallstruktur als günstigste erkannt werden konnte, weicht die von FTdock generierte und anschließend am besten bewertete Geometrie von der experimentell bestimmten deutlich ab (ein *rmsd*-Wert wurde von den Autoren nicht angegeben). Eine zuverlässige Erkennung „gut“ gedockter Protein-Ligand-Anordnungen ist jedoch eine notwendige Voraussetzung für eine verläßliche Vorhersage von Bindungsaffinitäten im Rahmen von *virtual screening*-Anwendungen (s.a. Kap. 5.5.3).

Bemerkenswerterweise kann z.B. für FlexX_DS2 die korrekte Erkennung nativ(-ähnlicher) Protein-Ligand-Konfigurationen von 27 Fällen (bei nur 30 möglichen (Tab. 21)) unter Verwendung ausschließlich gedockter Anordnungen auf 61 Fälle gesteigert werden, sofern die Kristallstruktur des Liganden mit berücksichtigt wird und somit in den verbleibenden 38 Fällen überhaupt eine geeignete Ligandgeometrie bewertet werden kann. Diese Tatsache weist auf die Notwendigkeit einer Verbesserung der Ligand*plazierung* in den verwendeten Dokkingprogrammen hin.

Für einige der Fälle, in denen die Kristallgeometrie des Liganden nicht am besten bewertet wird, lassen sich bei Betrachtung der experimentell bestimmten Struktur mögliche Ursachen identifizieren. Im Fall der R106Q-Mutante des Fettsäure-bindenden Proteins mit Ölsäure ((Z)-9-Octadecensäure) als Ligand (PDB-Code: 1icn, Auflösung 1.74 Å, Abb. 26) wird die Kristallstruktur auf dem 35. Rang bewertet. Der Ligand ist in der nativen Struktur so orientiert, daß seine Carboxylatgruppe in das Innere des Proteins steht und das nichtpolare Kohlenwasserstoffende zum Lösemittel hin ausgerichtet ist. Die Carboxylatgruppe bildet dabei keine gerichtete Wechselwirkung mit einer funktionellen Gruppe des Proteins innerhalb von 3.5 Å; lediglich ein Wassermolekül tritt in unmittelbarer Nachbarschaft (3.3 Å) auf. Anstelle der Kristallgeometrie wird dagegen ein Ligandbindungsmodus als am besten bewertet, der eine umgekehrte Orientierung in der Bindetasche ergibt (*rmsd* = 11.3 Å). Hierbei ist nicht nur der nichtpolare Teil im Inneren des Proteins vergraben, sondern die Carboxylatgruppe bildet darüber hinaus noch Wasserstoffbrücken zum Amid-Stickstoff von Ala73 aus. Beide Effekte bedingen die wesentlich günstigere Bewertung.

Interessanterweise ist die Carboxylatgruppe in der Kristallstruktur ungeordnet, wobei drei alternative Positionen in der PDB-Datei angegeben werden. Darüber hinaus wird von den Autoren (Eads *et al*., 1993) eine geringe, aber kontinuierliche, J-förmige Elektronendichte berichtet, die hinter der terminalen Methylgruppe der Ölsäure im Komplex mit dem Wildtyp-protein auftritt. Dieses hat am Boden der Bindetasche außerdem ein Arginin (Arg106), das in dem für den hier beschriebenen Komplex verwendeten Protein zu Glutamin mutiert wurde. Somit entfällt in diesem Fall auch die Möglichkeit, daß die Carboxylatgruppe der Ölsäure dort eine Salzbrücke ausbilden kann.

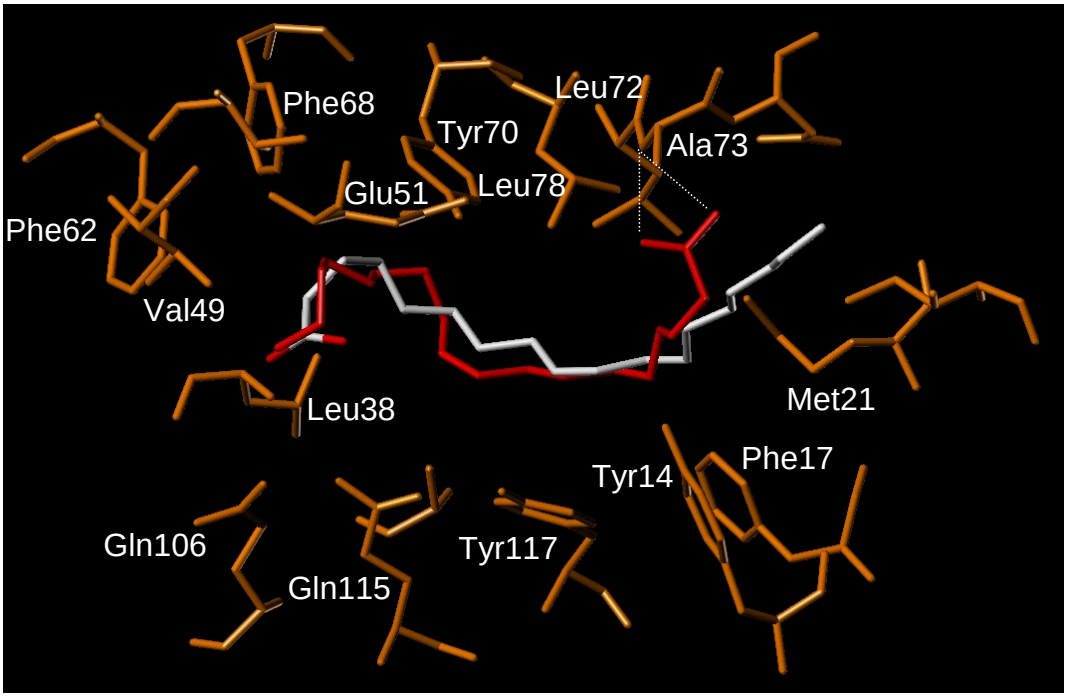



**Abb. 26:** Dargestellt ist ein Ausschnitt aus der Bindetasche von 1icn (R106Q-Mutante des Fettsäure-bindenden Proteins) zusammen mit der Kristallstruktur des Liganden Ölsäure ((Z)-9-Octadecensäure; Farbcodierung nach Atomtypen), die nach Gl. 36 auf dem 35. Rang gefunden wird, sowie einer mit *rmsd* = 11.3 Å davon abweichenden Ligandkonfiguration (rot), die am besten bewertet wurde. Der Eingang zu der Bindetasche befindet sich auf der rechten Seite der Abbildung, ober- und unterhalb der Bindetasche liegende Aminosäuren wurden aus Gründen der Übersichtlichkeit entfernt. Die Carboxylatgruppe des Liganden in der Kristallanordnung zeigt keine Wechselwirkung zum Protein, sondern lediglich zu einem 3.3 Å entfernten Wassermolekül (hier nicht dargestellt). Im Gegensatz dazu bildet die besser bewertete Ligandkonfiguration nicht nur Wasserstoffbrücken zwischen der Carboxylatgruppe und einem Amidstickstoff von Ala73 aus (gepunktete Linien; die Abstände zwischen den jeweiligen Atomen betragen 2.8 / 2.9 Å), sondern die nichtpolare Kohlenwasserstoffkette der Fettsäure ist auch noch zum Inneren des Proteins gerichtet.

In Anbetracht dieser Tatsachen kann die beobachtete Elektronendichte nun auch so interpretiert werden, daß die Ölsäure tatsächlich umgekehrt in der Bindetasche plaziert ist. Der

ungeordnete Teil im Proteininneren würde dann dem Ende der Kohlenwasserstoffkette entsprechen, was auch mit der vorhergesagten Ligandanordnung in Übereinstimmung wäre. Bemerkenswerterweise können sowohl Jones *et al.* (Jones *et al.*, 1997) als auch Hoffmann *et al.* (Hoffmann *et al.*, 1999) für diesen Komplex ebenfalls keine mit der experimentellen Struktur im Einklang stehende Ligandanordnung auf dem ersten Bewertungsrang erhalten.

Im Komplex aus dem Protein *Human Plasminogen Kringle 4* mit ε-Aminohexansäure (PDB-Code: 2pk4, Auflösung 2.25 Å, Abb. 27) liegt der Ligand in einer flachen Einbuchtung an der Proteinoberfläche. Während eine mit FlexX erzeugte Ligandkonfiguration mit *rmsd* = 1.3 Å bezüglich der Kristallstruktur am besten bewertet wird, zeigt die native Struktur (Rang 129) eine Wasserstoffbrücke zwischen der Aminogruppe des Liganden und der Carboxylatgruppe von Asp55, die lediglich 2.1 Å lang ist. Bedingt durch die kurze Distanz wird diese Wechselwirkung vom N.3-O.co2-Paarpotential als repulsiv gewertet. Dies steht auch im Einklang mit den beschriebenen Eigenschaften starker Wasserstoffbrücken mit hohem kovalentem Anteil, für die als untere Grenze für den Abstand zwischen den beteiligten Nichtwasserstoffatomen 2.2 Å angegeben wird (Jeffrey, 1997) (s. a. Kap. 2.2.1).

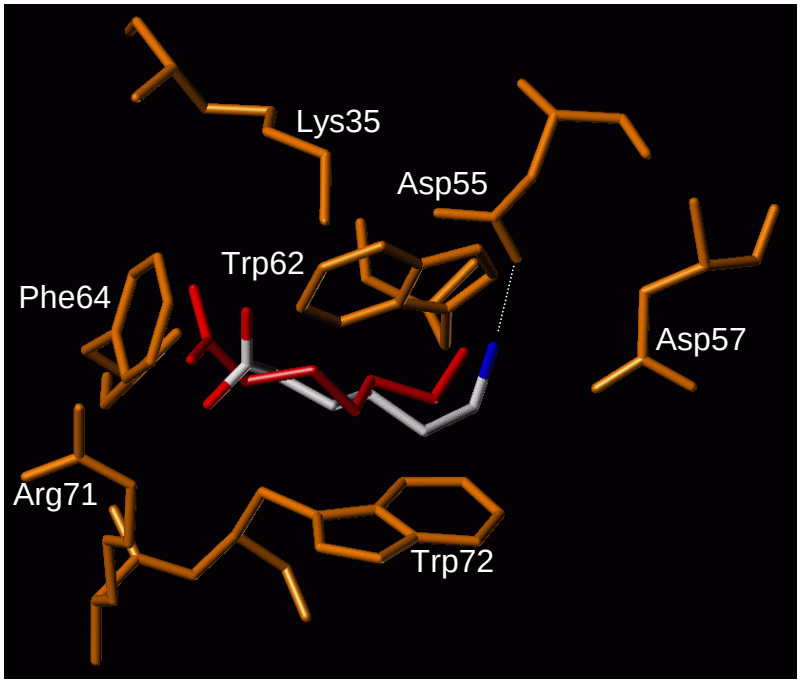



**Abb. 27:** Dargestellt ist ein Ausschnitt aus der Bindetasche von 2pk4 (*Human Plasminogen Kringle 4* Protein) zusammen mit der Kristallgeometrie des Liganden ε-Aminohexansäure (Rang 129 nach Bewertung gemäß Gl. 36; Farbcodierung nach dem Atomtyp). Die bestbewertete Ligandkonfiguration (*rmsd* = 1.3 Å) ist rot gefärbt. Die Ligandanordnung im Kristall führt zu einer Wasserstoffbrücke zwischen der terminalen Aminogruppe und einer Carboxylatgruppe von Asp55 mit einer Länge von nur 2.1 Å (gepunktete Linie).

### 5.4.5 Untersuchung von Faktoren, die einen Einfluß auf die Erkennung nativ-ähnlicher Ligandkonfigurationen haben

In dem hier entwickelten Ansatz zur Bewertung von Protein-Ligand-Wechselwirkungen werden zwei wissensbasierte Präferenzen verwendet, die einerseits auf den Paarverteilungsfunktionen von Atom-Atom-Kontakten und zum anderen auf den Anteilen vergrabener Oberflächen von Protein- und Ligandatomen beruhen. Damit stellt sich die Frage, wieviel redundante Information in beiden Termen enthalten ist. In Abb. 28 ist die Eignung der in Gl. 15 bzw. Gl. 35 beschriebenen statistischen Paar- und Einteilchenpräferenzen gezeigt, jeweils *einzeln* verwendet „gut“ gedockte Ligandanordnungen in einer Menge von generierten Alternativen für die Komplexe des Datensatzes FlexX_DS1 zu erkennen. Dabei ist wiederum die akkumulierte Anzahl der Komplexe gegen den *rmsd*-Wert aufgetragen worden, der jeweils für die bestbewerteten Ligandanordnungen ermittelt wurde. Während die Paarpotentiale alleine signifikant besser als die FlexX-Bewertung geeignet sind, nativ-ähnliche Geometrien zu erkennen, ergibt sich bei Verwendung der SAS-abhängigen Einteilchenpotentiale ein ähnliches Ergebnis wie für die FlexX-Bewertung. Letzteres ist bemerkenswert, da diese Einteilchenpotentiale keinerlei detaillierte Informationen über die Eigenschaften des jeweiligen molekularen Wechselwirkungspartners (d.h. der Proteinbindetasche im Falle der Potentiale für Ligandatome sowie umgekehrt) benutzen. Sie berücksichtigen nur, ob ein polarer Bereich eines Moleküls in einem ebenfalls polaren oder aber unpolaren Bereich des Wechselwirkungspartners vergraben wird. Weitergehende (implizite) Informationen wie etwa über spezifische Atom-Atom-Kontakte oder aber die Geometrie von Wechselwirkungen werden dagegen nicht betrachtet.

Die Verwendung der aus beiden Termen zusammengesetzten Bewertungsfunktion nach Gl. 36 ($\gamma = 0.5$ entsprechend einer 1:1-Gewichtung) ergibt für den Datensatz FlexX_DS1 hinsichtlich der Anzahl korrekt erkannter Protein-Ligand-Anordnungen mit einem *rmsd*-Wert < 2.0 Å die gleichen Ergebnisse wie die Verwendung der Paarpotentiale alleine (Abb. 22, Abb. 28, Tab. 24). Im Fall von Protein-Ligand-Anordnungen mit höchstens 1 Å bzw. höchstens 1.5 Å *rmsd*-Wert läßt sich durch eine 1:1-Gewichtung von Paar- und Einteilchenpräferenz jedoch der Anteil richtig erkannter Konfigurationen von 37 % bzw. 63 % auf 39 % bzw. 66 % steigern. Tab. 24 zeigt, daß ein gewählter $\gamma$-Wert von 0.5 dabei optimal ist.

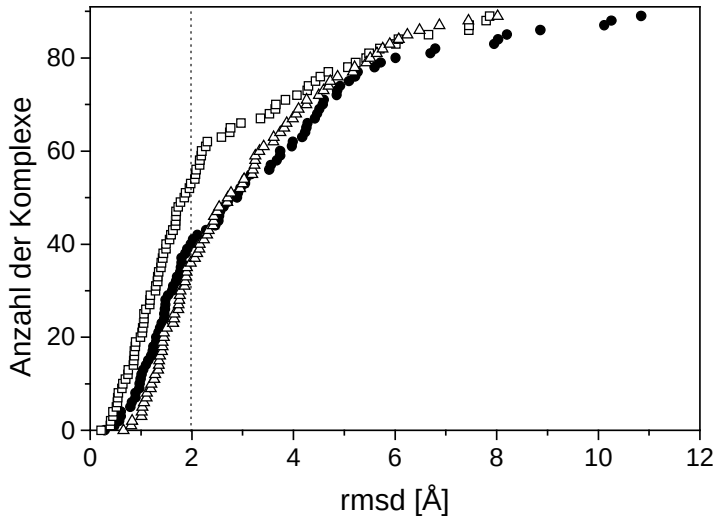



**Abb. 28:** Die Eignung der distanzabhängigen Paar- (□) sowie der SAS-abhängigen Einteilchenpotentiale (Δ) zur Erkennung nativ-ähnlicher Protein-Ligand-Anordnungen im Vergleich zur Bewertungsfunktion in FlexX (●) ist dargestellt für 91 Komplexe des Datensatzes FlexX_DS1. Der *rmsd*-Wert entspricht demjenigen der Protein-Ligand-Anordnungen auf dem ersten Rang unter Verwendung des jeweiligen Bewertungsverfahrens. Alle Kurven wurden unabhängig voneinander nach dem *rmsd*-Wert sortiert.

**Tab. 24:** Ergebnisse für von FlexX generierte Docking-Lösungen für 91 Komplexe des Datensatzes FlexX_DS1, die von der hier entwickelten Funktion (Gl. 36) unter Verwendung verschiedener Gewichtungen von Paar- zu Einteilchenpotential bewertet wurden.

| $\gamma$[a)] | Anteil[b)] Komplexe mit *rmsd* | | | |
|---|---|---|---|---|
| | **< 1.0 Å** | **< 1.5 Å** | **< 2.0 Å** | **≥ 2.0 Å** |
| **1.0 (1:0)** | 37 | 63 | 73 | 27 |
| **0.5 (1:1)** | 39 | 66 | 73 | 27 |
| **0.09 (1:10)** | 36 | 63 | 73 | 27 |
| **0.91 (10:1)** | 32 | 65 | 73 | 27 |

a) Faktor gemäß Gl. 36. In Klammern ist zusätzlich die dementsprechende Gewichtung von Paar- und Einteilchenpotentialen angegeben. b) In Prozent. Die Zahlen beziehen sich auf die Anzahl derjenigen Protein-Ligand-Komplexe in FlexX_DS1, für die überhaupt eine Lösung mit angegebenem *rmsd* (ohne Beachtung der Bewertung) generiert werden konnte (s.a. Tab. 20, S. 134, Zeile ‚Alle Bewertungsränge').

Ein illustratives Beispiel für einen Fall, der erst bei einer 10-fachen Gewichtung ($\gamma = 0.09$) der Oberflächen-abhängigen Potentiale gegenüber den distanzabhängigen Potentialen annähernd korrekt behandelt wird, ist in Abb. 29 gezeigt. Für den Komplex aus dem Protein

Elastase und dem Liganden Trifluoracetyl-Lys-Pro-Isopropylanilid (PDB-Code: 1ela, Auflösung 1.8 Å) ist dort die Kristallgeometrie des Liganden nach den Atomtypen coloriert. Eine unter Verwendung der 1:1-Gewichtung ($\gamma = 0.5$) am besten bewertete gedockte Lösung mit *rmsd* = 11.9 Å ist blau dargestellt; die rote Ligandanordnung (*rmsd* = 2.5 Å) ist dagegen die unter Verwendung von $\gamma = 0.09$ auf Rang 1 gefundene Lösung. Im Fall der stark von der nativen Struktur abweichenden Ligandkonfiguration (blau) bleibt die tiefe, hydrophobe, durch Methylgruppen von Thr221, Thr236 und Val224 gebildete Tasche Lösemittel-zugänglich. Zusätzlich weist auch die hydrophobe $CF_3$-Gruppe des Liganden ins Lösemittel. Beide Effekte werden nur durch die gesteigerte Gewichtung der SAS-abhängigen Potentiale korrekt behandelt; die „Bestrafung“ der dem Lösemittel zugewandten Molekülteile ergibt dann insgesamt eine schlechte Bewertung für die erzeugte Protein-Ligand-Anordnung.

Die bislang vorgestellten Ergebnisse sind unter Verwendung derjenigen Parameter für die Paar- sowie die Einteilchenpotentiale erhalten worden (s. a. Kap. 5.1 und 5.2), die für den Kalibrierungsdatensatz FlexX_DS1 die höchste Rate der Erkennung nativ-ähnlicher Ligandkonfigurationen ergaben. Die damit erhaltene Bewertungsfunktion nach Gl. 36 zeigte auch bei der (Kreuz-)Validierung an den Datensätzen FlexX_DS2 und DOCK_DS überzeugende Resultate. Hier nun sollen Ergebnisse für FlexX_DS1 gezeigt werden, die bei der Verwendung alternativer, ebenfalls in Kap. 4.2 vorgestellter Referenzzustände, Dämpfungsschemata sowie Intervallparameter erhalten wurden. Die Betonung liegt hierbei auf der Untersuchung der Paarpotentiale. Sie alleine bewerten - unter Verwendung der optimalen Parameter - im Falle von FlexX_DS1 in 73 % aller möglichen Fälle eine Ligandkonfiguration mit *rmsd* < 2.0 Å am besten.

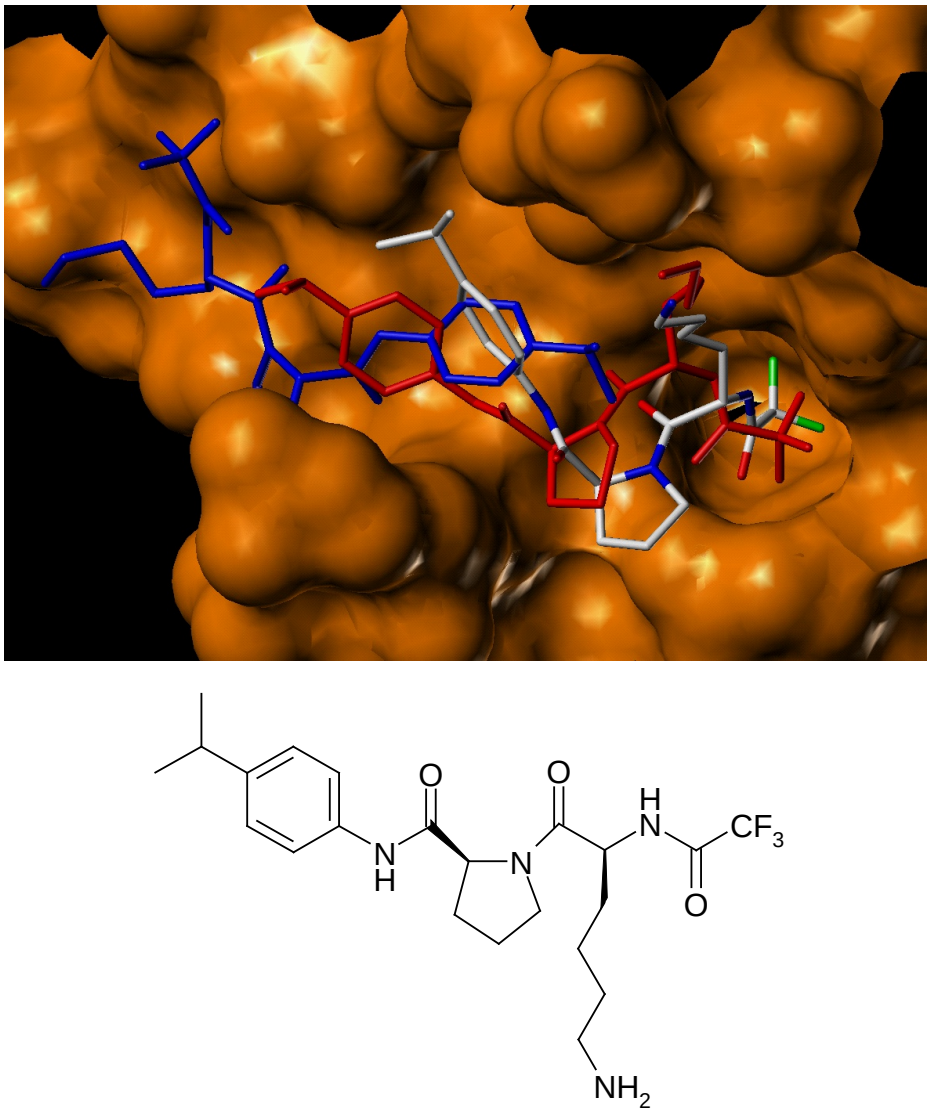



**Abb. 29:** Dargestellt ist die Connolly-Oberfläche (Connolly, 1983) der Bindetasche von Elastase im Komplex mit Trifluoracetyl-Lys-Pro-Isopropylanilid (s.a. Formel) (PDB-Code: 1ela). Die Ligandgeometrie der Kristallstruktur ist nach den Atomtypen farbcodiert, die unter Anwendung der 1:1-Gewichtung der distanzabhängigen Paar- sowie der Oberflächen-abhängigen Einteilchenpotentiale auf dem ersten Rang gefundene Ligandkonfiguration (*rmsd* = 11.9 Å) ist blau gefärbt, die unter Anwendung der 10-fachen Gewichtung der Einteilchenpotentiale gefundene bestbewertete Lösung (*rmsd* = 2.5 Å) ist rot dargestellt. Die tiefe Tasche auf der rechten Seite wird durch die Seitenkette von Val224 und die beiden Methylgruppen von Thr221 und Thr236 gebildet.

Verwendet man den in Gl. 23 aufgeführten und von Sippl (Sippl, 1990; Sippl, 1993) ursprünglich verwendeten Referenzzustand anstelle des in Gl. 24 benannten bei der Erzeugung der Paarpotentiale, so sinkt bei ansonsten gleichen Parametern die Rate der mit diesen Paarpotentialen auf Rang 1 bewerteten „guten“ Protein-Ligand-Anordnungen für FlexX_DS1 auf 54 %. Bei Verwendung der ausgezählten, absoluten radialen Auftrittshäufigkeiten anstelle der normierten wird der erhaltene Referenzzustand von zahlenmäßig in einem Distanzintervall

häufig auftretenden Kontakten zwischen zwei spezifischen Atomen dominiert. Als Folge weisen diese in den zur Ableitung verwendeten Komplexen häufig auftretenden Kontakte insgesamt nur eine geringe statistische Nettopräferenz auf (s. a. Kap. 4.2.2). Dies wird deutlich beim Vergleich der Paarpotentiale von C.3-C.3 bzw. O.co2-O.3. Die C.3-C.3-Verteilung ist dabei die mit Abstand am meisten populierte, wohingegen die O.co2-O.3-Verteilung nur knapp ein Zwanzigstel dieser Beobachtungen aufweist (Tab. 16, S. 103). Im ersteren Fall unterscheiden sich sodann auch die zum jeweiligen globalen Minimum gehörenden Potentialwerte um den Faktor drei, wobei der Betragswert für das unter Verwendung von Gl. 23 als Referenzzustand erhaltene Paarpotential erwartungsgemäß kleiner ist als der für das unter Verwendung von Gl. 24 erhaltene; im O.co2-O.3-Fall beträgt der Unterschied dagegen nur etwa ein Viertel.

Die hinter der Verwendung von Gl. 23 als Referenzzustand stehende Überlegung, implizit nicht nur Qualität, sondern auch Quantität der Kontakte bei den *vorherzusagenden* Protein-Ligand-Komplexen zu berücksichtigen („häufige Kontakte zeigen näher bei Null liegende Potentialwerte“), bedingt allerdings, daß die neu zu untersuchenden Fälle jeweils eine analoge Kontaktauftrittshäufigkeit aufweisen wie die, die im Mittel in der zur Ableitung verwendeten Datenbank gefunden wurden. Beachtet man nun, daß z. B. der Anteil polarer Atome bezogen auf alle Nichtwasserstoffatome für die Liganden im Datensatz FlexX_DS1 zwischen 5 und 80 % variiert (Mittelwert: 39 %, Standardabweichung: 17 %), so ist diese Bedingung für die vorliegende Datenauswahl sicher nicht erfüllt. In der zur Ableitung verwendeten Datenbank häufig auftretende, im jeweiligen untersuchten Komplex aber unterrepräsentierte Kontakte würden dann „unterschätzt“ werden und umgekehrt.

Die in Kap. 4.2.3 beschriebene Festlegung der oberen Abstandsgrenze auf 6 Å bedingt, daß ein Wassermolekül gerade nicht mehr als gegenseitiger Mediator von Wechselwirkungen *zwischen* einem Protein- und einem Ligandatom auftreten kann. Die bei der Verwendung alternativer oberer Schranken von 5 bzw. 8 Å mit den jeweiligen Paarpotentialen erhaltenen Erkennungsraten für „gute“ Ligandkonfigurationen bei FlexX_DS1 belaufen sich auf 61 % bzw. 65 % und liegen damit unter den mit $r_{max}$ = 6 Å erzielten Ergebnissen. Die Resultate für $r_{max}$ = 7 Å gleichen dagegen den mit 6 Å als oberer Schranke erhaltenen. Die Abnahme der Eignung der Potentiale zur Erkennung nativ-ähnlicher Ligandgeometrien bei Verwendung von $r_{max}$ = 8 Å zeigt, daß schon in diesem Fall spezifischen Wechselwirkungen zwischen Protein und Ligand ein zu geringes Gewicht eingeräumt wird. So liegt die mittlere Anzahl polarer Nachbarn eines polaren Atoms, bestimmt für alle Komplexe in FlexX_DS1, im Abstand bis zu

3.3 Å – der Obergrenze, bis zu der i.a. eine Wasserstoffbrückenbindung angenommen wird (Jeffrey, 1997) – bei eins. Für die Anzahl polarer Nachbarn um ein polares Atom bis 5.0, 6.0, 7.0 und 8.0 Å Abstand findet man dagegen die Mittelwerte 5, 9, 15 bzw. 24. Ein ähnliches Ergebnis wird von Muegge und Martin (Muegge & Martin, 1999) beschrieben, die zwar Paarpotentiale bis 12 Å aus kristallographisch bestimmten Protein-Ligand-Komplexen ableiten, allerdings bei der Berechnung von Bindungsaffinitäten die erhaltenen Potentiale nur bis zu Abständen von 6 Å (für Kohlenstoff-Kohlenstoff-Wechselwirkungen) bzw. 9 Å (für sonstige Wechselwirkungen) verwenden.

In Kap. 4.2.4 sind verschiedene Möglichkeiten zur Behandlung von Paarverteilungen bei geringer Beobachtungszahl aufgeführt worden. Die Verwendung der Reduzierung „lokaler Unsicherheit“ gemäß Gl. 26 (Absatz 2 in Kap. 4.2.4) mit $\chi = 10^{-4}$ erwies sich dabei als optimal. So ergaben sich bei Verwendung von $\chi = 0$, $10^{-3}$ bzw. $10^{-2}$ Erkennungsraten für „gut“ gedockte Protein-Ligand-Konfigurationen von 67 % , 68 % bzw. 65 %. Ein von Sippl (Sippl, 1990) ursprünglich vorgeschlagenes „Mischen“ einer normalisierten Verteilungsfunktion für ein spezifisches Atom-Atom-Paar mit der Paarverteilungsfunktion des Referenzzustandes in Abhängigkeit von der Anzahl der Beobachtungen für das betrachtete Paar (Gl. 25, Absatz 1 in Kap. 4.2.4) wurde für $\sigma$-Werte von $10^{-2}$, $2 \cdot 10^{-3}$ sowie $10^{-3}$ getestet. Diese Werte bedingen, daß bei Paarverteilungen mit 100, 500 bzw. 1000 Beobachtungen die Verteilungsfunktion des Referenzzustandes und der spezifischen Paarverteilungsfunktion den gleichen Anteil an der gemäß Gl. 25 entstehenden Verteilungsfunktion besitzen. Für $\sigma = 10^{-2}$ bzw. $2 \cdot 10^{-3}$ wurden für FlexX_DS1 mit 68 % schlechtere Erkennungsraten für „gut“ gedockte Protein-Ligand-Anordnungen festgestellt, für $\sigma = 10^{-3}$ betrug der Wert 65 %. Kombiniert man die in Absatz 1 und 2 in Kap. 4.2.4 aufgeführten Methoden zu Gl. 27 (Absatz 3 in Kap. 4.2.4) und verwendet als Parameter wiederum $\chi = 10^{-4}$ sowie $\sigma = 10^{-2}$, $2 \cdot 10^{-3}$ bzw. $10^{-3}$, so erhält man für die ersten beiden $\sigma$-Werte eine Erkennungsrate von 71 %, für $\sigma = 10^{-3}$ ergibt sich ein schlechteres Ergebnis (68 % richtig erkannter „gut“ gedockter Ligandkonfigurationen). Die in Absatz 4 in Kap. 4.2.4 angeführte lineare Variation der „Konstanten“ $\chi$ (Gl. 28) von einem Startwert $\chi_{Start}$ auf einen Endwert $\chi_{Ende}$ für Abstände der Atome $l$ und $p$ zwischen $VDW_{T(l)} + VDW_{T(p)} - dist$ und $VDW_{T(l)} + VDW_{T(p)}$ kann zu einer stärkeren Dämpfung lokaler Unsicherheiten in dem Bereich verwendet werden, wo besonders wenige Beobachtungen von Paarkontakten zu erwarten sind. Mit $dist$ = 1.0 Å, $\chi_{Start} = 10^{-4}$ und $\chi_{Ende} = 10^{-3}$, $10^{-2}$ sowie $10^{-1}$ wurde allerdings keine Verbesserung für den ersten $\chi_{Ende}$-Wert hinsichtlich der Erkennung nativ-ähnlicher Ligandkonfigurationen festgestellt (71 %), für $\chi_{Ende} = 10^{-2}$ bzw. $10^{-1}$ ergaben sich sogar nur Erken-

nungsraten von 67 % bzw. 59 %. Wird dagegen $\chi$ von einem Startwert $\chi_{Start}$ auf einen Endwert $\chi_{Ende}$ für Abstände der Atome *l* und *p* zwischen $VDW_{T(l)} + VDW_{T(p)} + dist$ und $r_{max}$ linear variiert (Gl. 29, Absatz 5 in 4.2.4), so können damit schneller gegen Null gehende statistische Nettopräferenzen für größere Abstände erzeugt werden. Für $dist = 1.0$ Å, $\chi_{Start} = 10^{-4}$ und $\chi_{Ende} = 10^{-3}$, $10^{-2}$ sowie $10^{-1}$ wurde mit den ersten beiden $\chi_{Ende}$-Werten allerdings keine Verbesserung der Erkennungsrate „guter" Ligandanordnungen bei FlexX_DS1 festgestellt (71 % bzw. 73 % Erkennungsrate), für $\chi_{Ende} = 10^{-1}$ ergab sich sogar eine Verschlechterung (63 %).

Insbesondere Paarverteilungen, bei denen ein Atomtyp F, Cl oder Br ist, sind gemäß Tab. 16 (S. 103) besonders schwach populiert. Um der Frage nachzugehen, ob die in ihnen enthaltene Information dennoch einen Einfluß auf die Erkennung nativ-ähnlicher Protein-Ligand-Anordnungen hat, wurden auch Paarpotentiale abgeleitet, die lediglich auf den 14 verbleibenden Atomtypen beruhen. Für den Datensatz FlexX_DS1 wurden bei Verwendung dieser statistischen Paarpräferenzen lediglich in 65 % aller möglichen Fälle eine „gut" gedockte Ligandgeometrie auf dem besten Bewertungsrang erkannt.

Um zu untersuchen, ob die mit der entwickelten Bewertungsfunktion erzielten Ergebnisse von molekularen Eigenschaften der untersuchten Liganden abhängen, wurde der *rmsd*-Wert der jeweils bestbewerteten Ligandkonfiguration für den Datensatz FlexX_DS1 gegen den Prozentsatz nichtpolarer Atome, den Prozentsatz von Wasserstoffbrückendonatoren und -akzeptoren (in beiden Fällen bezogen auf die Anzahl von Nichtwasserstoffatomen) sowie die Anzahl drehbarer Bindungen im Molekül aufgetragen. In allen Fällen konnte dabei keine signifikante Korrelation festgestellt werden; lediglich eine sehr schwache Abhängigkeit zeigte sich für die Anzahl drehbarer Bindungen. Diese gibt einen Hinweis auf die Komplexität des jeweiligen Docking-Problems. Da mit zunehmender Flexibilität des Liganden der zu durchmusternde Konformationsraum größer wird, erschwert dies die Vorhersage einer akzeptablen Bindungsgeometrie. Daher ist es nicht überraschend, daß für rigidere Liganden i.a. geringere *rmsd*-Werte gefunden werden.

### *5.5 Priorisierung von Liganden und Vorhersage von Bindungsaffinitäten*

Mit der hier entwickelten Bewertungsfunktion (Gl. 36) und unter Verwendung – sofern nicht anders vermerkt – der für die Erkennung von „guten" Ligandkonfigurationen für den Datensatz FlexX_DS1 gefundenen optimalen Parametern (s.a. Kap. 5.1 und 5.2) werden nun

für verschiedene Datensätze von Protein-Ligand-Komplexen mit bekannten Inhibitionskonstanten Bindungsaffinitäten vorhergesagt. Damit werden nicht nur die Gültigkeit der bei der Ableitung der Funktion angewendeten Annahmen getestet (s.a. Kap. 4.6.2), sondern auch, inwieweit für die Bindung eines Liganden an ein Protein zu berücksichtigende Beiträge implizit durch den verfolgten wissensbasierten Ansatz in den abgeleiteten statistischen Präferenzen enthalten sind.

Um zu untersuchen, inwiefern die zugrundegelegten Protein-Ligand-Strukturen einen Einfluß auf die erhaltenen Ergebnisse haben, werden sowohl kristallographisch bestimmte Komplexgeometrien als auch solche verwendet, bei denen der Ligand von FlexX in das jeweilige Protein gedockt wurde. Hierbei wird die erhaltene Bewertung mit Gl. 36 gegenüber dem experimentellen $pK_i$-Wert skaliert und der quadrierte Korrelationskoeffizient, die Standardabweichung sowie die maximale Abweichung bestimmt. Im letzten Kapitel dieses Abschnitts wird dann die Bewertungsfunktion an einem Fall getestet, der in seinem Umfang heutzutage durchgeführten virtuellen Screening-Ansätzen entspricht. Es geht dabei jeweils um die Erkennung „aktiver“ Liganden aus einer Menge gegebener Alternativen.

### 5.5.1 Bindungsaffinitätsvorhersage für Datensätze aus kristallographisch bestimmten Protein-Ligand-Komplexen

In Tab. 25 und Abb. 30 sind die Ergebnisse für die Vorhersage von Bindungsaffinitäten mit der hier entwickelten Bewertungsfunktion (Gl. 36) für Protein-Ligand-Komplexe aus der PDB mit experimentell bekannten Inhibitionskonstanten zusammengefaßt. Sofern nicht anders angegeben, wurde hierbei eine Gewichtung der Paar- zu den Einteilchenpotentialen von 1:1 ($\gamma = 0.5$ in Gl. 36) benutzt. Die Datensätze (s. a. Kap. 4.9.2) wurden dabei aus den Arbeiten von Eldridge *et al.* (Eldridge *et al.*, 1997), Head *et al.* (Head *et al.*, 1996) und Böhm (Böhm, 1994) entnommen, da sie bereits von Muegge und Martin (Muegge & Martin, 1999) zum Vergleich ihrer Bewertungsfunktion mit SCORE1 (Böhm, 1994) und SMoG-Score (DeWitte & Shaknovich, 1996) herangezogen wurden.

**Tab. 25:** Statistische Parameter der Korrelationen experimentell bestimmter und mit der hier entwickelten Bewertungsfunktion berechneter Bindungsaffinitäten. Die verwendeten Datensätze beruhen auf kristallographisch bestimmten Protein-Ligand-Komplexen.

| Datensatz | Anzahl Komplexe | $pK_i$-Bereich | $R^2$ | SD[a)] | MD[a)] | $\log(-c_s)$[b)] |
|---|---|---|---|---|---|---|
| **Serinproteasen** | 16 | 7 | 0.86 | 0.95 | 1.50 | -1.55 |
| **Metalloproteasen**[c)] | 15 | 10 | 0.70 | 1.53 | 3.32 | -1.49 |
| **Endothiapepsine** | 17 | 4 | 0.30 | 0.94 | 1.67 | -1.86 |
| **Arabinose-bindende Protine** | 18 (9)[d)] | 3 | 0.18 | 0.77 | 1.33 | -1.24 |
| **'Andere'**[e)] | 17 | 8 | 0.43 | 1.85 | 3.39 | -1.53 |
| **Böhm1998**[f)] | 71 | 13 | 0.33 | 2.21 | 7.22 | -1.48 |
| **Böhm1998(I)**[g)] | 49 | 10 | 0.44 | 1.79 | 4.52 | -1.44 |
| **Böhm1998(II)**[h)] | 46 | 10 | 0.56 | 1.53 | 4.36 | -1.44 |

a) Die Standardabweichung (SD) und die maximale Abweichung (MD) sind in logarithmischen Einheiten angegeben. b) Die Skalierungskonstante wird nach Gl. 42 berechnet. c) Verwendung von $\gamma = 0.09$ in Gl. 36. d) In den Kristallstrukturen von 9 Komplexen werden bedingt durch eine statische Unordnung jeweils beide Kohlenhydratepimere ($\alpha$- und $\beta$-Form) in der Bindetasche gefunden. Jedes dieser Epimere wird hier eigenständig behandelt. e) Dieser Datensatz umfaßt die Komplexe aus dem kombinierten Trainings- und Testdatensatz von Böhm (Böhm, 1994), die nicht in den vier vorherigen Sätzen auftreten und eine Auflösung $\leq 2.5$ Å besitzen. f) Nur Protein-Ligand-Komplexe, die auch in der PDB enthalten sind, werden von dem Trainings- und Testdatensatz von Böhm (Böhm, 1998) verwendet. g) Untermenge von Böhm1998, die nur die Komplexe mit einer Auflösung $\leq 2.5$ Å sowie $\leq 40$ Nichtwasserstoffatome für die Liganden enthält. h) Untermenge von Böhm1998(I), die nicht die Ausreißer 1cil, 1sbp und 1tnk enthält. Für nähere Erläuterungen siehe Text.

Der Testdatensatz „Serinproteasen" umfaßt 16 Trypsin- und Thrombin-Komplexe (Abb. 30 a). Für ihn wird die beste Korrelation aller hier untersuchten Datensätze mit $R^2 = 0.86$ erhalten. Allerdings werden Protein-Ligand-Komplexe mit experimentell bestimmter geringer Bindungsaffinität ($pK_i$-Werte < 2: 1bra, 1tni, 1tnj, 1tnk, 1tnl) um etwa 1.5 Größenordnungen zu gut vorhergesagt.

Der zweite Datensatz „Metalloproteasen" besteht aus 14 Carboxypeptidase A und Thermolysin-Komplexen sowie einem Collagenasekomplex (Abb. 30 b). Bei einem $R^2$-Wert von 0.70 wird die größte Abweichung mit + 3.3 Größenordnungen für den Komplex 6tmn beobachtet (für eine Erklärung dazu s. a. Kap. 5.6.1).

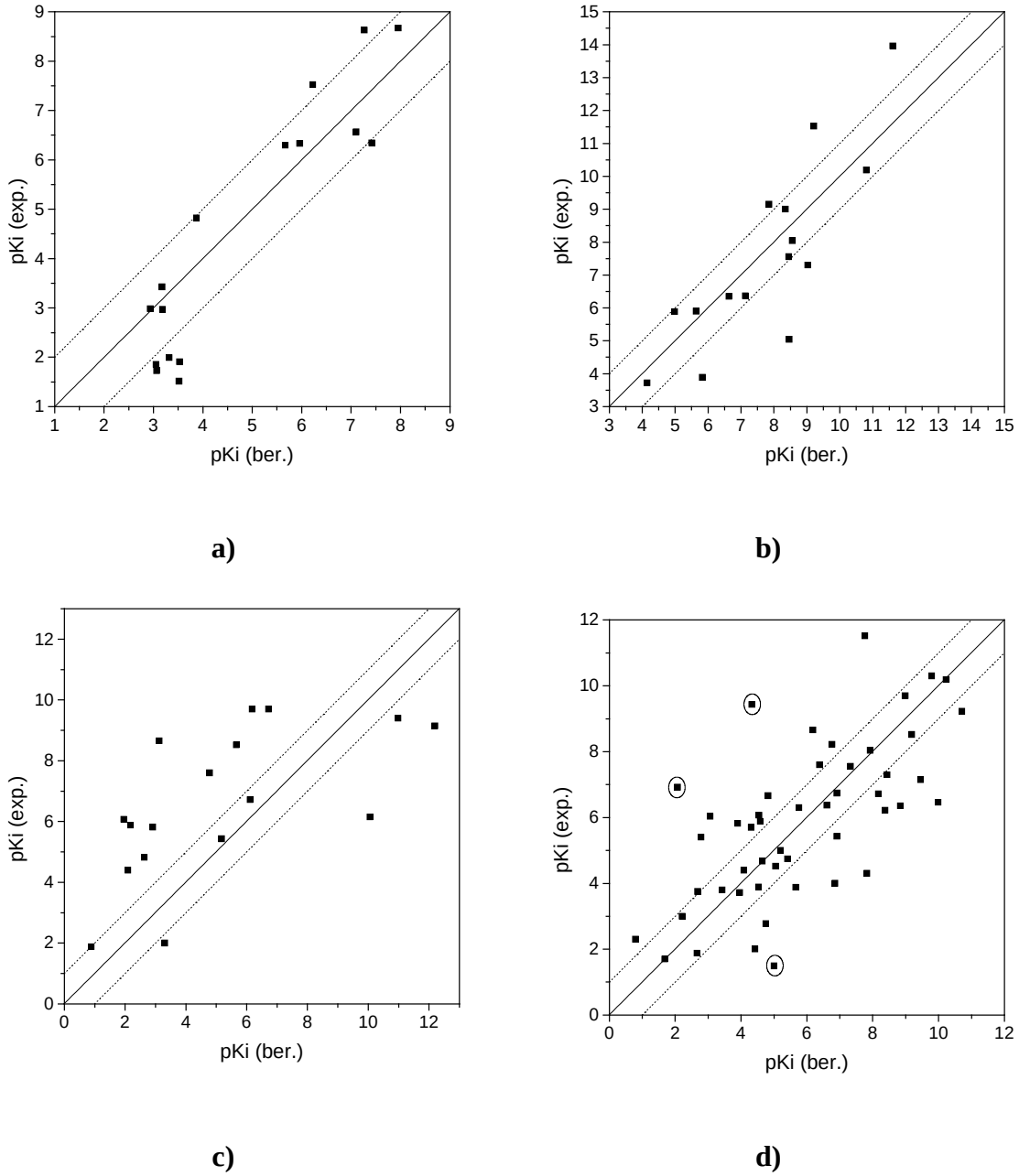



**Abb. 30:** Korrelationen experimentell bestimmter und nach Gl. 36 berechntere $pK_i$-Werte für vier Datensätze kristallographisch bestimmter Protein-Ligand-Komplexe, wie sie im Text beschrieben wurden: Serinproteasen (a), Metalloproteasen (b), 'Andere' (c), und Böhm1998(I / II) (d). Im letzteren Fall sind die drei Ausreißer 1cil, 1sbp und 1tnk gesondert gekennzeichnet. Die gestrichelt gezeichneten Linien geben Abweichungen von ± 1 logarithmischen Einheit an.

Der dritte Testdatensatz „Endothiapepsine" beinhaltet 11 Komplexe. Obwohl nur eine schwache Korrelation mit $R^2 = 0.30$ zwischen experimentellen und berechneten Bindungsaffinitäten gefunden wird, werden der am schwächsten bindende Ligand PD125754 (in 1eed) sowie der am stärksten bindende H-261 (in 2er7) im Vergleich zu den Komplexen mit mittle-

rer Bindungsaffinität ($pK_i$-Werte zwischen 6 und 8) korrekt identifiziert. Weiterhin ist zu beachten, daß die Standardabweichung zwischen experimentellen und berechneten Werten für diesen Datensatz nur 0.94 logarithmische Einheiten beträgt und daß die untersuchten Inhibitoren nicht nur insgesamt relativ groß (45 bis 79 Schweratome) sind, sondern sich in ihrer Größe auch deutlich voneinander unterscheiden.

Datensatz Nr. 4 (‚Arabinose-bindende Proteine') besteht aus 9 Komplexen. Da jeder dieser Komplexe jeweils zwei Kohlenhydratepimere als Liganden enthält, werden für beide Bindungsaffinitäten berechnet; die erhaltenen Ergebnisse für jedes Epimerpaar sind jedoch in allen Fällen nahezu identisch. Obwohl jeweils unterschiedliche Anteile von α- und β-Formen der jeweiligen Kohlenhydrate bedingt durch die Mutarotation während der experimentellen Bestimmung der Bindungsaffinität anwesend waren, hat dieses daher keinen Einfluß auf die bei der Berechnung erhaltenen Ergebnisse. Die für diesen Satz erhaltene Korrelation ist mit $R^2 = 0.18$ die schlechteste, wobei auch hier die Standardabweichung nur 0.77 logarithmische Einheiten beträgt. Dieses Ergebnis (kleiner $R^2$-Wert bei moderater Standardabweichung) wird bedingt durch den Bereich der experimentellen $pK_i$-Werte, der mit 3 Größenordnungen der kleinste aller hier untersuchten Datensätze ist. Eine Verwendung der kristallographisch bestimmten Wassermoleküle 309, 310 und 311 als Teil des Proteins während der Affinitätsvorhersage bringt keine Verbesserung. Diese Wassermoleküle werden von Quiocho *et al.* als wesentlich für die beobachtete Spezifität des Proteins bzgl. L-Arabinose, D-Fucose und D-Galactose angesehen.

Datensatz Nr. 5 (‚Andere') umfaßt 17 Protein-Ligand-Komplexe mit verschiedenen Proteinen (Cytochrom P450, FKB-bindendes Protein, HIV-Protease, Renin, DHFR, Galactose-bindendes Protein, Thymidilat-Synthase, Retinol-bindendes Protein, Triosephosphat-Isomerase, Myoglobin und Concavalin) (Abb. 30 c). Unter Verwendung von Kofaktoren als Teil des Proteins während der Berechnung der Affinitäten wird ein quadrierter Korrelationskoeffizient von 0.43 erhalten. Drei Bindungsaffinitäten werden dabei als zu hoch vorhergesagt: für den Reninkomplex 1rne wie auch für die HIV-Protease-Komplexe 4hvp und 4phv. In diesen Fällen bestehen die Liganden aus 51, 54 und 46 Nichtwasserstoffatomen. Die entwikkelte Bewertungsfunktion wurde nur unter Verwendung von Protein-Ligand-Komplexen abgeleitet, die Liganden mit ≤ 50 Nichtwasserstoffatomen enthielten, was als obere Grenze für wirkstoffartige Moleküle angesehen wird (Lipinski *et al.*, 1997). Es ist daher möglich, daß Informationen über Liganden mit einer Größe nahe bei oder oberhalb dieser Grenze nur unvollständig in den erhaltenen statistischen Präferenzen enthalten sind.

Der letzte Datensatz (‚Böhm1998') besteht aus 71 kristallographisch bestimmten Komplexen der Trainings- und Testdatensätze von Böhm (Böhm, 1998). Von Böhm modellierte Komplexe wurden dagegen nicht verwendet. ‚Böhm1998' ergibt einen $R^2$-Wert von 0.33. Der Ausschluß aller Komplexe mit einer Auflösung größer als 2.5 Å oder Liganden mit mehr als 40 Nichtwasserstoffatomen (s. o.) resultiert in einer Korrelation mit $R^2 = 0.44$ (‚Böhm1998(I)', Abb. 30 d). Dabei können drei Ausreißer erklärt werden: in 1sbp ist der Ligand Sulfat und fällt damit unter die untere Grenze von 6 Nichtwasserstoffatomen, die für die Ableitung der statistischen Präferenzen verwendet wurde. In 1cil (Inhibitor ETS gebunden an Carboanhydrase II) bildet der Sulfonamid-Stickstoff des Liganden mit dem Zink-Kation des Proteins eine starke Wechselwirkung aus, da ersterer vermutlich deprotoniert vorliegt. Die Bewertungsfunktion nach Gl. 36 behandelt diese Wechselwirkung allerdings als Kontakt zwischen einem ungeladenen Amidstickstoff und Zink, so daß sie den dadurch bedingten Bindungsbeitrag unterschätzt. Für den Liganden des Komplexes 1tnk kann eine stark deformierte Konformation festgestellt werden, wobei intramolekulare Bindungslängen und -winkel von den Standardwerten deutlich abweichen. Dies kann auf eventuelle Probleme im abschließenden Schritt der Verfeinerung der Kristallstruktur hindeuten. Entfernt man diese drei Ausreißer, ergibt sich für ‚Böhm1998(II)' eine deutlich verbesserte Korrelation mit $R^2 = 0.56$ und einer Standardabweichung von 1.53 logarithmischen Einheiten.

Anstelle der hier mit Gl. 36 berechneten und unmittelbar mit den experimentellen $pK_i$-Werten korrelierten (Gl. 42) Bewertungen verwenden DeWitte und Shakhnovich (DeWitte & Shaknovich, 1996) die auf die Anzahl der Nichtwasserstoffatome des Liganden bezogenen Werte, um so eine Abhängigkeit ihrer „Design Energien" von der Größe des Liganden zu korrigieren. Eine analoge Vorgehensweise erbrachte mit den hier berechneten Bewertungen jedoch keinerlei Verbesserung der Korrelation zwischen vorhergesagten und experimentellen Bindungsaffinitäten. Das gleiche Ergebnis wurde auch bei Verwendung der Anzahlen drehbarer Bindungen anstelle der Anzahl von Nichtwasserstoffatomen erhalten.

Die auf den distanzabhängigen Paarpotentialen beruhenden Beiträge zur Bewertung sind gemäß Kap. 4.2.1 bis auf eine vom jeweiligen betrachteten Komplex abhängige Konstante bestimmt. Diese Konstante ist daher beim Vergleich der Bewertungen *verschiedener* Komplexe von Bedeutung. Nur wenn sie in diesen Fällen eine jeweils ähnliche Größe annimmt, können die auf Gl. 36 beruhenden Bewertungen miteinander verglichen werden. Betrachtet man Korrelationen dann als signifikant, wenn ihr $R^2$-Wert größer als 0.3 ist (Cramer III *et al.*,

1993), so scheint im Fall der ‚Serinproteasen'-, ‚Metalloproteasen'-, ‚Andere'-, und ‚Böhm1998'-Datensätze (Tab. 25) die Voraussetzung der Ähnlichkeit der erwähnten Konstanten erfüllt. Für diese Testdatensätzen sind die erhaltenen $c_S$-Werte zudem untereinander relativ ähnlich (Mittelwert: $-3.08 \cdot 10^{-2}$, Standardabweichung: $2.32 \cdot 10^{-3}$ entsprechend 7.6 % bezogen auf den Mittelwert), so daß auch hier eine annähernd gleiche Größe für die unbestimmte Konstante angenommen werden kann. Bei dem mit $-1.38 \cdot 10^{-2}$ kleinsten $c_S$-Wert im Fall der Endothiapepsine besteht ein möglicher Zusammenhang zur Größe der in diesem Datensatz betrachteten Liganden (s.o.). Im Fall der ‚Arabinose-bindenden Proteine' wird die Interpretation des erhaltenen Wertes dagegen durch die in diesem Fall vorliegende schlechte Korrelation erschwert. Unter Beachtung, daß eine Bewertungsfunktion im Rahmen des strukturbasierten Designs v.a. zur *Priorisierung* verschiedener Liganden gegenüber *einem* Protein und weniger zur unmittelbaren Affinitätsvorhersage verschiedener Liganden gegenüber *verschiedenen* Proteinen verwendet wird, erscheint die Annahme der Gleichheit der unbestimmten Konstanten für Datensätze des ersteren Typs daher gerechtfertigt.

Die hier zur Vorhersage experimentell bestimmter Bindungsaffinitäten verwendete Bewertungsfunktion entspricht derjenigen, die auch für die Erkennung nativ-ähnlicher Protein-Ligand-Anordnungen verwendet wurde und deren Parameter an FlexX_DS1 kalibriert wurden. Nur für den Datensatz ‚Metalloproteasen' wurde hierbei der $\gamma$-Wert auf 0.09 gesetzt. Die bei Verwendung des ursprünglichen Wertes von $\gamma = 0.5$ erhaltenen statistischen Parameter ($R^2 = 0.64$, $SD = 1.69$) weichen allerdings nur wenig von den in Tab. 25 für diesen Datensatz aufgeführten ab. Da für den hier vorliegenden Zweck der Affinitätsvorhersage keinerlei zusätzliche Parameteranpassungen mehr durchgeführt wurden, können die erhaltenen Ergebnisse als „echte" Vorhersagen betrachtet werden.

Obwohl ein Bezug der erhaltenen Bewertungen auf die Anzahl der Nichtwasserstoffatome bzw. die der drehbaren Bindungen keine Verbesserung der Korrelationen erbrachten (s.o.), wirft doch der im Fall der Endothiapepsine nach unten abweichende $c_S$-Wert (s.o.) die Frage auf, ob zusätzliche Terme für die während der Ligandbindung erfolgende Einschränkung von Flexibilität und Mobilität verwendet werden müssen. Dies folgt auch in Anbetracht der zu hoch vorhergesagten Bindungsaffinitäten für 1rne, 4hvp und 4phv im Datensatz ‚Andere'. Die im Fall von Datensätzen mit Liganden ähnlicher Größe erhaltenen guten Korrelationen weisen ansonsten jedoch auf eine vollständige Berücksichtigung der bei der Protein-Ligand-Bindung auftretenden Beiträge hin.

### 5.5.2 Vergleich der erhaltenen Ergebnisse mit denen anderer Bewertungsfunktionen

Die mit der hier entwickelten Funktion für die ersten fünf Testdatensätze erhaltenen Ergebnisse bei der Vorhersage von Bindungsaffinitäten für Protein-Ligand-Komplexe (Tab. 25) werden mit denen von vier derzeit verwendeten Bewertungsfunktionen verglichen, von denen zwei ebenfalls wissensbasiert (PMFScore (Muegge & Martin, 1999), SMoGScore (DeWitte & Shaknovich, 1996)), die beiden anderen (SCORE1 (Böhm, 1994), ChemScore (Eldridge *et al.*, 1997)) jedoch regressionsbasiert sind (s.a. Kap. 3.2). Die Vergleichsdaten für PMFScore, SCORE1 und SMoGScore werden einer Zusammenstellung in (Muegge & Martin, 1999) entnommen; die Daten für ChemScore stammen von (Eldridge *et al.*, 1997).

Beim Vergleich verschiedener Verfahren auf Grundlage statistischer Deskriptoren ist zu beachten, daß der $R^2$-Wert stark von der jeweiligen Zusammensetzung des verwendeten Testdatensatzes abhängt. Dies gilt nicht für die Standardabweichung. Eindrückliche Beispiele hierfür ergeben sich mit den Datensätzen 3 und 4 (‚Endothiapepsine' und ‚Arabinose-bindende Proteine') aus dem vorherigen Kapitel. Obwohl im Fall der ‚Endothiapepsine' insgesamt ein $pK_i$-Bereich von 4 logarithmischen Einheiten abgedeckt wird, befinden sich 9 der 11 Komplexe in einem $pK_i$-Bereich zwischen 6 und 8, d.h. die Standardabweichung im Testdatensatz beträgt lediglich 1.1 logarithmische Einheiten. Mit 0.85 logarithmischen Einheiten noch geringer ist die Standardabweichung im Datensatz 4, bei dem sich 10 der 18 Protein-Ligand-Kombinationen in einem $pK_i$-Bereich von 7 bis 8 logarithmischen Einheiten befinden. Zieht man zusätzlich in Betracht, daß der Fehler für experimentell bestimmte Bindungsaffinitäten aus verschiedenen (Literatur-)Quellen einen Faktor 5 bis 10 in der Inhibitionskonstanten ausmachen kann (Böhm, 1998; Hosur *et al.*, 1994; Murray *et al.*, 1998), so liegen die Streuungen in den oben erwähnten Datensätzen bereits an dieser Grenze.

Aus diesem Grund werden im folgenden die Ergebnisse der verschiedenen Bewertungsfunktionen bei der Vorhersage von Bindungsaffinitäten im Hinblick auf die Standardabweichungen zwischen experimentellen und berechneten Daten verglichen (Abb. 31). In diesem Fall hier werden die dabei beobachteten Trends allerdings auch in den $R^2$-Werten widergespiegelt, da identische Datensätze als Grundlage des Vergleichs herangezogen werden.

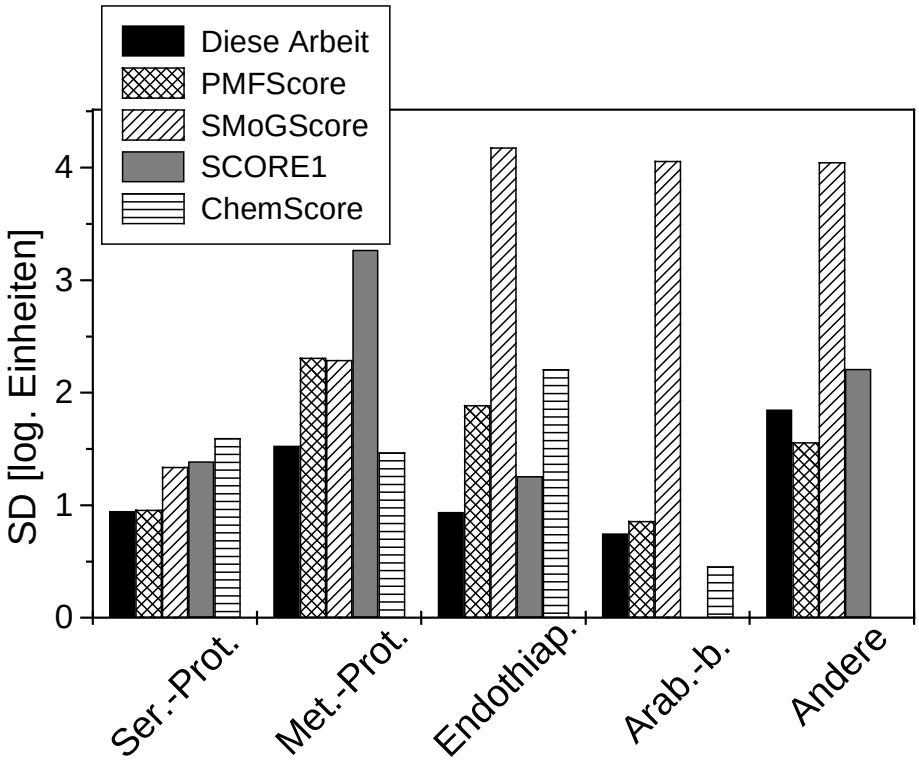



**Abb. 31:** Vergleich verschiedener Bewertungsfunktionen für Protein-Ligand-Wechselwirkungen im Hinblick auf die bei der Vorhersage von Bindungsaffinitäten erhaltenen Standardabweichungen zwischen experimentellen und berechneten Werten. Fünf in Kap. 5.5.1 beschriebene Testdatensätze (‚Serinproteasen', ‚Metalloproteasen', ‚Endothiapepsine', ‚Arabinose-bindende Proteine', ‚Andere') wurden untersucht. Die Angaben für PMFScore, SCORE1 und SMoGScore wurden (Muegge & Martin, 1999) entnommen, die für ChemScore stammen von (Eldridge *et al.*, 1997) (für den Datensatz ‚Andere' ist dort keine Angabe vorhanden). Im Fall der ‚Arabinose-bindenden Proteine' wird von Muegge und Martin (Muegge & Martin, 1999) ein Wert von 69.7 für SCORE1 berichtet.

Verglichen mit den anderen beiden wissensbasierten Ansätzen (PMFScore und SMoGScore) ergeben sich mit der hier entwickelten Funktion teilweise deutlich geringere Standardabweichungen für alle vier Datensätze mit einheitlichem Protein. Im Falle des ‚Andere'-Datensatzes weist PMFScore dagegen eine geringfügig kleinere Standardabweichung auf. Vergleicht man die regressionsbasierten Ansätze SCORE1 und ChemScore mit der hier entwickelten Methode, so ist letztere in allen Fällen besser als SCORE1, wohingegen ChemScore nur im Fall der ‚Arabinose-bindenden Proteine' eine merklich geringere Standardabweichung liefert. Unter der Beachtung, daß im Rahmen von *virtual screening*-Anwendungen die exakte Vorhersage verschiedener Liganden gegenüber *einem* biologischen

Zielmolekül wichtiger ist als die von Liganden gegenüber *verschiedenen* Proteinen, zeigt die hier entwickelte Bewertungsfunktion verglichen mit den vier anderen ein besseres Ergebnis.

### 5.5.3 Bindungsaffinitätsvorhersage für Datensätze gedockter Protein-Ligand-Strukturen

Wie das vorhergehende Kapitel gezeigt hat, liefert die entwickelte Bewertungsfunktion zufriedenstellende Ergebnisse bei der Bindungaffinitätsvorhersage basierend auf experimentell bestimmten Protein-Ligand-Anordnungen. Bei Anwendungen von virtuellen Screeningverfahren werden jedoch modellierte Ligandkonfigurationen zugrundegelegt. Dies gilt insbesondere bei der Verwendung von Docking-Methoden. Hier muß die Bewertungsfunktion zusätzlich die wahrscheinlichste Bindungsgeometrie des Liganden aus einer Menge gegebener Alternativen finden, sicherlich eine ungleich schwerere Aufgabe. Um die hier entwickelte Funktion auch für diese Anwendungen zu testen, wurden drei Datensätze verwendet, für die experimentell bestimmte Bindungsaffinitäten bekannt sind und bei denen die Protein-Ligand-Geometrien mit FlexX erzeugt wurden (s.a. Kap. 4.9.2). Der erste umfaßt 53 Komplexe aus FlexX_DS1 und FlexX_DS2, der zweite eine zusammenhängende Serie von 32 Inhibitoren von Thrombin bzw. Trypsin (Murray *et al.*, 1998) und der letzte eine Serie von 61 bzw. 15 Thermolysin-Inhibitoren (Klebe *et al.*, 1994) (s.a. Kap. 4.9.2). Eine Zusammenstellung der erhaltenen statistischen Parameter ist in Tab. 26 gegeben.

**Tab. 26:** Statistische Parameter der Korrelationen zwischen experimentell bestimmten Bindungsaffinitäten und mit der hier entwickelten Bewertungsfunktion berechneten für Datensätze mit von FlexX generierten Protein-Ligand-Anordnungen.

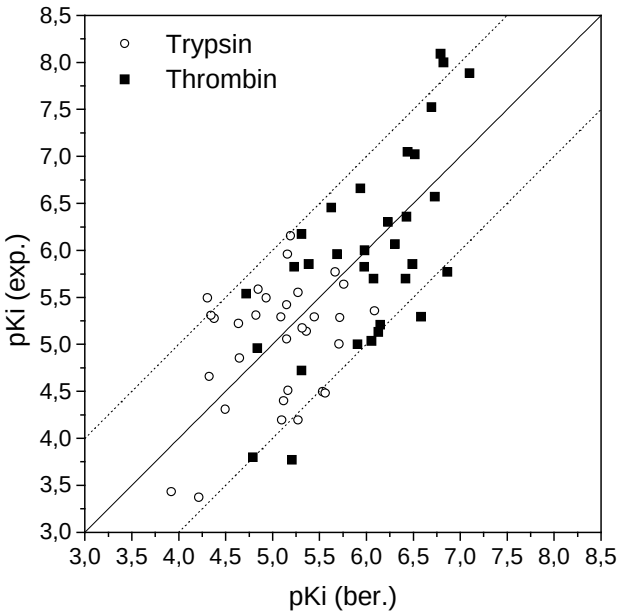
| Datensatz | Anzahl Komplexe | $pK_i$-Bereich | $R^2$ | SD[a)] | MD[a)] | $\log(-c_s)$[b)] |
|---|---|---|---|---|---|---|
| **'Gemischter Satz'[c)]** | 53 | 10 | 0.44 | 1.80 | 4.27 | -1.43 |
| **Thrombin- / Trypsin-Inhib.[d)]** | 64 | 5 | 0.48 | 0.71 | 1.25 | -1.59 |
| **Thermolysin[e)]** | 61 | 10 | 0.35 | 1.70 | 4.00 | -1.59 |
| **Thermolysin (I)[f)]** | 43 | 10 | 0.43 | 1.68 | 3.90 | -1.58 |
| **Thermolysin (II)[g)]** | 15 | 5 | 0.36 | 1.53 | 3.00 | -1.58 |
| **Thermolysin (III)[h)]** | 14 | 5 | 0.50 | 1.39 | 3.27 | -1.60 |

a) Die Standardabweichung (SD) und die maximale Abweichung (MD) sind in logarithmischen Einheiten angegeben. b) Die Skalierungskonstante wird nach Gl. 42 berechnet. c) Untermenge beider Validierungsdatensätze FlexX_DS1 und FlexX_DS2, für die eine Bindungsaffinität in der Literatur (Böhm, 1994; Böhm, 1998; Eldridge *et al.*, 1997; Head *et al.*, 1996) angegeben wurde. Bindungsaffinitäten wurden jeweils für die am besten von der hier entwickelten Funktion bewerteten und von FlexX generierten Protein-Ligand-Anordnung berechnet. d) Datensatz von verwandten Thrombin- und Trypsin-Inhibitoren aus der Arbeit von Obst *et al.* (Obst, 1997; Obst *et al.*, 1997). Die Protein-Ligand-Anordnungen wurden mit FlexX generiert, indem die Liganden in die Proteinstrukturen von Obst *et al*. (Obst *et al.*, 1997) bzw. 1pph gedockt wurden. e) 61 Thermolysininhibitoren aus dem Trainingsdatensatz in (Klebe *et al.*, 1994), für die Ligandgeometrien mit FlexX durch Docking in die Proteinstruktur von 1tlp erzeugt wurden. f) Untermenge des Datensatzes 'Thermolysin', die nur diejenigen Liganden umfaßt, für die überhaupt eine "vernünftige" (für Erläuterungen siehe Text) Geometrie von FlexX erzeugt wurde. g) 15 Thermolysin-Inhibitoren aus dem Testdatensatz in (Klebe *et al.*, 1994), für die Protein-Ligand-Konfigurationen wie in (e) beschrieben erzeugt wurden. h) Untermenge von Thermolysin(II), die nicht den Ausreißer ZGPLNH2 enthält.

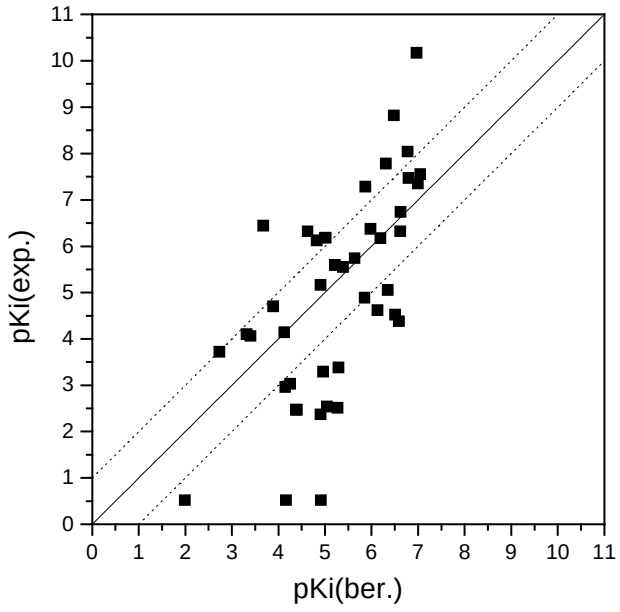
Für den ersten Datensatz (‚Gemischter Satz') wird eine Standardabweichung von 1.8 logarithmischen Einheiten und ein quadrierter Korrelationskoeffizient von 0.44 erhalten. Zum Vergleich beläuft sich der $R^2$-Wert bei der Bindungsaffinitätsvorhersage basierend auf den *Kristall*geometrien dieser Komplexe auf 0.34, d.h. die Affinitätsvorhersage ist in diesem Fall mit gedockten Ligandgeometrien sogar genauer als mit experimentell bestimmten Protein-Ligand-Anordnungen. Dies kann aber nicht verallgemeinert werden. Im Fall der Thrombin- und Trypsininhibitoren (Abb. 32 a) weichen die vorhergesagten Affinitäten von den experimentell bestimmten nur um 0.71 logarithmische Einheiten ab und der $R^2$-Wert beträgt 0.48. Dabei ist kein Unterschied in der Genauigkeit im Hinblick auf das zugrundegelegte Protein zu

bemerken. Für die 61 gedockten Thermolysininhibitoren (‚Thermolysin') ergibt sich zunächst ein $R^2$-Wert von 0.35. Beschränkt man den Satz auf solche Liganden, für die FlexX überhaupt eine Geometrie mit einem *rmsd*-Wert < 3 Å von mit Hand modellierten Referenzstrukturen (diese basieren auf Kristallgeometrien als Templaten) erzeugen kann, so steigt der $R^2$-Wert auf 0.43 (Standardabweichung: 1.68 logarithmische Einheiten) (‚Thermolysin (I)', Abb. 32 b). Für den getrennt betrachteten Satz von 15 Thermolysininhibitoren des Testsatzes aus (Klebe *et al.*, 1994) (‚Thermolysin(II)') ergibt sich $R^2 = 0.36$. Der Ausschluß des Ausreißers ZGPLNH2 liefert $R^2 = 0.50$ sowie eine Standardabweichung von 1.39 logarithmischen Einheiten (‚Thermolysin(III)').

Die für die Trypsin- und Thrombininhibitoren gefundene Standardabweichung von deutlich weniger als einer logarithmischen Einheit demonstriert die Eignung der hier entwickelten Bewertungsfunktion zur Vorhersage von Bindungsaffinitäten sogar auf Grundlage computergenerierter Protein-Ligand-Geometrien. Da die experimentell bestimmten Bindungsaffinitäten allerdings aus einer Quelle stammen, ist davon auszugehenden, daß ihr Fehler sogar noch geringer ist, d.h., daß in diesem Fall das experimentell vorgegebene Limit noch nicht erreicht wurde. Für die größere Abweichung von 1.5 bis 1.7 logarithmischen Einheiten im Fall der Thermolysininhibitoren muß dagegen beachtet werden, daß diese Daten aus unterschiedlichen Quellen stammen und mit unterschiedlichen Assaymethoden bestimmt wurden, so daß diese experimentellen Werte stärker fehlerbehaftet sind (Böhm, 1998; Hosur *et al.*, 1994; Murray *et al.*, 1998). Erschwerend kommt noch hinzu, daß durch die konformative Flexibilität der Liganden nicht in allen Fällen von FlexX überhaupt „vernünftige" Lösungen erzeugt wurden. Von der Bewertung unwahrscheinlicher Geometrien kann allerdings keine verläßliche Affinitätsvorhersage erwartet werden. Dies zeigt sich sehr deutlich für die erzielten Ergebnisse beim Übergang von ‚Thermolysin' zu ‚Thermolysin(I)', da im letzteren Fall nur noch „vernünftige" Ligandanordnungen in Betracht gezogen werden.




**a)**




**b)**

**Abb. 32:** Korrelationen zwischen experimentell bestimmten und mit der hier entwickelten Bewertungsfunktion berechneten Bindungsaffinitäten für einen Datensatz von Thrombin- und Trypsin-Inhibitoren (mit FlexX in die Proteinstrukturen von Thrombin (aus (Obst *et al.*, 1997)) und Trypsin (aus 1pph) gedockt) (a) sowie einen Datensatz von Thermolysininhibitoren (mit FlexX in die Proteinstruktur von 1tlp gedockt) (b). Im letzteren Fall sind nur die Daten dargestellt, die Liganden entsprechen, für die FlexX eine „vernünftige“ Ligandkonfiguration (für Erläuterungen siehe Text) gefunden hat (‚Thermolysin(I)‘). Die gestrichelt gezeichneten Linien geben eine Abweichung von ± 1 logarithmischen Einheit von dem experimentellen Wert an.

### 5.5.4 Virtuelles Screening

Eine mit der Priorisierung von verschiedenen Liganden gegenüber einem Protein direkt in Zusammenhang stehende Anwendung für die hier entwickelte Bewertungsfunktion ergibt sich im Rahmen des virtuellen Screenings von Substanzbibliotheken. Aus einer (großen) Menge zur Verfügung stehender Verbindungen sollen diejenigen herausgefunden werden, die eine Aktivität gegenüber dem untersuchten Rezeptor aufweisen („Aktiven"). Neben Verfahren, die auf der Ähnlichkeit von Molekülen zu schon bekannten Liganden des Zielmoleküls beruhen (Klebe, 1998a), wird hierbei im Rahmen der strukturbasierten Wirkstoffsuche insbesondere eine Kombination aus Docking-Verfahren zur Geometriegenerierung und anschließender Bewertung der erhaltenen Protein-Ligand-Anordnungen eingesetzt.

Für das hier untersuchte Beispiel wurden als „Aktive" 31 Verbindungen aus der Arbeit von Murray *et al.* (Murray *et al.*, 1998) verwendet (s.a. Kap. 4.9.3), die mit $pK_i$-Werten zwischen 2 und 7 schwache bis mittelstarke Thrombin- und Trypsin-Inhibitoren sind. Von diesen verfügen 10 über eine terminale Guanidinogruppe, weitere 11 besitzen eine primäre aliphatische Aminogruppe. Der Satz der „Inaktiven" besteht aus 824 aus dem ACD extrahierten Molekülen, die alle eine terminale Amidino- bzw. Guanidinofunktion aufweisen. Nach Docking dieser Liganden in die Rezeptorstrukturen von 1dwd (Thrombin) und 1pph (Trypsin) wurden die erhaltenen Rezeptor-Ligand-Anordnungen mit Gl. 36 bewertet und Anreicherungsfaktoren nach Gl. 43 berechnet. Diese Werte sowie zum Vergleich mit der von FlexX verwendeten Bewertungsfunktion berechnete sind in Tab. 27 aufgeführt. Der nach Gl. 44 berechnete maximale Anreicherungsfaktor beträgt 27.6. In beiden Fällen zeigt sich dabei ein deutlich höherer Anreicherungsfaktor bei der Bewertung mit der hier entwickelten Funktion im Vergleich zur FlexX-Bewertung. Nach Untersuchung der bestbewerteten 10% aller Moleküle würden so im Thrombinfall die 5-fache Anzahl „Aktiver", im Trypsinfall sogar die 23-fache Anzahl bei einer Bewertung mit Gl. 36 verglichen mit FlexX gefunden werden. Dieses wird in Abb. 33 deutlich: während die mit der hier entwickelten Bewertungsfunktion erhaltenen Anteile an „Aktiven" sich in beiden Fällen deutlich über der bei einer Zufallsauswahl zu erwartenden Anreicherung befinden, liegen die mit der FlexX-Bewertung erhaltenen Ergebnisse nur geringfügig darüber (Thrombinfall) bzw. sogar darunter (Trypsinfall). Alle „Aktiven" sind im Fall der Bewertung mit Gl. 36 bereits nach dem Durchmustern von 20 % (Thrombin) bzw. 30 % (Trypsin) der Gesamtdatenbank gefunden, wohingegen mit der FlexX-Bewertung 75 % bzw. 90 % untersucht werden müssen. Dies stellt eine signifikante Verbesserung dar.

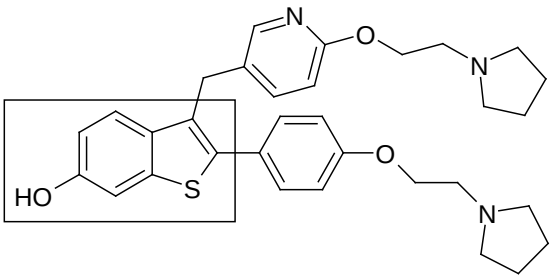
**Tab. 27:** Nach Gl. 43 berechnete Anreicherungsfaktoren für 31 „aktive“ (Murray *et al.*, 1998) und 824 „inaktive“ Verbindungen, die mit FlexX in die Proteinstrukturen von Thrombin (1dwd) und Trypsin (1pph) gedockt wurden.

| **Datensatz** | **Bewertung** | **Anteil der untersuchten Moleküle insgesamt** | | |
|---|---|---|---|---|
| | | **2 %** | **5 %** | **10 %** |
| **Thrombin** | **Gl. 36[a)]** | 4.8 | 8.3 | 6.5 |
| | **FlexX[b)]** | 1.6 | 1.3 | 1.3 |
| **Trypsin** | **Gl. 36[a)]** | 11.2 | 11.6 | 7.4 |
| | **FlexX[b)]** | 0 | 0 | 0.3 |

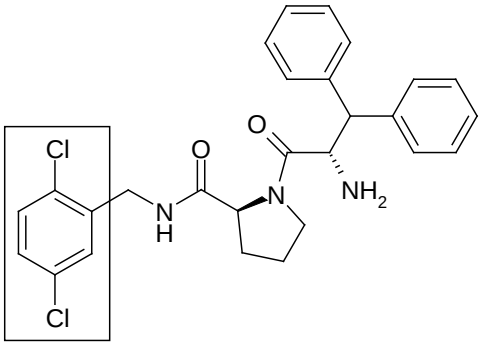
a) Die Bewertung der erhaltenen Protein-Ligand-Anordnungen erfolgte jeweils mit der hier entwikkelten Bewertungsfunktion. b) Die Bewertung der erhaltenen Protein-Ligand-Anordnungen erfolgte jeweils mit der von FlexX verwendeten Bewertungsfunktion, die auf Arbeiten von Böhm (Böhm, 1994) zurückgeht.

Bei der Untersuchung der Eignung einer Bewertungsfunktion zur Identifikation von „Aktiven“ in einer Menge von „Inaktiven“ im Rahmen eines *virtual screening*-Ansatzes sind die Eigenschaften der hierzu ausgewählten Datensätze von maßgeblicher Bedeutung. Während i.a. die Identifikation von Liganden mit Bindungsaffinitäten im nanomolaren Bereich oder gar darüber ($pK_i > 9$) in einem Satz von zufällig ausgewählten Verbindungen ohne Schwierigkeiten gelingt (Murray *et al.*, 1998), entspricht dieses Vorgehen nicht den in einem tatsächlichen Fall vorliegenden Gegebenheiten. Hierbei werden als Leitstrukturen auch solche Moleküle angesehen, die nur mit einer Affinität von 1 bis 10 µM ($pK_i = 5$ bis 6) an die biologische Zielstruktur binden (Teague *et al.*, 1999). Im Anbetracht dieser Tatsache entspricht die Wahl des Inhibitordatensatzes aus Murray *et al.* (Murray *et al.*, 1998) mit $pK_i$-Werten zwischen 2 und 7 den realen Erfordernissen. Zusätzlich wurden im ACD für den Satz der „Inaktiven“ nur solche Verbindungen gesucht, die auch eine nicht-zyklische Amidino- oder Guanidino-Funktion aufweisen. Im Fall der Serinproteasen Thrombin und Trypsin ergibt sich nämlich eine Randbedingung für die Bindung eines Liganden durch die Erfordernisse der Spezifitätstasche S1: Liganden müssen diese Tasche nicht nur optimal ausfüllen, sondern mit den entsprechenden Seitenketten (Asp189, Gly216, Gly219) auch Wechselwirkungen eingehen können. Während bei Substraten die basischen Aminosäuren Arginin und Lysin diese Bedingungen erfüllen, werden Inhibitoren häufig mit (Benz-)Amidino- oder Guanidinofunktionen für diesen Zweck versehen. Somit erfüllt der Satz der „Inaktiven“ zumindest diese strukturelle Voraussetzung zur Proteinbindung. Allerdings ist die Anwesenheit einer solchen basischen Gruppe keines-

wegs notwendig, wie potente Inhibitoren von Ely Lilly (**6**) sowie Merck (**7**) mit lipophilen S1-Substituenten (in den Formeln eingerahmt) zeigen.




**6**




**7**

Stahl *et al*. (Stahl *et al*., 2000) berichten bei der Verwendung von FlexX im Zusammenhang mit der Durchmusterung von Substanzbibliotheken nach Thrombininhibitoren, daß alle von dem Programm gut bewerteten Moleküle über eine an das Asp189 bindende Benzamidino-Funktion verfügen. Fehlt diese spezifische Wechselwirkung allerdings oder wird der entsprechende Term in der von FlexX verwendeten Bewertungsfunktion (Böhm, 1994) auf Null gesetzt, kann keine signifikante Anreicherung der „Aktiven“ mehr festgestellt werden. Ein Grund für die Dominanz dieser Wechselwirkung hängt damit zusammen, wie anhand einer multiplen linearen Regression eine Bewertungsfunktion kalibriert wird. Bei unausgewogener Zusammensetzung des Trainingsdatensatzes können so einzelne Terme über- bzw. unterbewertet werden. In Anbetracht dieser Befunde und unter Beachtung, daß nur etwa ein Drittel der verwendeten „Aktiven“ über eine terminale Guanidinofunktion verfügen (wohingegen alle „Inaktiven“ eine solche oder aber eine Amidinofunktion besitzen), läßt sich so das unbefriedigende Abschneiden der Bewertungsfunktion in FlexX erklären. Wie oben aufgeführte Ergebnisse zeigen, ist alleine schon aus methodischen Gründen eine „Abhängigkeit“ wegen der Überbetonung spezifischer Wechselwirkungen bei der Verwendung der in dieser Arbeit entwickelten Bewertungsfunktion nicht gegeben. Bei dem verfolgten wissensbasierten Prozeß beruhen die Potentialwerte auch nicht explizit auf einem Trainingssatz.

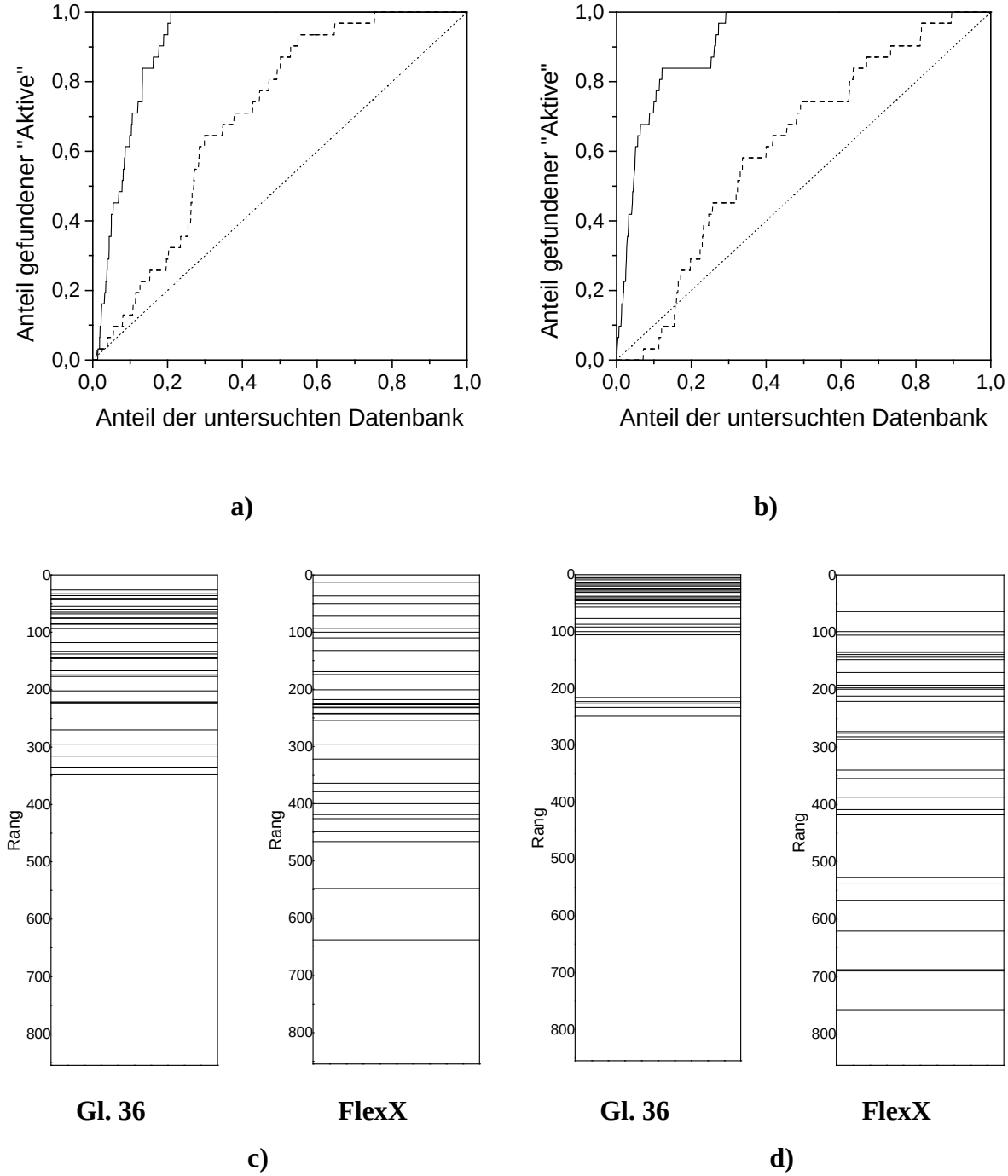
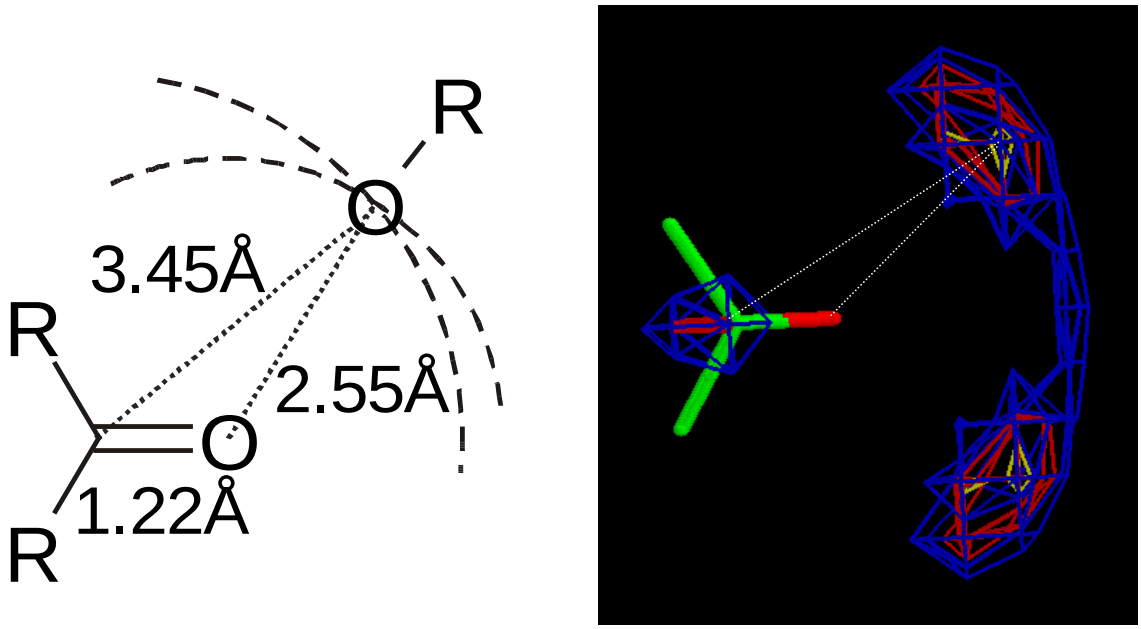



**Abb. 33:** Anteil gefundener „aktiver“ Moleküle als Funktion des Anteils der insgesamt (31 „Aktive“ + 824 „Inaktive“) untersuchten Datenbank für mit FlexX in Thrombin (a) bzw. Trypsin (b) gedockte Liganden, die anschließend mit dem hier entwickelten Ansatz (durchgezogene Linie) bzw. mit der FlexX-Bewertungsfunktion (gestrichelte Linie) bewertet wurden. Zusätzlich mit angegeben ist der bei einer Zufallsauswahl zu erwartende Anteil an „Aktiven“ (gepunktetet Linie). Außerdem ist der Rang der „Aktiven“ dargestellt, wie er nach dem Docking der gesamten Datenbank („Aktive“ + „Inaktive“) mit FlexX in Thrombin (c) und Trypsin (d) und Bewertung der erhaltenen Protein-Ligand-Anordnungen mit dem hier entwickelten Ansatz bzw. der FlexX-Funktion erhalten wurde.

## *5.6 Visualisierung von Wechselwirkungsfeldern und Untersuchung der impliziten Berücksichtigung von Direktionalität in Paar-Potentialen*

Die folgenden Betrachtungen beschränken sich auf die in dieser Arbeit abgeleiteten, distanzabhängigen statistischen *Paar*präferenzen. Jedes dieser Paarpotentiale *alleine* weist einen kugelsymmetrischen Verlauf auf, der keinerlei Winkelabhängigkeit der Wechselwirkung zwischen den betrachteten Atomen beschreibt. Hier soll daher untersucht werden, inwiefern durch die lokale Überlagerung *mehrerer* Paarpotentiale an einem Ort in der Bindetasche eines Proteins die Direktionalität von Wechselwirkungen zwischen (Paaren von) Atomen dennoch implizit beschrieben wird.

Dieses Zusammenwirken mehrerer Paarpotentiale soll am Beispiel der Wasserstoffbrükkenbindung zwischen einer Carbonylgruppe und einem Sauerstoffatom des Typs O.3 deutlich gemacht werden (Abb. 34). Die günstigste Wechselwirkung für ein O.2-O.3-Atompaar tritt hierbei in der Paarpräferenz im gegenseitigen Abstand von 2.55 Å auf, die für ein C.2-O.3-Paar im Abstand von 3.45 Å. Nimmt man zusätzlich für die C.2-O.2-Bindungslänge einen Wert von 1.22 Å an, so ergibt sich bei den vorgegebenen Distanzen ein C.2-O.2-O.3-Winkel für die günstigste Anordnung beider Atomgruppen von 128° (Abb. 34 a). Dieser Wert liegt nicht nur in dem erwarteten Bereich (s. a. Kap. 2.2.1), sondern eine analoge Vorzugsgeometrie kann auch in der Datenbank IsoStar (Bruno *et al.*, 1997) für nichtbindende Wechselwirkungen zwischen einer Carbonylgruppe und Sauerstoffatomen vom Hydroxyltyp gefunden werden (Abb. 34 b). Obwohl durch Verwendung von nur zwei Paarpotentialen für dieses Beispiel die günstigste räumliche Anordnung beider Gruppen bislang nur auf einen senkrecht zur Carbonylgruppenebene stehenden Kreisbogen beschränkt ist, kann angenommen werden, daß weitere Wechselwirkungen zu benachbarten Atomen diese Wechselwirkungsgeometrie zusätzlich beschränken werden.




**a)** **b)**

**Abb. 34:** Von Atom-Atom-Paarpräferenzen des Typs O.2-O.3 bzw. C.2-O.3 implizit bedingte Eingrenzung der geometrischen Anordnung zwischen einer Carbonylgruppe und einem Sauerstoffatom des Typs O.3 (a). Unter Verwendung der jeweils günstigsten Abstände zwischen O.2-O.3 (2.55 Å) und C.2-O.3 (3.45 Å) sowie der C.2-O.2-Bindungslänge von 1.22 Å kann ein C.2-O.2-O.3-Winkel von 128° berechnet werden. Zum Vergleich sind Wahrscheinlichkeitsdichten für die Anordnung von Hydroxylgruppen-Sauerstoffatomen um eine Carbonylgruppe angegeben, wie sie in der Datenbank IsoStar gefunden werden (b). Die blaue Kontur umfaßt die Bereiche, in denen die Wahrscheinlichkeitsdichte der Hydroxylgruppen mind. 20 % des für diese Wechselwirkung festgestellten Maximums entspricht, die rote umfaßt Bereiche mit mind. 40 % Wahrscheinlichkeitsdichte und die gelbe Bereiche mit mind. 80 % Wahrscheinlichkeitsdichte.

### 5.6.1 Visualisierung von „ausgezeichneten Punkten“ (*hot spots*) in Proteinbindetaschen

An den Gitterpunkten eines in der Bindetasche von Proteinen erzeugten kubischen Gitters werden für die in Tab. 2 (S. 65) angegebenen Ligandatomtypen mit den statistischen Paarpräferenzen die jeweiligen Bewertungen in *Abwesenheit* des nachfolgend mit dargestellten Liganden berechnet (Gl. 45). Anschließend werden die Ergebnisse für jeden Ligandatomtyp einzeln konturiert. Für das Arabinose-bindende Protein sowie Phospholipase A2 sind die dabei erhaltenen Ergebnisse in Abb. 35 dargestellt.

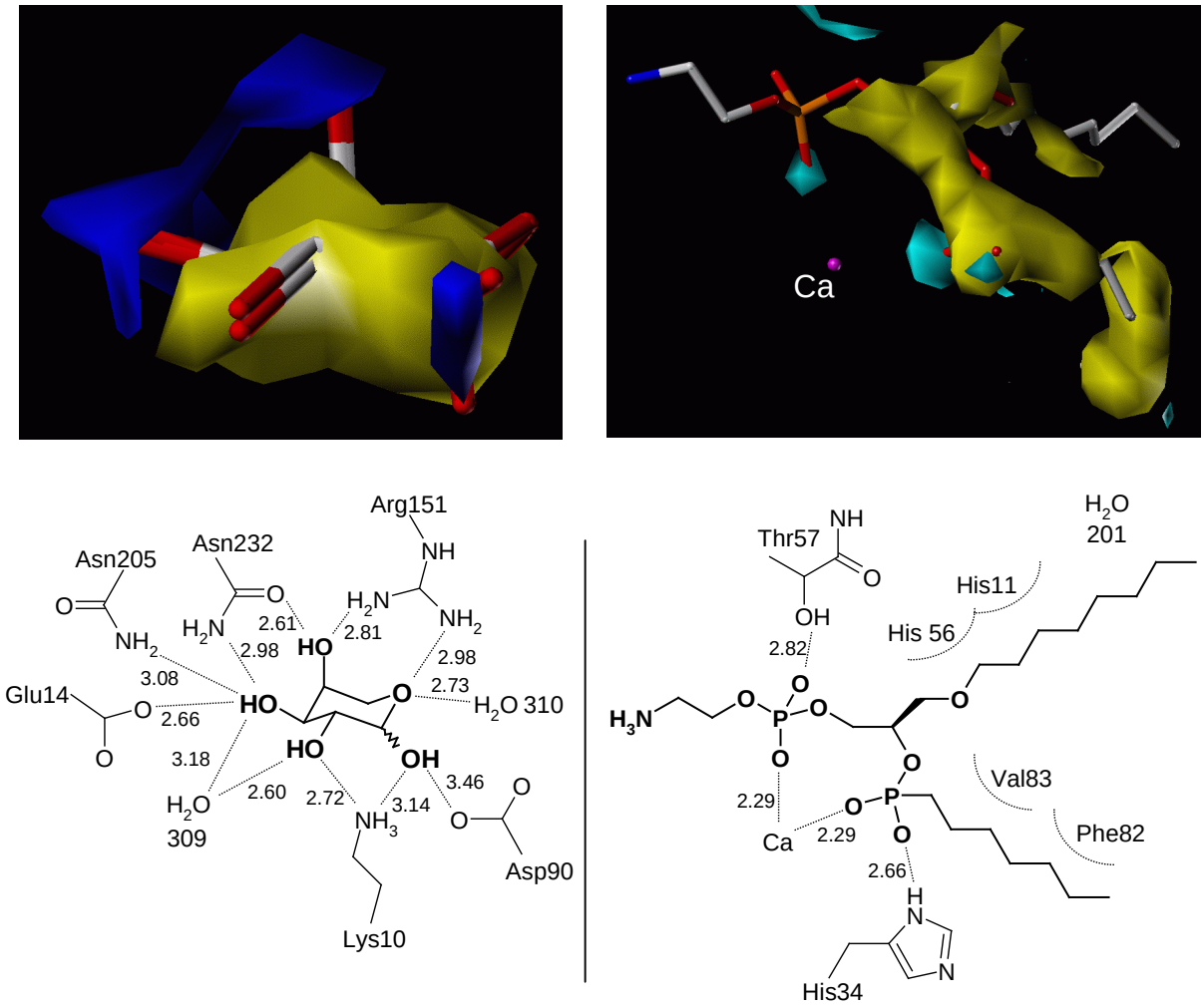



**Abb. 35:** Isokonturflächen, die mit der hier entwickelten Funktion in der Bindetasche von Proteinen berechnete Bewertungen für einzelne Ligandatomtypen einschließen (oben). Die Bewertungen wurden an den Gitterpunkten eines kubischen Gitters der Weite 0.5 Å ermittelt. Die eingeschlossenen Bewertungen liegen jeweils bis zu 10 % über dem globalen Minimum für einen Atomtyp. Im Fall des Arabinose-bindenden Proteins (PDB-Code: 1abe) (a) sind zusätzlich die in der Kristallstruktur enthaltenen Epimere von L-Arabinose dargestellt; im Fall von Phospholipase (PDB-Code: 1poc) (b) ist ein Übergangszustandsanalog als Ligand gezeigt. Dunkelblaue Konturen umfassen Werte, die für Ligandatome des Typs O.3 berechnet wurden, gelbe Werte für „C.3-Atome“ und hellblaue Werte für „O.2-Atome“. Im unteren Teil der Abbildung sind jeweils charakteristische Wechselwirkungen zwischen den Liganden und sie in der Bindetasche umgebenden Aminosäuren bzw. Wassermolekülen gezeigt. Die Längenangaben für gerichtete Wechselwirkungen erfolgen in Å; gestrichelte Kreisbögen zeigen hydrophobe Wechselwirkungen an.

Das Arabinose-bindende Protein nimmt beide Epimere der L-Arabinose (PDB-Code: 1abe) bevorzugt vor anderen Kohlenhydratliganden in seiner Bindetasche auf (Quiocho *et al.*, 1989). Die in Abb. 35 a dargestellte Isokonturoberfläche für C.3 (gelb) umfaßt alle Gitter-

punkte, an denen die Bewertung für diesen Ligandatomtyp bis zu 10 % über dem dafür ermittelten globalen Minimum in der Bindetasche liegt. Von ihr umschlossen werden die Positionen aller Ligandatome dieses Typs. Die O.3-Kontur (ebenfalls für die 10 %-Grenze; dunkelblau) zeigt drei ausgezeichnete Bereiche im untersuchten Raum; alle drei umschließen Hydroxylgruppen der Liganden in der Kristallstruktur. Interessanterweise erstreckt sich die Kontur in der Nähe des O-1-Sauerstoffs dabei über den Bereich, der von den Sauerstoffatomen der jeweiligen Epimere besetzt wird. An den Positionen, wo sich in der Struktur O-2 und O-5 befinden, werden bei Verwendung der 10 %-Grenze keine *hot spots* ausgewiesen. In beiden Fällen kann dieser Befund damit erklärt werden, daß diese Sauerstoffatome außer einer Wasserstoffbrücke zu den Seitenkettenstickstoffatomen von Lys10 (O-2) bzw. Arg151 (O-5) nur Wechselwirkungen mit umgebenden Lösemittelmolekülen ($H_2O$-309 bei O-2; $H_2O$-310 bei O-5) eingehen und daher keine günstigen Wechselwirkungen mit dem Protein bestimmt werden können.

Phospholipase A2 bindet im Komplex mit dem PDB-Code 1poc ein Übergangszustandsanalog. Werden in diesem Fall die Konturen für den Ligandatomtyp O.2 dargestellt (Abb. 35 b; 10 %-Grenze; hellblau), so stimmen ihre Positionen mit den im Kristall gefundenen Anordnungen von Sauerstoffatomen der Phosphat- bzw. Phosphonateinheiten sehr gut überein. Insbesondere die zum Calcium-Ion gewandten Positionen werden sehr gut repräsentiert. Große Teile der Kohlenwasserstoffketten liegen ebenfalls innerhalb der gefundenen günstigsten Aufenthaltsbereiche für Atome des Typs C.3 (10 %-Grenze; gelb). Das am oberen rechten Bildrand nicht umschlossene Ende einer dieser Ketten sowie die Aminoethylgruppe sind wiederum zum Lösemittel hin orientiert.

In Abb. 36 sind die Konturen für die Ligandatomtypen C.3, O.3 und N.am jeweils bei der 10 %-Grenze in der Bindetasche von Thermolysin dargestellt, zusammen mit dem Phosphoranalogen (ZGPLL) des Carbobenzoxy-Gly-Leu-Leu-Peptids (PDB-Code: 5tmn). Für dieses Phosphonamid wurde von Bartlett und Marlowe (Bartlett & Marlowe, 1987) eine Inhibitionskonstante von 9.1 nM berichtet, während das Phosphonat-Isomer (Substitution von P-NH durch P-O: ZGPOLL) etwa 990-fach schwächer an Thermolysin bindet. Im Gegensatz dazu wurde für das Phosphinat-Isomer (Substitution von P-NH durch P-$CH_2$: ZGPCLL) eine Inhibitionskonstante von 180 nM gefunden (Grobelny *et al.*, 1989), nur etwa 20-fach größer als für das Phosphonamid. Diese Unterschiede wurden mit (De-)Solvatationseffekten und der An- bzw. Abwesenheit einer Wasserstoffbrücke zwischen dem zum Phosphor benachbarten Atom und dem Carbonylsauerstoff der Aminosäure Ala113 im Thermolysin erklärt (Bash *et al.*,

1987; Grobelny *et al.*, 1989), da ansonsten die Bindungsmodi aller drei Inhibitoren nahezu identisch sind (Tronrud *et al.*, 1987). Bemerkenswerterweise wird auch von der in dieser Arbeit entwickelten Bewertungsfunktion in Nachbarschaft zum Ligandphosphoratom ein *hot spot* für Atome des Typs N.am bzw. C.3 gefunden, wohingegen ein Atom des Typs O.3 hier als nicht günstig angesehen wird. Für den letzteren Typ werden dagegen günstige Positionen in der Nähe der terminalen, zum Zink hin orientierten Sauerstoffatome der $-PO^{-}_{2}$-Gruppierung gefunden.

Die in Abb. 36 dargestellten Konturen für einen Ligandatomtyp geben lediglich eine *relative* Beschreibung der jeweiligen Wechselwirkungen wieder. Eine *absolute* Beschreibung der Verhältnisse erhält man jedoch, wenn der „Pro-Atom-Beitrag“ an der Gesamtbewertung der Protein-Ligand-Bindung für die Typen N.am, C.3 und O.3 am Ort des jeweiligen Atoms herangezogen wird. Die Beiträge für C.3 (-11.11) und N.am (-9.33) in Nachbarschaft zum Phosphoratom des Liganden weichen nur um 15 % voneinander ab; ein O.3-Atom an dieser Stelle trägt jedoch deutlich weniger zur Bindungsaffinität bei (-4.34). Damit resultieren für das Phosphonamid (ZGPLL) (5tmn) und das Phosphinat (ZGPCLL, aus 5tmn durch Austausch von N.am und C.3 erhalten) jeweils vergleichbare Bewertungen, wohingegen das Phosphonat (ZGPOLL) (6tmn) als am schwächsten bindender Ligand vorhergesagt wird. Obwohl mit der hier entwickelten Bewertungsfunktion der Unterschied zwischen ZGPOLL und seinen beiden Isomeren also *qualitativ* richtig vorhergesagt wird, ergibt sich bei *quantitativer* Betrachtung allerdings eine zu groß vorhergesagte Bindungsaffinität für ZGPOLL. Dies ist vermutlich auf die Zusammenfassung von Hydroxyl- sowie an zwei Nichtwasserstoffatome gebundenen Sauerstoffatomen unter den gemeinsamen Obertyp O.3 während der Ableitung der Potentiale zurückzuführen. Dementsprechend wird die eigentlich ungünstige Estersauerstoff / Carbonylsauerstoff-Wechselwirkung durch die eher günstigen Hydroxylsauerstoff / Carbonylsauerstoff-Wechselwirkungen im hier verwendeten O.3-O.2-Paarpotential überdeckt.

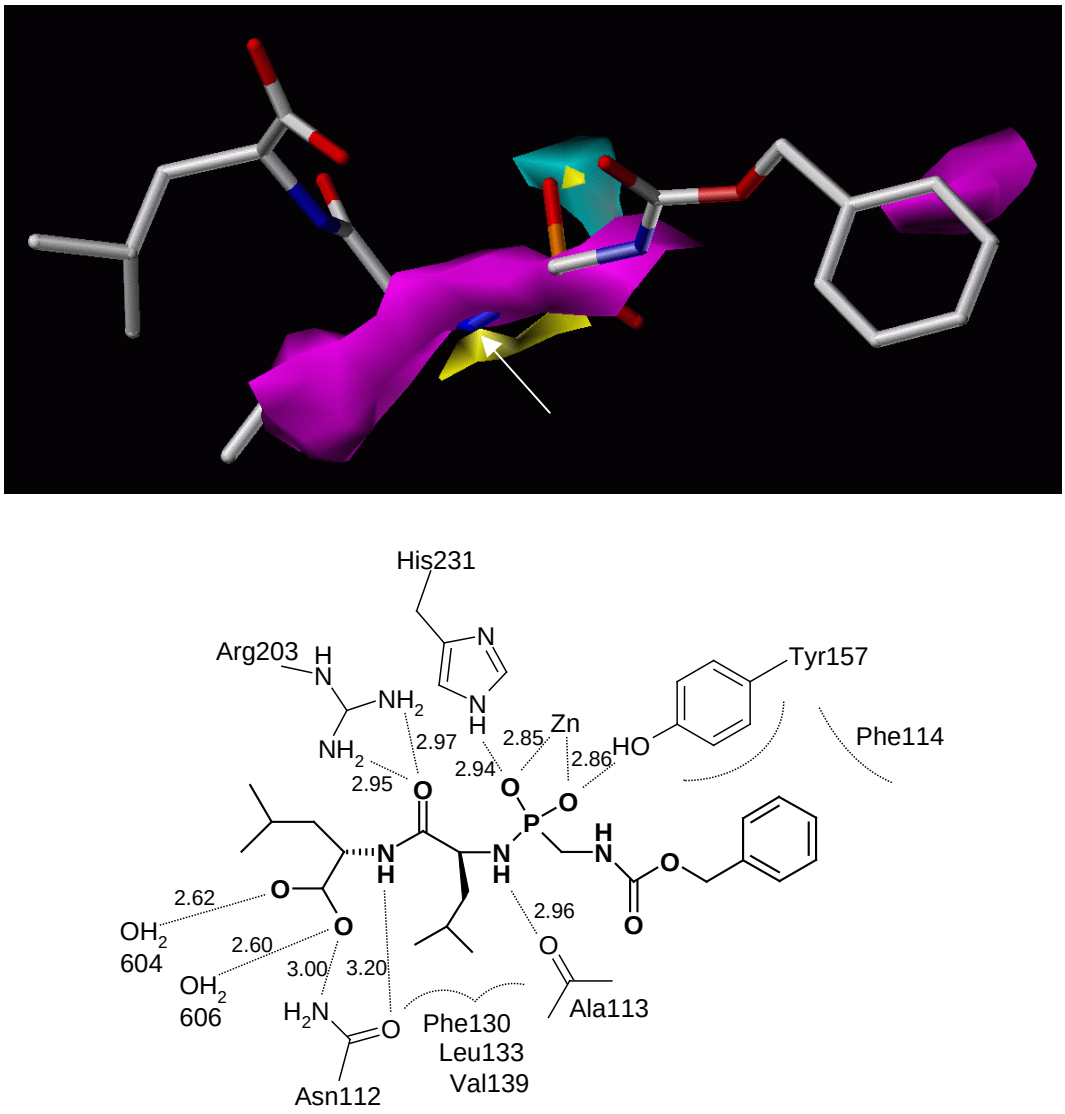



**Abb. 36:** Isokonturflächen, die mit der hier entwickelten Funktion in der Bindetasche von Thermolysin (PDB-Code: 5tmn) berechnete Bewertungen für einzelne Ligandatomtypen einschließen (oben). Die Bewertungen wurden an den Gitterpunkten eines kubischen Gitters der Weite 0.5 Å ermittelt. Die eingeschlossenen Bewertungen liegen jeweils bis zu 10 % über dem globalen Minimum für einen Atomtyp. Zusätzlich ist der Ligand ZGPLL mit angegeben. Violette Konturen umfassen Werte, die für Ligandatome des Typs C.3 berechnet wurden, gelbe Werte für „N.am-Atome“ und hellblaue Werte für „O.3-Atome“. Der Pfeil zeigt auf den Phosphonamid-Stickstoff, der in ZGPOLL durch Sauerstoff und in ZGPCLL durch Kohlenstoff ersetzt ist. Im unteren Teil der Abbildung sind charakteristische Wechselwirkungen zwischen dem Liganden und ihn in der Bindetasche umgebenden Aminosäuren bzw. Wassermolekülen gezeigt. Die Längenangaben für gerichtete Wechselwirkungen erfolgen in Å; gestrichelte Kreisbögen zeigen hydrophobe Wechselwirkungen an.

### 5.6.2 Quantitative Untersuchung der Übereinstimmung von *hot spots* mit in Kristallstrukturen tatsächlich gefundenen Atomtypen von Liganden

Für die Untersuchung, wie oft von der hier entwickelten Bewertungsfunktion vorhergesagte *hot spots* für Ligandatome in Proteinbindetaschen mit experimentell gefundenen Ligandatomanordnungen übereinstimmen, wurden die insgesamt 159 kristallographisch bestimmten Protein-Ligand-Komplexe der Datensätze FlexX_DS1 und FlexX_DS2 verwendet. Hierbei werden an den Gitterpunkten eines in der Bindetasche erzeugten kubischen Gitters für Sondenatome des Typs C.3, O.3, O.2, O.co2 und N.3 Bewertungen anhand der Paarpotentiale mit Gl. 15 bzw. 45 berechnet. Die Auswahl dieser Ligandatomtypen ergibt einen Satz von Sondenatomen, mit denen verschiedene Wechselwirkungstypen untersucht werden können: hydrophobe Wechselwirkungen, Wasserstoffbrückenbindungen sowie ladungsunterstützte Wechselwirkungen. Zusätzlich ähnelt diese Auswahl derjenigen, die für die Validierung von SuperStar (Verdonk *et al.*, 1999), X-SITE (Laskowski *et al.*, 1996) und einer Methode von Nissink *et al.* (Nissink *et al*., 2000) verwendet wurde und erlaubt daher einen direkten Vergleich dieser Verfahren. Für Ligandatome mit einem dieser Typen wird daraufhin ermittelt, wie häufig ihr Typ mit dem Typ der günstigsten Bewertung am nächsten Gitterpunkt übereinstimmt. Für das hier verwendete Gitter mit der Weite von 0.5 Å ist der maximal mögliche Abstand eines Ligandatoms von einem Gitterpunkt gleich der halben Raumdiagonale (≈ 0.43 Å). Da dieser Wert in etwa der mittleren Unsicherheit in den Ortskoordinaten bei Proteinstrukturen entspricht (Dauber-Osguthorpe *et al.*, 1988; Kossiakoff *et al.*, 1992; Wlodawer *et al.*, 1987), kann das Gitter als ausreichend fein betrachtet werden. Prinzipiell ist es möglich, unter Verwendung der Paarpotentiale für *jeden* Punkt des Raums einen Potentialwert zu berechnen, also auch an dem Ort des untersuchten Ligandatoms selbst. Die Verwendung des Gitteransatzes entspricht allerdings eher einer möglichen Anwendung im *de novo*-Design bzw. bei der Optimierung von Liganden bzgl. der Bindung an ein Protein.

Die Ergebnisse der Übereinstimmung zwischen vorhergesagten und experimentell gefundenen Atomtypen bei der Betrachtung von ausschließlich vergrabenen Ligandatomen sind in Tab. 28 zusammengefaßt.

**Tab. 28:** Statistik der Übereinstimmung zwischen mit den hier abgeleiteten Paarpotentialen (Gl. 15 bzw. Gl. 45) vorhergesagten und in Kristallstrukturen von Protein-Ligand-Komplexen gefundenen Atomtypen für vollständig vergrabene Ligandatome.

| **Beobachtet** | | **Vorhergesagt**[a)] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Typ** | **Anzahl** | **C.3**[b)] | **O.3**[b)] | **O.2**[b)] | **O.co2**[b)] | **N.3**[b)] | **Korrekt**[c)] | **Hydrophob-hydrophil**[d)] | **Ähnliche Wechselw.**[e)] |
| **C.3**[b)] | 745 | 92 | 3 | < 1 | < 1 | 4 | | 92 | 92 |
| **O.3**[b)] | 168 | 40 | 37 | 5 | 7 | 11 | | 60 | 60 |
| **O.2**[b)] | 124 | 18 | 26 | 27 | 23 | 6 | | 82 | 76 |
| **O.co2**[b)] | 67 | 6 | 27 | 17 | 44 | 6 | | 94 | 88 |
| **N.3**[b)] | 15 | 20 | 7 | 0 | 0 | 73 | | 80 | 80 |
| **Gesamt** | 1119 | | | | | | 74 | 86 | 85 |

a) Alle Angaben in Prozent. b) Die Bezeichnung der Atomtypen folgt der SYBYL-Notation (Tab. 2, S. 65): $sp^3$-hybridisierter Kohlenstoff (C.3), $sp^3$-hybridisierter Sauerstoff (O.3), Carbonylsauerstoff (O.2), Carboxyl(at)-Sauerstoff (O.co2), (protonierter) Aminostickstoff (N.3). c) Exakte Übereinstimmung zwischen vorhergesagten und experimentell gefundenen Ligandatomtypen. d) Die Atomtypen sind hinsichtlich hydrophober bzw. hydrophiler Eigenschaften gruppiert: C.3 bzw. O.3 / O.2 / O.co2 / N.3. e) Die Atomtypen sind hinsichtlich ähnlicher Wechselwirkungseigenschaften gruppiert (N.3 wird dabei als protoniert betrachtet): C.3 (hydrophob); O.3, N.3 (Wasserstoffbrückendonatoren); O.3, O.2, O.co2 (Wasserstoffbrückenakzeptoren).

Bei der Verwendung von fünf Atomtypen zur Evaluierung beträgt die Rate der Übereinstimmung zwischen berechneten und experimentell gefundenen Typen im Fall einer Zufallsvorhersage 20 %. Demgegenüber werden gemäß Tab. 28 (grau unterlegte Felder) deutlich höhere Raten für die *korrekte* Vorhersage des Typs erzielt. Dies gilt insbesondere für die Vorhersage von C.3 und N.3, die in 92 % bzw. 73 % der Fälle mit dem gefundenen Typ exakt übereinstimmt. Im Fall des Carbonylsauerstoffs (O.2) beträgt diese Rate nur 27 %. Hier ist aber zu beachten, daß sie unter Hinzunahme von O.3 und O.co2 auf 76 % steigt. Die beiden letzten Atomtypen sind ebenfalls Wasserstoffbückenakzeptoren und können daher ähnliche Wechselwirkungen wie O.2 eingehen. Ähnliches gilt für O.co2 mit 44 % exakter Übereinstimmung. Zieht man auch hier die Atomtypen mit in Betracht, die ähnliche Wechselwirkungen ausbilden können, so steigt die Rate sogar auf 88 %. Interessant ist, daß im Fall von O.3 (37 % exakte Übereinstimmung) in 40 % der beobachteten Fälle ein C.3-Atom als an dieser Position am günstigsten vorhergesagt wird. Ein so hoher Anteil wird in keinem der anderen untersuchten Fälle hydrophiler Atomtypen gefunden und kann darauf hindeuten, daß eine

Wechselwirkung zwischen einer funktionellen Gruppe mit einem O.3-Atom und dem Protein weniger spezifisch ist als die mit O.2, O.co2 und N.3. Insgesamt wird eine exakte Übereinstimmung zwischen vorhergesagten und experimentell gefundenen Typen für vollständig vergrabene Ligandatome in 74 % der Fälle gefunden.

Werden die Atomtypen nach ihren hydrophoben bzw. hydrophilen Eigenschaften zusammengefaßt (C.3 gegenüber O.3 / O.2 / O.co2 / N.3), so wird in 86 % aller Fälle die korrekte Gruppe vorhergesagt (Wahrscheinlichkeit einer Zufallsvorhersage: 50 %). Wird angenommen, daß Atome des N.3-Typs protoniert vorliegen und klassifiziert man die verwendeten Typen dann nach den von ihnen potentiell ausgebildeten Wechselwirkungen (hydrophob: C.3; Wasserstoffbrückendonatoren: O.3, N.3; Wasserstoffbrückenakzeptoren: O.3, O.2, O.co2), so wird in 85 % aller Fälle der korrekte Wechselwirkungstyp vorhergesagt (Wahrscheinlichkeit einer Zufallsvorhersage: 33.3 %). Insbesondere im Fall von N.3 und O.co2 ist zu bemerken, daß der jeweilige Typ mit entgegengesetzter Ladung (O.co2 im Vergleich zu N.3; N.3 im Vergleich zu O.co2) in keinem bzw. nur in 6 % der Fälle vorgeschlagen wird.

Im Hinblick auf die erreichten Übereinstimmungsraten vorhergesagter korrekter Atom- bzw. Wechselwirkungstypen ist zu bedenken, daß nicht für alle Atome eines Liganden uneingeschränkt erwartet werden kann, daß sie sich in einer für sie optimalen Proteinumgebung befinden. So ergibt sich die Position eines *einzelnen* Atoms als Resultat der Wechselwirkungen *aller* Ligandatome mit dem Protein. Stärker bindende funktionelle Gruppen können – bedingt durch vom Molekülgerüst des Liganden vorgegebene Einschränkungen – schwächer bindende daher durchaus in eine für sie nichtoptimale Umgebung des Proteins zwingen.

Abb. 37 zeigt die Abhängigkeit der Übereinstimmungsrate zwischen vorhergesagten und experimentell gefundenen Ligandatomtypen für alle untersuchten 159 Komplexe als Funktion der verwendeten Gitterweite. Bei einer Verdopplung dieser Weite von 0.5 Å auf 1.0 Å sinkt die Rate in allen Fällen (exakte Übereinstimmung bzw. Übereinstimmung hinsichtlich des Wechselwirkungstyps) um etwa 2 bis 3 %. Zusätzlich mitangegeben ist der Einfluß der Vergrabung eines Ligandatoms. Für vollständig vergrabene Atome liegt in allen Fällen die Übereinstimmungsrate um 5 bis 10 % über der bei Betrachtung aller Atome erhaltenen. Letztere Beobachtung resultiert aus der Tatsache, daß Wechselwirkungen von Lösemittelzugänglichen Ligandatomen mit dem Lösemittel nicht explizit in den verwendeten Paarpräferenzen berücksichtigt werden. Daher werden nicht in unmittelbarem Kontakt mit dem Protein stehende Ligandatome von den Paarpotentialen nur unvollständig beschrieben.

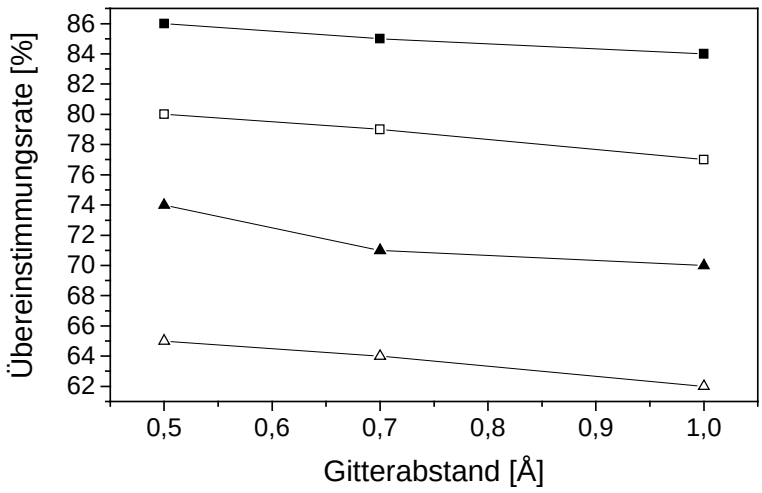



**Abb. 37:** Abhängigkeit der Rate der Übereinstimmung zwischen mit Gl. 15 bzw. Gl. 45 vorhergesagten und in kristallographisch bestimmten Protein-Ligand-Komplexstrukturen gefundenen Atomen des Typs C.3, O.3, O.2, O.co2 und N.3 als Funktion der Gitterweite. Quadratische Symbole stehen für eine Übereinstimmung bezüglich ähnlicher Wechselwirkungen (für eine Definition siehe Text), Dreiecke für eine exakte Übereinstimmung des Typs. Ausgefüllte Symbole gelten dabei für vollständig vergrabene Ligandatome, unausgefüllte bei Betrachtung aller Ligandatome.

### 5.6.3 Vergleich mit anderen Verfahren

Tab. 29 faßt die Ergebnisse von vier verschiedenen Verfahren zur Vorhersage von atomtypspezifischen „ausgezeichneten Punkten“ (*hot spots*) in Bindetaschen von Proteinstrukturen zusammen (s.a. Kap. 3.2.5). ‚Diese Arbeit(I)‘ entspricht dabei den Werten für die im vorherigen Kapitel beschriebenen Untersuchungen. Für ‚Diese Arbeit(II)‘ sind dagegen nur die Atomtypen C.3, O.3, O.2 und N.3 verwendet worden, da sie auch zur Vorhersage mit den auf Arbeiten von Verdonk *et al.* (Verdonk *et al.*, 1999) (‚SuperStar(I)‘) bzw. Nissink *et al.* (Nissink *et al.*, 2000) (‚SuperStar(II)‘) beruhenden Verfahren eingesetzt wurden. Beide Methoden verwenden im Gegensatz zu der hier entwickelten Methode aus der Kristallstruktur-Datenbank niedermolekularer Verbindungen (CSD, (Allen *et al.*, 1991)) gewonnene räumliche Wahrscheinlichkeitsdichten nichtkovalenter Wechselwirkungen. Diese Daten sind in IsoStar (Bruno *et al.*, 1997) zusammengefaßt. Das von Nissink *et al.* (Nissink *et al.*, 2000) entwickelte Verfahren paßt dazu anisotrope Gauss-Funktionen an diese Rohdaten an, wohingegen das Verfahren von Verdonk *et al.* (Verdonk *et al.*, 1999) mit den ursprünglichen, diskreten Wahrscheinlichkeitsdichten arbeitet. Das Verfahren X-SITE (Laskowski *et al.*, 1996)

verwendet dagegen aus der PDB gewonnene räumliche Wahrscheinlichkeitsdichten, wobei allerdings nur Protein-Protein-Wechselwirkungen berücksichtigt werden.

**Tab. 29:** Vergleich der Übereinstimmungsrate zwischen vorhergesagten und experimentell gefundenen Ligandatomtypen in Protein-Ligand-Komplexe.

| **Methode** | **Lösemittel-unzugängliche Atome** | | **Alle Atome** | |
|---|---|---|---|---|
| | **Korrekt[a)]** | **Ähnlich[b)]** | **Korrekt[a)]** | **Ähnlich[b)]** |
| **Diese Arbeit(I)[c)]** | 74 | 85 | 65 | 80 |
| **Diese Arbeit(II)[d)]** | 77 | 86 | 71 | 80 |
| **SuperStar(I)[e)]** | 82 | 91 | 67 | 81 |
| **SuperStar(II)[f)]** | 74 | 89 | 59 | 75 |
| **X-SITE[g)]** | -[h)] | -[h)] | 12 | 66 |

a) Exakte Übereinstimmung (in Prozent) für vorhergesagte und berechnete Ligandatomtypen. b) Übereinstimmung (in Prozent) hinsichtlich der potentiell ausgebildeten Wechselwirkungen. c) Testsatz mit 159 Protein-Ligand-Komplexen; 5 Ligandatomtypen: C.3, O.3, O.2, O.co2, N.3. d) Testsatz mit 159 Protein-Ligand-Komplexen; 4 Ligandatomtypen: C.3, O.3, O.2, N.3. e) Testsatz mit 122 Protein-Ligand-Komplexen (Verdonk *et al.*, 1999); 4 Ligandatomtypen: C.3, O.3, O.2, N.3. f) Testsatz mit 130 Protein-Ligand-Komplexen (Nissink *et al.*, 2000); 4 Ligandatomtypen: C.3, O.3, O.2, N.3. g) Testsatz mit 6 Protein-Ligand-Komplexen (Laskowski *et al.*, 1996); 26 Atomtypen gemäß der Definition von Engh und Huber (Engh & Huber, 1991). h) Keine Angabe vorhanden.

Im Fall Lösemittel-unzugänglicher Atome weist ‚SuperStar(I)' bei dem verwendeten Testdatensatz sowohl für die korrekte Vorhersage des Atomtyps als auch die unter Beachtung ähnlicher Wechselwirkungen (s.a. Kap. 5.6.2) eine um 5 % bessere Vorhersagerate auf als bei Verwendung der hier abgeleiteten statistischen Paarpotentiale (‚Diese Arbeit(II)'). Bei Betrachtung aller Atome werden dagegen für die Vorhersage von Atomtypen mit ähnlichen Wechselwirkungen für ‚Diese Arbeit(II)' als auch ‚SuperStar(I)' ähnliche Ergebnisse erzielt, wohingegen im Fall der Vorhersage des korrekten Atomtyps die hier entwickelte Methode das beste Ergebnis zeigt. ‚SuperStar(II)' und X-SITE zeigen dabei deutlich schlechtere Ergebnisse. Im Fall von X-SITE ist allerdings zu berücksichtigen, daß bei der Vorhersage des korrekten Atomtyps die Wahl zwischen 26 Atomtypen besteht, was eine Wahrscheinlichkeit bei einer Zufallsvorhersage von 3.8 % bedingt. Das erzielte Ergebniss von 12 % übertrifft diesen Wert um den Faktor drei. Beim Vergleich der im Rahmen dieser Arbeit erzielten Ergebnisse (‚Diese Arbeit(II)') mit denen von ‚SuperStar(I)' und ‚Superstar(II)' ist zudem zu beachten,

daß die verwendeten Datensätze in allen drei Fällen nicht identisch sind, was ebenfalls zu den auftretenden Unterschieden in den Übereinstimmungsraten beitragen kann.

Das im Vergleich mit bestehenden Methoden sehr gute Ergebnis des hier auf der Basis der statistischen Paarpräferenzen entwickelten Verfahrens zeigt, daß selbst bei Verwendung distanzabhängiger Paarpotentiale mit sphärischer Symmetrie durch die Gesamtbetrachtung aller Wechselwirkungen in einer Proteinbindetasche lokal begrenzte, für einen spezifischen Atomtyp günstige Bereiche identifiziert werden können. Mit dem gleichen Argument wird bei der Verwendung isotroper Einzelbeiträge (Lennard-Jones- bzw. Coulomb-Wechselwirkungen) in Kraftfeldern deren Beschreibung anisotroper Wechselwirkungen (etwa Wasserstoffbrückenbindungen) begründet (Weiner *et al.*, 1984). Daß keines der Verfahren eine Vorhersagerate von deutlich über 90 % aufweist, hängt sicher damit zusammen, daß die Positionen von Atomen eines Liganden in einer Bindetasche keineswegs unabhängig voneinander sind und nicht für alle Ligandatome erwartet werden kann, daß sie sich in einer optimalen Proteinumgebung befinden (s.a. Kap. 5.6.2).

Die implizite Berücksichtigung von Direktionalität in den hier verwendeten Paarpotentialen kann auf die Verwendung der nur 6 Å großen oberen Abstandsgrenze ($r_{max}$) zurückgeführt werden. So ist die Abnahme des Einflusses spezifischer Wechselwirkungen mit zunehmender Berücksichtigung von Wechselwirkungen bei größeren Distanzen in der Literatur beschrieben worden (Bahar & Jernigan, 1997). Dieser Punkt ist im Fall von Protein-Ligand-Komplexen von großer Bedeutung, denn obwohl die Protein-Ligand-Bindungsaffinität hauptsächlich durch den Betrag an vergrabener nichtpolarer Oberfläche bestimmt wird, sind es gerichtete Wechselwirkungen wie Wasserstoffbrückenbindungen, die die Spezifität einer Ligandbindung ausmachen (Davis & Teague, 1999). Bedingt durch die Art der Ableitung der statistischen Paarpräferenzen kann zudem davon ausgegangen werden, daß die hier erzeugten *hot spots* nicht nur Bereiche günstiger *Energie* umschließen, sondern daß zudem auch *entropische* Beiträge berücksichtigt werden (s.a. Kap. 5.3). Somit bietet sich eine mögliche Verwendung zur Ligandoptimierung bzw. Ligandplazierung im Rahmen von Docking-Programmen an (s.a. Kap. 6.2).

## *5.7 Problem-spezifische Adaptierung der Bewertungsfunktion*

### 5.7.1 Datensatz und Überlagerung

Die zur proteinspezifischen Anpassung der statistischen Paarpräferenzen verwendeten 61 Verbindungen des Trainingsdatensatzes sowie die 15 Verbindungen des Testdatensatzes sind in Tab. 13 (S. 95) und Tab. 14 (S. 95) aufgeführt; sie entstammen der Arbeit von Klebe *et al.* (Klebe *et al.*, 1994) und wurden bereits in den Arbeiten von DePriest *et al.* (De Priest *et al.*, 1993), Waller und Marshall (Waller & Marshall, 1993) und Klebe und Abraham (Klebe & Abraham, 1999) zur Validierung von 3D-QSAR-Verfahren verwendet.

Bei allen Verfahren, die relative Unterschiede in beobachteten „Wirkungen“ (hier: Bindungsaffinitäten) mit den relativen Unterschieden in molekularen Strukturen erklären wollen, ist die Frage der gegenseitigen Orientierung dieser Verbindungen (auch *Überlagerung* bzw. *Alignment* genannt) von besonderer Bedeutung. Dies schließt sowohl die Frage nach der bioaktiven Konformation der betrachteten Moleküle wie auch die nach der relativen Anordnung dieser Konformationen ein, wobei insbesondere beachtet werden muß, daß bioaktive Konformationen nicht notwendigerweise mit der des (berechneten) globalen Minimums übereinstimmen müssen (Nicklaus *et al.*, 1995). Unter den alternativen Ansätzen zur Lösung des Überlagerungsproblems (Klebe, 1993; Lemmen & Lengauer, 2000) ist die Verwendung von experimentell bestimmten bioaktiven Konformationen – etwa von Templaten aus Protein-Ligand-Kristallstrukturen – hervorzuheben (Waller *et al.*, 1993). Die Einbeziehung der Proteinstruktur wird sogar unerläßlich, wenn, wie in dieser Arbeit, die strukturellen Unterschiede in den Molekülen in Form von Wechselwirkungsfeldern in der Bindungsregion abgebildet werden.

Für 8 der insgesamt 76 hier betrachteten Thermolysin-Inhibitoren sind die Kristallstrukturen der Protein-Ligand-Komplexe in der PDB verfügbar (1tlp, 1tmn, 2tmn, 4tln, 4tmn, 5tln, 5tmn, 6tmn). Eine Überlagerung der Proteinketten von 7 dieser Strukturen auf die von 1tlp hinsichtlich der Atome des Proteinrückgrats ergibt dabei einen gemittelten *rmsd*-Wert von nur 0.14 Å. Diese große Ähnlichkeit in den Proteinstrukturen spiegelt sich trotz der Bindung unterschiedlicher Liganden auch im Bereich der Proteintasche wider (Abb. 38). Somit kann angenommen werden, daß die für 3D-QSAR-Verfahren allgemein zu erfüllende Voraussetzung gültig ist, nach der die Binderegion des biologischen Zielmoleküls annähernd starr sein muß. In Abb. 38 zeigt sich zudem, daß alle Ligandmoleküle in einem ähnlichen Bereich der Tasche an das Protein binden.

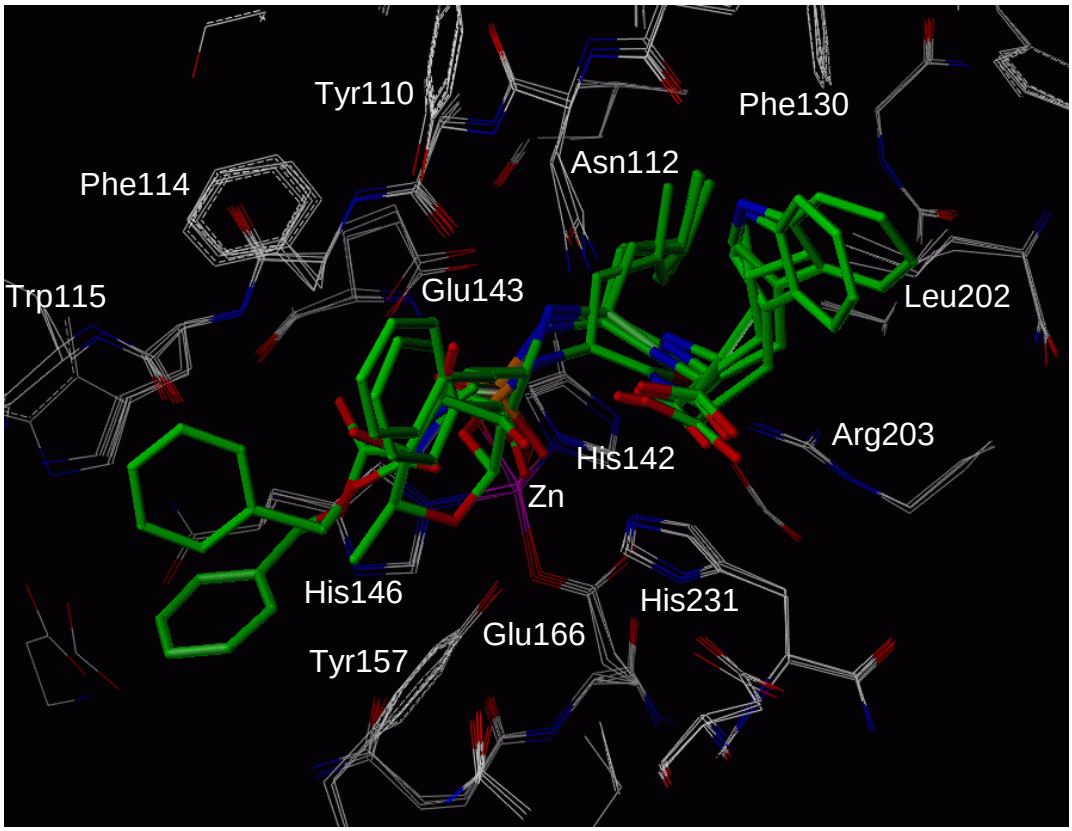



**Abb. 38:** Überlagerung der Bindetaschen der Thermolysinkomplexe 1tlp, 1tmn, 4tmn und 5tmn, die als Template für den Aufbau der Trainings- und Testdatensätze der in Tab. 13 (S. 95) und Tab. 14 (S. 95) angegebenen Verbindungen dienen. Mit grün eingefärbten Kohlenstoffatomen sind die jeweiligen Inhibitoren Phosphoramidon, CLT, ZFPLA und ZGPLL dargestellt. Auffällig sind die geringen Unterschiede in der Anordnung der Aminosäuren, die auf eine starre Bindetasche hindeuten. Für eine schematische Darstellung der Wechselwirkungen zwischen den Liganden und Thermolysin am Beispiel ZGPLL (5tmn) s.a. Abb. 36 (S. 171).

Aus diesem Grund wurden die Bindungsgeometrien von 1tlp, 1tmn, 4tmn und 5tmn als Template für den Aufbau der restlichen Verbindungen des in sich homogenen Datensatzes gewählt (s.a. Kap. 4.9.5). Die abschließende Geometrieoptimierung in der Bindetasche von 1tlp unter Verwendung des Kraftfeldes MAB (Gerber & Müller, 1995) dient der Beseitigung von ungünstigen intra- wie auch intermolekularen Wechselwirkungen; Durchdringungen von Liganden und Protein werden somit ausgeschlossen. Dies ist auch in Übereinstimmung mit dem Befund der starren Thermolysinbindetasche, durch den auf *induced fit* beruhende Veränderungen bei Bindung unterschiedlicher Liganden weitgehend ausgeschlossen werden können. Bei der hier vorgestellten Methode wird somit *bewußt* die Proteinbindetasche bei der Überlagerung der Moleküle im Sinne eines *strukturbasierten* Ansatzes mit einbezogen, im Gegensatz zu den in Kap. 3.1 beschriebenen Verfahren.

Abb. 39 zeigt die Geometrien der auf diesem Weg überlagerten 61 Liganden des Trainingssatzes in der Bindetasche von Thermolysin.

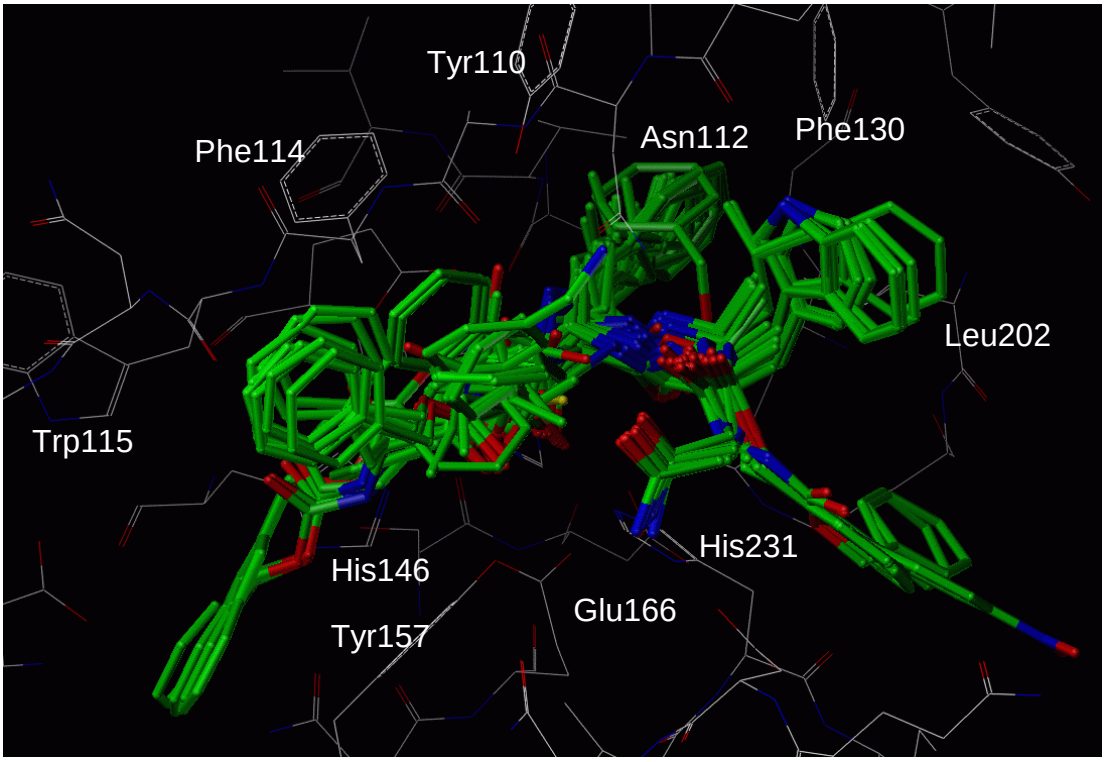



**Abb. 39:** Unter Anwendung der Template 1tlp, 1tmn, 4tmn und 5tmn erhaltene Überlagerung der Geometrien der 61 Thermolysininhibitoren des Trainingsdatensatzes (Tab. 13, S. 95) in der Bindetasche der Proteinstruktur aus dem Komplex 1tlp.

### 5.7.2 Ergebnisse der vergleichenden Feldanalyse und Signifikanz der statistischen Ergebnisse

Ausgehend von den überlagerten Strukturen des Trainingsdatensatzes und den jeweiligen Bindungsaffinitäten (Tab. 13, S. 95) wurden gemäß Gl. 48 Wechselwirkungsfelder unter Betrachtung der Atomtypen C.3, C.ar, O.3, O.2, O.co2, N.am und S.3 für Gitter der Weite 1.0, 1.5 und 2.0 Å berechnet. Die dazugehörigen Gitterparameter sind in Tab. 30 aufgeführt.

**Tab. 30:** Gitterparameter zur Berechnung der Wechselwirkungsfelder

| | **Gitter I** | | | **Gitter II** | | | **Gitter III** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **x** | **y** | **z** | **x** | **y** | **z** | **x** | **y** | **z** |
| **Untere Grenze** | -7.19 | -12.72 | -10.69 | -7.19 | -12.72 | -10.69 | -7.19 | -12.72 | -10.69 |
| **Obere Grenze** | 16.81 | 10.28 | 8.31 | 16.81 | 11.28 | 8.81 | 18.81 | 11.28 | 9.31 |
| **Schrittanzahl** | 24 | 23 | 19 | 16 | 16 | 13 | 13 | 12 | 10 |
| **Schrittweite** | | 1.0 | | | 1.5 | | | 2.0 | |
| **Gitterpunkte** | | 10488 | | | 3328 | | | 1560 | |

Die Ergebnisse der (SAM)PLS-Analysen mit bzw. ohne Kreuzvalidierung sind in Tab. 31 aufgeführt. Gemäß Gl. 55 werden während der Analysen nur die auf die Wechselwirkungsfelder zurückgehenden $pK_i$'-Werte berücksichtigt; zusätzlich mitangegeben sind jedoch auch die Statistikwerte bei Betrachtung der eigentlichen Bindungsaffinität ($pK_i$ nach Gl. 54). Für den $c_{S,Paar}$-Wert wird der Mittelwert der für die Datensätze ‚Serinproteasen', ‚Metalloproteasen', ‚Endothiapepsine', ‚Andere' und ‚Böhm1998' (s.a. Kap. 5.5.1) nach Gl. 42 erhaltenen Skalierungskonstanten verwendet, wobei die experimentellen $pK_i$-Werte ausschließlich mit den auf die *Paar*präferenzen zurückgehenden Bewertungen korreliert werden. Es ergibt sich ein Wert von $c_{S,Paar} = -3.11 \cdot 10^{-2}$. Für die Berechnung wurde eine Gitterweite von 1.0 Å, ein $\sigma$-Wert von 0.7 und ein künstlicher Abstoßungsterm mit dem Wert 10 (s.a Kap. 4.8.1) für verschwindenden Atom-Atom-Abstand gewählt (Ergebnisse bei Verwendung anderer Parameter s.u.). Bei der Kreuzvalidierung wurde die Anzahl der optimalen Komponenten gewählt, die den kleinsten $s_{PRESS}$-Wert ergaben. Mit dieser festgelegten Anzahl wurde anschließend eine nichtkreuzvalidierte Analyse durchgeführt. Die Berechnungen werden dabei sowohl mit autoskalierten unabhängigen Variablen (d.h. durch den jeweiligen Wert der Spalten-Standardabweichung geteilten Wechselwirkungsbeiträgen) wie auch mit nicht-skalierten unabhängigen Variablen durchgeführt.

**Tab. 31:** Zusammenfassung der Ergebnisse der PLS-Analysen

| | **Ohne Skalierung** | **Mit Autoskalierung[a)]** |
|---|---|---|
| $q^{2}$ **b)** | 0.62 (0.65) | 0.53 (0.57) |
| $s_{PRESS}$ **c)** | 1.34 | 1.43 |
| $r^{2}$ **d)** | 0.97 (0.97) | 0.96 (0.96) |
| $S$ **e)** | 0.37 | 0.43 |
| $F$ **f)** | 166 (181) | 201 (219) |
| **Komponenten** | 10 | 6 |
| **Beitrag[g)]** | | |
| **C.3** | 0.35 | 0.29 |
| **C.ar** | 0.30 | 0.35 |
| **O.3** | 0.08 | 0.09 |
| **O.2** | 0.07 | 0.08 |
| **O.co2** | 0.07 | 0.06 |
| **N.am** | 0.10 | 0.10 |
| **S.3** | 0.04 | 0.03 |

a) Die Werte der unabhängigen Variablen werden durch die jeweilige Spalten-Standardabweichung geteilt. b) Gemäß Gl. 56. In Klammern mit angegeben ist das Ergebnis bei Betrachtung der gesamten $pK_i$-Werte, zusammengesetzt aus dem in der PLS-Analyse angepaßten Anteil $pK_i$‘ und dem auf dabei nicht berücksichtigte Atomtypen zurückgehenden Anteil $pK_i$‘‘ (s.a. Gl. 54). c) Gemäß Gl. 57. d) Gemäß Gl. 58. In Klammern mit angegeben ist das Ergebnis bei Betrachtung der gesamten $pK_i$-Werte, zusammengesetzt aus dem in der PLS-Analyse angepaßten Anteil $pK_i$‘ und dem auf dabei nicht berücksichtigte Atomtypen zurückgehenden Anteil $pK_i$‘‘ (s.a. Gl. 54). e) Gemäß Gl. 59. f) Gemäß Gl. 60. In Klammern mit angegeben ist das Ergebnis bei Betrachtung der gesamten $pK_i$-Werte, zusammengesetzt aus dem in der PLS-Analyse angepaßten Anteil $pK_i$‘ und dem auf dabei nicht berücksichtigte Atomtypen zurückgehenden Anteil $pK_i$‘‘ (s.a. Gl. 54). g) Gemäß Gl. 61.

In beiden Fällen liegt der auf diese Weise erhaltene $q^2$-Wert über der Grenze von 0.5, ab der ein erhaltenes Modell zur Erklärung der Unterschiede in den Bindungsaffinitäten mit Unterschieden in der Struktur der betrachteten Verbindungen als „gut“ bezeichnet wird (s.a. Kap. 4.8.2). Der mit 0.62 um nahezu 17 % größere $q^2$-Wert wird bei Verwendung der nichtskalierten unabhängigen Variablen erhalten, verglichen mit dem unter Verwendung der autoskalierten Variablen berechneten Betrag von 0.53. Bei Betrachtung der Ergebnisse der PLS-Analyse ohne Kreuzvalidierung sind die $r^2$-Werte zwar annähernd gleich (0.96 bzw. 0.97); für die Bewertung der Qualität des erhaltenen Modells ist jedoch der kreuzvalidierte $q^2$-Wert ausschlaggebend, da nur er etwas über die mit diesem Modell zu erwartenden *Vorhersage*ergeb-

nisse aussagt. Das bessere Ergebnis unter Verwendung der Wechselwirkungsfelder ohne Skalierung deutet darauf hin, daß die dabei verwendeten Variablen mit unveränderter Gewichtung mehr Information enthalten als die nach einer Einheitsgewichtung durch Autoskalierung erhaltenen. Dieser Befund erklärt sich damit, daß die der Berechnung zugrundeliegenden Wechselwirkungsfelder auf die statistischen Paarpräferenzen zurückgehen, diese durch die Art ihrer Ableitung (s.a. Kap. 4.2) aber schon zueinander gewichtet sind. Somit ist eine zusätzliche Einflußnahme auf die Gewichtung der verschiedenen Variablen ungünstig; alle weiteren Berechnungen werden daher ohne Skalierung durchgeführt.

Der für die Berechnung ohne Skalierung ermittelte $s_{PRESS}$-Wert von 1.34 entspricht bei einer Anzahl von 10 Komponenten einer Standardabweichung zwischen experimentell bestimmten und in der Kreuzvalidierung vorhergesagten Bindungsaffinitäten für die 61 Thermolysininhibitoren von 1.22 logarithmischen Einheiten. Berücksichtigt man, daß die aus der Arbeit von Klebe *et al.* (Klebe *et al.*, 1994) extrahierten Daten ursprünglich aus verschiedenen Originalquellen stammen, so kann der Fehler in den experimentellen Bindungsaffinitäten zu etwa einer logarithmischen Einheit abgeschätzt werden (Böhm, 1998; Hosur *et al.*, 1994; Murray *et al.*, 1998). Somit reicht das Ergebnis der Vorhersage bei der Kreuzvalidierung an diese vorgegebene Grenze heran.

Um die prinzipiell für Regressionsverfahren bestehende Möglichkeit auszuschließen, daß die im vorherigen Abschnitt beschriebenen Ergebnisse auf einer Zufallskorrelation beruhen, wurden die biologischen Daten der Verbindungen des Trainingssatzes zufällig miteinander vertauscht und für das so erhaltene Gleichungssystem (s.a. Kap. 4.8.2) erneut eine PLS-Analyse durchgeführt. Aus Tab. 32 wird ersichtlich, daß die sich für die ersten 10 Komponenten ergebenden $q^2$-Werte deutlich kleiner als Null sind. Das mit den zufallsverteilten Werten erhaltene Modell weist also eine schlechtere Vorhersagefähigkeit auf als jenes, daß bei der Verwendung des Mittelwertes der abhängigen Variablen des Datensatzes erhalten würde (dies entspräche einem $q^2$-Wert von Null). Somit kann der Verdacht einer Zufallskorrelation in obigen Ergebnissen ausgeschlossen werden.

**Tab. 32:** PLS-Analyse mit zufallsverteilten biologischen Aktivitäten

| **Anzahl der Komponenten** | $q^{2}$ **a)** | $s_{PRESS}$ **b)** |
|---|---|---|
| 1 | -0.51 (-0.39) | 2.45 |
| 2 | -0.30 (-0.20) | 2.30 |
| 3 | -0.28 (-0.18) | 2.30 |
| 4 | -0.26 (-0.17) | 2.31 |
| 5 | -0.41 (-0.30) | 2.46 |
| 6 | -0.45 (-0.33) | 2.51 |
| 7 | -0.47 (-0.35) | 2.56 |
| 8 | -0.47 (-0.35) | 2.58 |
| 9 | -0.41 (-0.30) | 2.56 |
| 10 | -0.36 (-0.25) | 2.54 |

a) Gemäß Gl. 56. In Klammern mit angegeben ist das Ergebnis bei Betrachtung der gesamten $pK_i$-Werte, zusammengesetzt aus dem in der PLS-Analyse angepaßten Anteil $pK_i$‘ und dem auf dabei nicht berücksichtigte Atomtypen zurückgehenden Anteil $pK_i$‘‘ (s.a. Gl. 54). b) Gemäß Gl. 57.

In Abb. 40 sind die aus der nicht-kreuzvalidierten PLS-Analyse berechneten $pK_i$-Werte gegen die experimentell bestimmten Bindungsaffinitäten aufgetragen. Die $pK_i$-Werte setzen sich hierbei aus dem angepaßten Anteil $pK_i$‘ und dem auf nicht berücksichtigte Atomtypen zurückgehenden Anteil $pK_i$‘‘ zusammen (Gl. 54). Bei ausschließlicher Betrachtung der $pK_i$‘-Werte ergibt sich eine nahezu identische Abbildung, bei der alle Punkte jedoch um etwa 0.7 Einheiten zum Ursprung hin verschoben sind. Zusätzlich zur idealen Korrelationsgeraden sind Bereiche der Abweichung von ± 1 logarithmischen Einheit gekennzeichnet. Alle berechneten $pK_i$-Werte liegen dabei innerhalb dieser Grenzen.

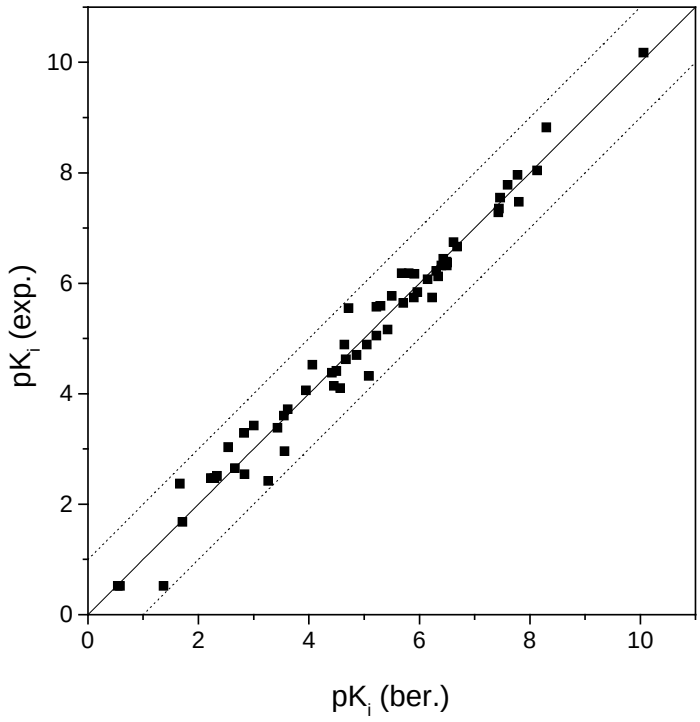



**Abb. 40:** Auftragung der ohne Anwendung einer Skalierung der unabhängigen Variablen mit dem Modell aus der PLS-Analyse (ohne Kreuzvalidierung) berechneten $pK_i$-Werte (Gl. 54) gegen die experimentell bestimmten Größen. Die gepunkteten Linien geben eine Abweichung von ± 1 logarithmischen Einheit gegenüber der idealen Korrelationsgeraden an.

Der Beitragsanteil der einzelnen Wechselwirkungsfelder zur Erklärung der Bindungsaffinitätsunterschiede ist ebenfalls in Tab. 31 angegeben. Es fällt auf, daß die Beiträge der auf C.3 und C.ar zurückgehenden Felder etwa drei bis vier mal so hoch sind wie die der polaren Atome, im Fall des S.3-Atomtyps sogar etwa acht mal. Dies gilt sowohl mit als auch ohne zusätzliche Skalierung der unabhängigen Variablen. Die Auswahl der verwendeten Atomtypen erfolgte zunächst so, daß insbesondere nach der Überlagerung beobachtbare Wechselwirkungs*unterschiede* mit dem umgebenden Protein repräsentiert werden. Ausgehend von der C.3 / O.3 / O.2 / O.co2 / N.am-Kombination wurde dabei ein $q^2$-Wert von 0.43 hinsichtlich der betrachteten $pK_i$'-Werte erhalten (Tab. 33). Werden zusätzlich C.ar-Atome berücksichtigt, findet eine deutliche Steigerung auf einen $q^2$-Wert von 0.57 statt (33 % bezogen auf den Wert von 0.43). Die Hinzunahme des S.3-Feldes bewirkt eine erneute Zunahme dieses Parameters auf 0.62 (9 % bezogen auf den Wert von 0.57), die ebenfalls als signifikant angesehen werden kann. Wird statt des S.3-Feldes ein auf den Atomtypen C.2, N.3, N.ar oder P.3 beruhendes Wechselwirkungsfeld mit einbezogen, kommt es allerdings zu keiner Steigerung des $q^2$-

Wertes. Somit erweist sich die Verwendung der auf der Kombination C.3 / C.ar / O.3 / O.2 / O.co2 / N.am / S.3 beruhenden Wechselwirkungsfelder als optimal.

**Tab. 33:** Statistische Parameter der bei einer kreuzvalidierten (SAM)PLS-Analyse unter Verwendung verschiedener Wechselwirkungsfeld-Kombinationen für den Trainingsdatensatz (Tab. 13, S. 95) erhaltenen Ergebnisse.

| **Kombinationen der Wechselwirkungsfelder** | $q^{2}$ **a)** | $s_{PRESS}$ **b)** |
|---|---|---|
| C.3 / O.3 / O.2 / O.co2 / N.am | 0.43 (0.48) | 1.58 |
| C.3 / C.ar / O.3 / O.2 / O.co2 / N.am | 0.57 (0.60) | 1.44 |
| C.3 / C.ar / O.3 / O.2 / O.co2 / N.am / S.3 | 0.62 (0.65) | 1.34 |
| C.3 / C.ar / O.3 / O.2 / O.co2 / N.am / C.2 | 0.57 (0.59) | 1.47 |
| C.3 / C.ar / O.3 / O.2 / O.co2 / N.am / N.3 | 0.58 (0.61) | 1.41 |
| C.3 / C.ar / O.3 / O.2 / O.co2 / N.am / N.ar | 0.57 (0.60) | 1.43 |
| C.3 / C.ar / O.3 / O.2 / O.co2 / N.am / P.3 | 0.58 (0.59) | 1.45 |

a) Gemäß Gl. 56. In Klammern mit angegeben ist das Ergebnis bei Betrachtung der gesamten $pK_i$-Werte, zusammengesetzt aus dem in der PLS-Analyse angepaßten Anteil $pK_i$‘ und dem auf dabei nicht berücksichtigte Atomtypen zurückgehenden Anteil $pK_i$‘‘ (s.a. Gl. 54). b) Gemäß Gl. 57.

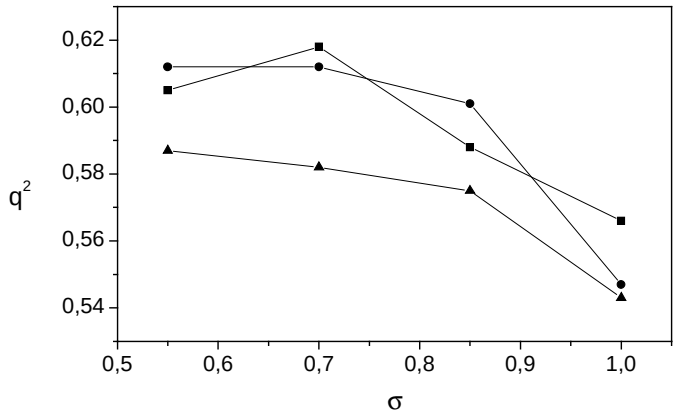
Im Fall der 3D-QSAR Verfahren wie CoMFA und CoMSIA ist diskutiert worden (Kim *et al.*, 1998), daß die Verwendung zusätzlicher Eigenschaftsfelder die statistische Signifikanz der bei der PLS-Analyse erhaltenen Ergebnisse erhöhen kann. Diese Annahme konnte bei der Einführung von zusätzlichen Feldern zur Beschreibung von Wasserstoffbrückenbindungs-Eigenschaften im CoMSIA-Verfahren nicht bestätigt werden (Klebe & Abraham, 1999). Hierbei ist allerdings zu beachten, daß die in diesen Fällen verwendeten Eigenschaftsbeschreibungen (sterisch, elektrostatisch, hydrophob, Wasserstoffbrückendonor und -akzeptor) *generisch* und nicht atom(typ)spezifisch sind. Aus der Tatsache, daß z.B. das Hinzufügen einer Hydroxylgruppe zu einem Molekül einen Beitrag zu *allen* oben angeführten physikochemischen Eigenschaftsfeldern liefert, folgt, daß diese Eigenschaften sicher nicht unabhängig voneinander, sondern auf komplexe Weise miteinander korreliert sind. Somit kann nicht erwartet werden, daß die Bindungsaffinitätsdaten der untersuchten Verbindungen *additiv* durch die einzelnen Felder erklärt werden. In dem Fall der hier verwendeten atomtypspezifischen Wechselwirkungsbeschreibungen ist die Unabhängigkeit der resultierenden Felder jedoch gegeben, da etwa das erwähnte Hinzufügen der Hydroxylgruppe sich auch nur im O.3-Feld auswirkt. Eine Beschränkung der verwendeten Feldanzahlen ergibt sich hierbei nur durch die Anwesenheit der jeweiligen Atomtypen im Trainingsdatensatz sowie den zur Ver-

fügung stehenden Arbeitsspeicher zur Aufnahme der zusätzlichen Spaltenwerte. Nicht berücksichtigte Atomtyp-Felder stellen jedoch prinzipiell keinen Informationsverlust dar, wird doch bei der Vorhersage von Affinitätsdaten unbekannter Verbindungen ihr auf den ursprünglich abgeleiteten Paarpotentialen beruhender Beitrag verwendet (s.u. sowie Gl. 54). Ein zusätzlicher Vorteil atomtypspezifischer gegenüber generischen Eigenschaftsfeldern sollte bei der graphischen Analyse der erzielten Ergebnisse zur Verbesserung der Bindungseigenschaft bekannter Verbindungen zum Tragen kommen. Während im CoMFA-Verfahren für einen Teil der Bindetasche z.B. vorhergesagt werden könnte, daß ein „sterisch anspruchsvoller Rest mit negativem elektrostatischen Feld“ dort vorteilhaft ist, würde in dem hier vorgestellten Verfahren eine darüber hinausgehende Information in Form von konkreten Vorschlägen zu günstigen Atomtypen gegeben werden.

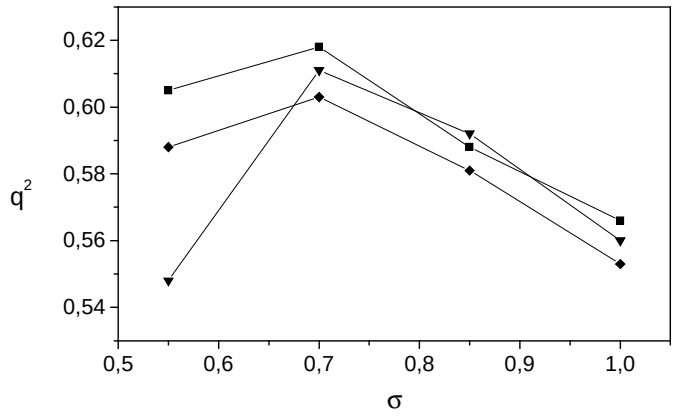
In Abb. 41 ist die Abhängigkeit des bei der kreuzvalidierten (SAM)PLS-Analyse für den Trainingsdatensatz der 61 Thermolysininhibitoren (Tab. 13, S. 95) erhaltenen $q^2$-Wertes von der Steilheit der zur Abbildung der Wechselwirkungen auf die umliegenden Gitterpunkte verwendeten sphärischen Gauss-Funktion angegeben. Zusätzlich wurde dazu die Höhe des künstlichen Abstoßungsterms (Abb. 41 a) bzw. die Gitterweite variiert (Abb. 41 b). In beiden Fällen erweist sich dabei ein $\sigma$-Wert von 0.7 als optimal. Die Wahl des $\sigma$-Parameters bestimmt hierbei, welchen Beitrag die Potentialwerte auf den um ein Atom liegenden Gitterpunkten zum Gesamtwechselwirkungsbeitrag des betrachteten Atoms leisten. So wird bei einem Gitter der Weite 1 Å für ein auf einem Gitterpunkt befindliches Atom der Beitrag der *nächsten* Gitterpunkte bei Verwendung von Gl. 47 und $\sigma = 0.7$ um den Faktor 0.21 abgeschwächt, für 2 Å entfernte Gitterpunkte beträgt er nur noch etwa ein Hundertstel. Hieraus wird deutlich, daß v.a. lokale Wechselwirkungseigenschaften in dem hier verwendeten Ansatz berücksichtigt werden. Der aus Abb. 41 a) ersichtliche Unterschied der $q^2$-Werte bei Verwendung verschiedener Höhen des künstlich angefügten Abstoßungsbeitrags (s.a. Kap. 4.8.1) ist bei Verwendung der Werte 10 bzw. 20 gering.

Daß auch bei der Verwendung von Gitterweiten von 1.5 bzw. 2.0 Å ein $\sigma$-Wert von 0.7 optimal ist, vermag auf den ersten Blick überraschen. Allerdings darf hierbei nicht vergessen werden, daß die in der Bindetasche liegenden Ligandmoleküle Kontakt zum Protein besitzen. Die Wahl eines größeren $\sigma$-Wertes würde also nicht nur zur Einbeziehung von Gitterpunkten im eigentlich freien Raum der Tasche führen, sondern vermehrt auch von solchen im Bereich der Proteinatome. Da hier jedoch nur ein geringfügig für den jeweiligen Atomtyp spezifischer, durch den Abstoßungsterm begründeter Potentialwert vorliegt, ergibt sich insgesamt

eine Erniedrigung des zur Verfügung stehenden Informationsgehaltes. Eine Variation der Gitterweite bei einem festen $\sigma$-Wert von 0.7 zeigt im Gegensatz zu in der Literatur erwähnten Beispielen (Cramer III *et al.*, 1993; Folkers *et al.*, 1993) keine Verschlechterung der statistischen Ergebnisse für kleinere Abstandwerte. Für eine Weite von 1.0 Å wird sogar ein um 7 % größerer $q^2$-Wert berechnet als für eine Weite von 2.0 Å. Aus diesem Grund wird trotz der um den Faktor 8 höheren Rechenzeit dieser Gitterabstand verwendet.




**a)**




**b)**

**Abb. 41:** Abhängigkeit der bei der kreuzvalidierten (SAM)PLS-Analyse für den Trainingsdatensatz erzielten Ergebnisse von der Steilheit der für die Abbildung der Wechselwirkungen auf die umliegenden Gitterpunkte verwendeten sphärischen Gauss-Funktion. Als Maß für die Steilheit wird hierbei der $\sigma$-Wert verwendet; sein Zahlenwert entspricht gemäß Gl. 47 der Lage der Wendepunkte der Funktionen. In Teil a) wird bei festgehaltenem Gitterabstand von 1 Å die Höhe des in Kap. 4.8.1 beschriebenen künstlichen Abstoßungsterms variiert (10: ■; 20: ●; 40: ▲), in Teil b) bei festgehaltener Abstoßungshöhe von 10 der Gitterabstand (1.0 Å: ■; 1.5 Å: ◆; 2.0 Å: ▼).

### 5.7.3 Vorhersagefähigkeit des erhaltenen Modells bei Variation der Einflußnahme der proteinspezifischen Information

Zur Überprüfung der Vorhersagefähigkeit des aus der PLS-Analyse für 61 Verbindungen des Trainingsdatensatzes erhaltenen Modells wird damit die Bindungsaffinität nach Gl. 54 für 15 Verbindungen eines Testdatensatzes (Tab. 14, S. 95) berechnet und mit experimentell bekannten Affinitäten verglichen (Abb. 42). Hierbei liegen die vorhergesagten Werte für acht der 15 Verbindungen innerhalb der zusätzlich mit eingezeichneten Abweichung vom tatsächlichen Wert von ± 1 logarithmischen Einheit. Drei der Verbindungen (ZFGNH2, ZLGNH2 und ZYGNH2) zeigen dagegen deutliche Abweichungen von mehr als 2.5 logarithmischen Einheiten; die für sie berechnete Bindungsaffinität wird dabei zu gering vorhergesagt. Hierbei kommen zwei Ursachen zusammen. Zum einen weisen alle drei Verbindungen keinen Substituenten auf, der in die hydrophobe S1‘-Tasche zeigt, zum anderen binden alle über eine terminale Amidgruppe an das Zinkion. Im Trainingsdatensatz trifft ersteres nur noch für die Verbindungen NHOHMALAGNH2 ($pK_i$ = 2.96), ZGGNHOH ($pK_i$ = 3.03) und PAAOH ($pK_i$ = 4.06) zu, von denen die ersten beiden mit einer Hydroxamsäuregruppe eine Wechselwirkung mit dem Zink eingehen, die letzte dagegen mit einer Phosphatgruppe an das Metallkation bindet. Umgekehrt binden die Verbindungen ACE_OHLEU_AGNH2 ($pK_i$ = 2.47), ZGLNH2 ($pK_i$ = 1.68) und Z_NH_GLNH2 ($pK_i$ = 3.42) mit einer terminalen Amidgruppe an das Zink, weisen allerdings auch einen Substituenten für die S1‘-Tasche auf. Wird die in allen sechs Trainingsfällen zumindest schwache Bindung an das Protein auf die Anwesenheit mindestens *eine* der beiden Eigenschaften (S1‘-Substituent oder *Nicht*-Amid-Gruppe für die Bindung an das Zink) zurückgeführt, lassen sich die zu gering vorhergesagten Bindungsaffinitäten der drei „Ausreißer“ ZFGNH2, ZLGNH2 und ZYGNH2 auf das Fehlen *beider* Komponenten in ihren Strukturen - kein Substituent für die S1‘-Tasche und nur eine über eine Amidbindung vermittelte Wechselwirkung zum Zink - zurückführen. Hierbei wird deutlich, daß ein auf einer Regression basierendes Verfahren immer nur *inter*polieren, nicht jedoch *extra*polieren kann.

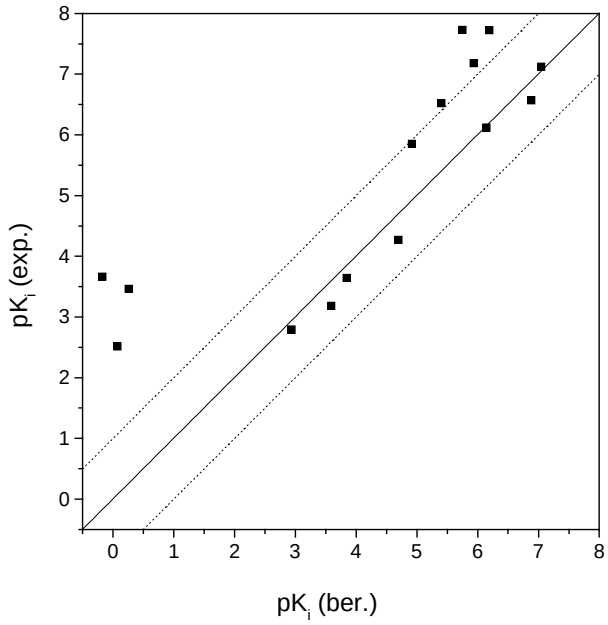



**Abb. 42:** Auftragung der mit dem aus der PLS-Analyse erhaltenen Modell vorhergesagten gegen die experimentellen $pK_i$-Werte für die 15 Verbindungen des Testdatensatzes (Tab. 14, S. 95). Zusätzlich mit eingezeichnet sind Abweichungsgrenzen von ± 1 logarithmischen Einheit (gepunktete Linien).

Der nach Gl. 63 berechnete $r^2_{pred}$-Wert, die Standardabweichung zwischen vorhergesagten und experimentellen Werten sowie die maximale Abweichung sind in Tab. 34 aufgeführt. Zusätzlich werden die statistischen Parameter der Vorhersageergebnisse der Bindungsaffinitäten unter Verwendung der ursprünglichen Paarpotentiale nach Gl. 15 sowie Gl. 42 angegeben. Die dazugehörige Auftragung der experimentellen gegen die berechneten Werte ist in Abb. 43 gezeigt. Sowohl der visuelle Vergleich als auch der der statistischen Werte zeigt, daß unter Verwendung des aus der PLS-Analyse erhaltenen Modells die Vorhersagegenauigkeit beträchtlich ansteigt (hinsichtlich der bei der Verwendung des PLS-Modells auftretenden größeren maximalen Abweichung s.u.). Ein Vergleich mit den in Tab. 26 (S. 159) aufgeführten Ergebnissen für diesen Testdatensatz (‚Thermolysin(II)‘) ergibt dagegen, daß die alleinige Verwendung der Paarpotentiale gegenüber der gesamten Bewertungsfunktion nach Gl. 36 sowie der durch die Überlagerung erhaltenen Protein-Ligand-Anordnungen im Gegensatz zu den mit FlexX erzeugten nur zu einer unwesentlichen Verschlechterung der Ergebnisse führt.

**Tab. 34:** Statistik der Vorhersage der Bindungsaffinitäten für 15 Verbindungen des Testdatensatzes (Tab. 14, S. 95) unter Verwendung des aus der PLS-Analyse (61 Trainingsverbindungen) erhaltenen Modells sowie der ursprünglichen Paarpotentiale (Gl. 15 / Gl. 42).

| **Berechnungsmethode** | $r^2_{pred}$ [a)] | $SD$ [b)] | $MD$ [c)] |
|---|---|---|---|
| PLS-Modell ($\theta = 1.0$)[d)] | 0.69 | 1.11 | 3.83 |
| PLS-Modell ($\theta = 0.8$)[d)] | 0.69 | 1.11 | 2.91 |
| Paarpotentiale | 0.34 | 1.55 | 2.56 |

a) Nach Gl. 63. b) Standardabweichung der berechneten von den experimentellen Bindungsaffinitäten. c) Maximale Abweichung zwischen berechneter und experimenteller Bindungsaffinität. d) Bezüglich der Bedeutung von $\theta$ s.a. Gl. 62.

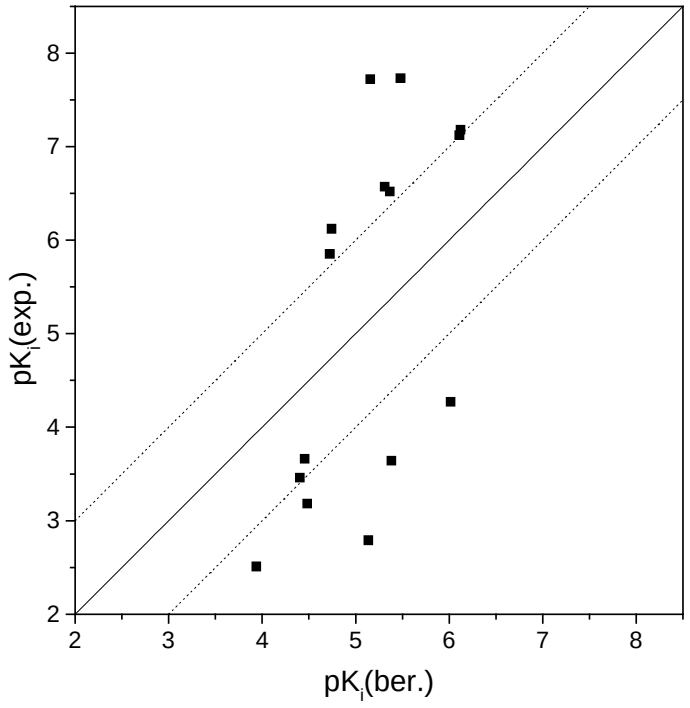



**Abb. 43:** Auftragung der unter Verwendung der ursprünglichen Paarpotentiale vorhergesagten gegen die experimentellen $pK_i$-Werte für die 15 Verbindungen des Testdatensatzes (Tab. 14, S. 95). Zusätzlich mit eingezeichnet sind Abweichungsgrenzen von ± 1 logarithmischen Einheit (gepunktete Linien).

Um der Frage nachzugehen, *wieviel* Zusatzinformation in Form von Protein-Ligand-Komplexen mit bekannter Struktur und Bindungsaffinität benötigt wird, um eine gegenüber der Verwendung der ursprünglichen Paarpräferenzen gesteigerte Verläßlichkeit bei der Vorhersage neuer Verbindungen zu erlangen, wurden jeweils 100 mal 5, 15, 30, 45 und 53 Verbindungen aus dem Trainingsdatensatz zufällig ausgewählt und mit ihnen nach dem in

Kap. 4.8.2 beschriebenen Verfahren eine PLS-Analyse durchgeführt (s.a. Kap. 4.8.3). Mit den für jede der Teilmengen erhaltenen 100 Modellen sowie mit dem unter Verwendung aller 61 Trainingsdaten aufgestellten Modell wurden sodann unter Variation des Parameters $\theta$ zur Beimischung der adaptierten Potentialwerte in Gl. 62 zwischen 0.1 und 1 die Bindungsaffinitäten für die 15 Verbindungen des Testdatensatzes berechnet. Die dabei erhaltenen, über alle Läufe gemittelten $r^2_{pred}$-Werte (Gl. 63) sind in Abb. 44 dargestellt. Zusätzlich ist der bei Verwendung der ursprünglichen Paarpotentiale erhaltene Wert (*) von 0.34 (Tab. 34) mit aufgeführt. Schon bei Verwendung von nur 5 Trainingsdaten ergibt sich bei $\theta = 0.5$ (d.h. bei einer 1:1-Gewichtung der Beiträge der ursprünglichen Paarpräferenzen und der auf die PLS-Modelle zurückgehenden Beiträge) ein $r^2_{pred}$-Wert von 0.43. Unter Berücksichtigung der für diesen Punkt erhaltenen Standardabweichung von 0.08 liegt die Verbesserung der Vorhersagegenauigkeit jedoch an der unteren Signifikanzgrenze. Eine signifikant verbesserte Vorhersagegenauigkeit ist allerdings bereits bei der Verwendung von 15 Trainingsverbindungen gegeben, wird hierbei doch ein $r^2_{pred}$-Wert von 0.48 (Standardabweichung: 0.05) bei $\theta = 0.6$ erzielt. Für 30, 45 und 53 Trainingsverbindungen ergeben sich dementsprechend Werte von 0.56 (Standardabweichung: 0.08; $\theta = 0.7$), 0.66 (Standardabweichung: 0.07; $\theta = 0.8$) und 0.68 (Standardabweichung: 0.05; $\theta = 0.9$). Der für die PLS-Analyse mit allen Trainingsverbindungen erhaltene maximale $r^2_{pred}$-Wert beträgt 0.69 ($\theta = 1.0$, s.a. Tab. 34). Wie zu erwarten nimmt mit zunehmender Zahl berücksichtigter Trainingsdaten auch der zu dem jeweiligen maximalen $r^2_{pred}$-Wert gehörige $\theta$-Wert zu: je mehr Daten aus dem Trainingssatz verwendet werden, desto sicherer wird auch die auf den PLS-Modellen beruhende Vorhersage. Daß allerdings auch bei 61 Verbindungen eine alleinige Berücksichtigung der PLS-Ergebnisse aufgrund der vorliegenden $r^2_{pred}$-Werte nicht optimal sein muß, wird bei der Betrachtung des *maximalen* Fehlers deutlich. Bei einem Unterschied in den $r^2_{pred}$-Werte für $\theta = 0.8$ bzw. $\theta = 1.0$ lediglich in der dritten Nachkommastelle kann bei einer dem ersten $\theta$-Wert entsprechenden Gewichtung dieser maximale Vorhersagefehler um nahezu eine logarithmische Einheit gegenüber der $\theta = 1.0$-Gewichtung gesenkt werden (s.a. Tab. 34).

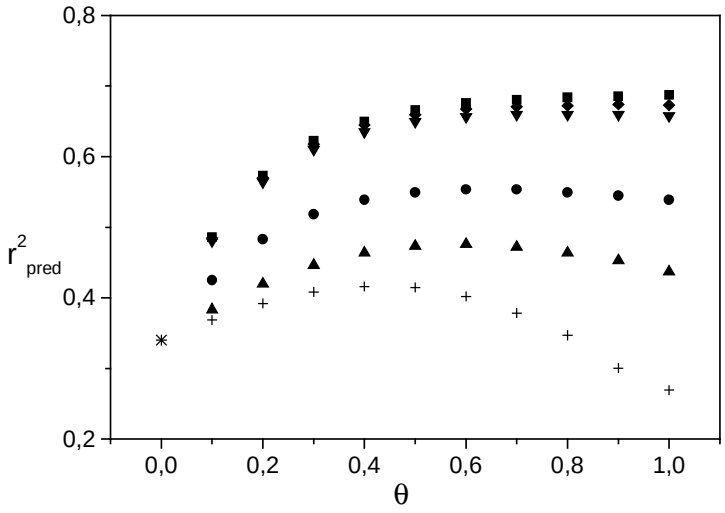



**Abb. 44:** Abhängigkeit des $r^2_{pred}$-Wertes (Gl. 63) bei der Vorhersage von Bindungsaffinitäten für die 15 Verbindungen des Testdatensatzes (Tab. 14, S. 95) von der Gewichtung $\theta$ der auf die ursprünglichen Paarpräferenzen bzw. die erhaltenen PLS-Modelle zurückgehenden Beiträge (Gl. 62). Zusätzlich ist die Anzahl der zur Berechnung der PLS-Analysen verwendeten Trainingsdaten (Tab. 13, S. 95) variiert worden (0 (d.h. ausschließliche Verwendung der Paarpotentiale): *; 5: +; 15: ▲; 30: ●; 45: ▼; 53: ◆; 61: ■).

### 5.7.4 Vergleich mit anderen Methoden

Die in Tab. 13 (S. 95) und Tab. 14 (S. 95) aufgeführten Trainings- und Testdatensätze wurden bereits in den Arbeiten von DePriest *et al.* (De Priest *et al.*, 1993), Waller und Marshall (Waller & Marshall, 1993) und Klebe *et al.* (Klebe & Abraham, 1999; Klebe *et al.*, 1994) zur Validierung von 3D-QSAR-Verfahren verwendet. Sie stellen daher eine solide Vergleichsgrundlage für die in dieser Arbeit mit ihnen erhaltenen Ergebnisse dar. Tab. 35 gibt eine Übersicht der statistischen Parameter für die (SAM)PLS-Analyse mit und ohne *leave-one-out*-Kreuzvalidierung des Trainingsdatensatzes sowie für die Vorhersage von Bindungsaffinitäten der Testdatensatzverbindungen.

**Tab. 35:** Zusammenfassung der statistischen Parameter der PLS-Analysen und Affinitätsvorhersagen für die in Tab. 13 (S. 95) und Tab. 14 (S. 95) angegebenen Datensätze von Thermolysininhibitoren.

| | **De Priest[a)]** | **Waller[b)]** | **Klebe I[c)]** | **Klebe II[d)]** | **Diese Arbeit** |
|---|---|---|---|---|---|
| **N[e)]** | 61 (11) | 61 (11) | 61 (15) | 61 | 61 (15) |
| $q^{2}$ **f)** | 0.70 | 0.54 | 0.59 | 0.59 | 0.62 |
| $s_{PRESS}$ **g)** | 1.30[k)] | 0.83[k)] | 1.41 | 1.41 | 1.34 |
| **Komponenten** | 11 | 5 | 7 | 7 | 10 |
| $r^{2}$ **h)** | 0.98 | 0.85 | 0.90 | 0.93 | 0.97 |
| $S$ **i)** | 0.31[l)] | - | 0.71 | 0.60 | 0.37 |
| $r^{2}_{pred}$ **j)** | 0.29 | 0.44 | 0.56 | - | 0.69 |

a) Die Werte stammen aus Tab. IX in (De Priest *et al.*, 1993), Zeile ‚CoMFA only'. b) Die Werte stammen aus Tab. X in (Waller & Marshall, 1993), Zeile ‚CoMFA (S,E)'. c) Die Werte stammen aus Tab. 3 in (Klebe *et al.*, 1994), Spalte ‚CoMSIA(6)'. d) Die Werte stammen aus Tab. 1 in (Klebe & Abraham, 1999), Spalte ‚CoMSIA(3)'. e) Anzahl der Trainings- und Testverbindungen (letztere in Klammern). f) Gemäß Gl. 56. g) Gemäß Gl. 57. h) Gemäß Gl. 58. i) Gemäß Gl. 59. j) Gemäß Gl. 63. k) Angegeben wurde der „*standard error of estimate*". l) Angegeben wurde die Standardabweichung.

Die der Arbeit von DePriest *et al.* (De Priest *et al.*, 1993) entnommenen und in Tab. 35 aufgeführten Werte resultieren aus einer dort durchgeführten CoMFA-Analyse unter Verwendung des sterischen und elektrostatischen Feldes sowie einer Überlagerung der Verbindungen unter Verwendung von Kristallstrukturtemplaten. Der auffällig hohe $q^2$-Wert von 0.70 geht auf eine mit nur 30 anstelle der 61 möglichen Kreuzvalidierungsgruppen durchgeführte Analyse zurück. Der angegebene $r^2_{pred}$-Wert von 0.29 weist somit zusätzlich auf eine nur eingeschränkte Vorhersagefähigkeit mit diesem Modell hin. In einer Folgestudie von Waller und Marshall (Waller & Marshall, 1993) wurde mit dem gleichen Datensatz erneut unter Anwendung einer CoMFA-Analyse eine auf 61 Gruppen beruhende Kreuzvalidierung durchgeführt, die ein Modell mit 5 Komponenten und einem $q^2$-Wert von 0.536 ergab. Die damit durchgeführte Vorhersage für 11 Testverbindungen ergab eine signifikante Korrelation (> 0.3). Die aus der Arbeit von Klebe *et al.* (Klebe *et al.*, 1994) (‚Klebe(I)') entnommenen Zahlenwerte beruhen auf einer mit einer modifizierten Version des Programms SEAL (Klebe *et al.*, 1999) durchgeführten Überlagerung der Strukturen und der Anwendung der CoMSIA-Methode unter Verwendung eines sterischen, elektrostatischen sowie hydrophoben Feldes. Die damit bei der Vorhersage für 15 Testverbindungen erzielten Ergebnisse ($r^2_{pred}$ = 0.56) können nach den in Kap. 4.8.2 aufgeführten Kriterien als „gut" bezeichnet werden. In einer Folgearbeit (Klebe &

Abraham, 1999) (‚Klebe(II)') wurden zusätzliche Felder zur Beschreibung von Wasserstoffbrückendonator und -akzeptoreigenschaften eingeführt; sie brachten allerdings keinen Gewinn der statistischen Signifikanz des damit erhaltenen Modells. Die mit dem in dieser Arbeit entwickelten Verfahren erzielten Ergebnisse zeigen eine im Vergleich zu den vorherigen Verfahren bessere Statistik. Sieht man von dem $q^2$-Wert in der ersten Spalte ab (s.o.), so ergibt sich hier der dafür beste Wert von 0.62. Daß dieses Modell tatsächlich zur Vorhersage von Bindungsaffinitäten geeignet ist, wird insbesondere auch durch den erzielten $r^2_{pred}$-Wert von 0.69 deutlich.

Welche Gründe können dafür angeführt werden? Sicher spielt die Berücksichtigung der Proteinumgebung in dem hier verwendeten Verfahren eine maßgebliche Rolle, werden doch Beiträge von zum Lösemittel hin orientierten Gruppen der Liganden von vornherein geringer bzw. gar nicht berücksichtigt, da bei ausreichendem Abstand der Gitterpunkte vom Protein der mit Gl. 45 dort errechnete Potentialwert gegen Null geht. Bei den Verfahren ohne Berücksichtigung der Rezeptorstruktur werden jedoch alle Atome des Liganden gleich behandelt. Eine weitere Ursache mag in der gegenseitigen Unabhängigkeit der verwendeten Wechselwirkungsfelder gesehen werden (s.a. Kap. 5.7.2). Während bei der Verwendung *generischer* Felder die Hinzunahme eines neuen Eigenschaftsfeldes nicht notwendigerweise zu einer Informationszunahme führen muß, trifft dies für die hier entwickelte Methode nicht zu. Schließlich kann auch die Verwendung der wissensbasierten Paarpräferenzen als vorteilhaft angesehen werden. Wie in Kap. 4.6.2 und 4.7 gezeigt, vermögen sie auch ohne vorherige PLS-Analyse, Bindungsaffinitäten und ‚*hot spots*' vorherzusagen.

# 6 Zusammenfassung und Ausblick

## *6.1 Zusammenfassung*

In dieser Arbeit ist die Entwicklung einer neuen, wissensbasierten, allgemein anwendbaren und schnell berechenbaren Bewertungsfunktion zur Vorhersage von Bindungsmodus und Affinität von Protein-Ligand-Komplexen vorgestellt worden. Zusätzlich wurde ihre Eignung zur graphischen Identifikation günstiger Wechselwirkungsbereiche in Proteinbindetaschen und eine Methode zur proteinspezifischen Adaptierung dieser Funktion unter Verwendung struktureller und energetischer Zusatzinformationen aufgezeigt.

Im Rahmen eines wissensbasierten Ansatzes wurden zur Ableitung der Bewertungsfunktion mit der Datenbank ReLiBase extrahierte Strukturinformationen aus kristallographisch bestimmten Protein-Ligand-Komplexen in atombasierte, statistische Präferenzen umgewandelt (Kap. 4.1 und 5.3). Die Notwendigkeit, sowohl *spezifische Wechselwirkungen* als auch hauptsächlich entropisch bedingte *Lösemittelbeiträge* bei der Protein-Ligand-Bindung berücksichtigen zu müssen, führte dabei zur Verwendung von zwei Termen, einer distanzabhängigen *Paarpräferenz* (Kap. 4.2 und 5.1) sowie einer von der Lösemittel-zugänglichen Oberfläche abhängigen *Einteilchenpräferenz* (Kap. 4.3 und 5.2). Eine Bewertung für gegebene 3D-Anordnungen von Protein und Ligand ergibt sich daraus durch Summation aller Paarwechselwirkungs- sowie Einteilchenbeiträge der Protein- und Ligandatome. Unter Verwendung eines approximierenden, auf einem Gitteralgorithmus beruhenden Verfahrens zur Berechnung der Lösemittel-zugänglichen Oberfläche wird je Protein-Ligand-Konfiguration eine *Rechenzeit* von nur etwa einer Sekunde benötigt.

Bei der Ableitung der statistischen Präferenzen hat sich die Wahl des jeweiligen *Referenzzustandes* als bedeutsam für die in den Präferenzen enthaltene Information und damit auch für das Ergebnis der Struktur- und Affinitätsvorhersagen gezeigt. Während im Fall der Paarpräferenzen ein auf einer kompakten Protein-Ligand-Konfiguration mit nichtspezifischen Wechselwirkungen beruhender Zustand gewählt wurde, wurde für die von der Lösemittel-zugänglichen Oberfläche abhängigen Einteilchenpräferenzen eine vollständige Separation von Protein und Ligand als Bezugspunkt verwendet. Weiterhin ist die Behandlung von in den *experimentellen* Daten auftretenden *Ungenauigkeiten* sowie eine adäquate Berücksichtigung von Verteilungen mit zu *geringer Anzahl an Beobachtungen* notwendig. Bedingt durch die Beschränkung auf Nichtwasserstoffatome während der Ableitung der Präferenzen sind außer-

dem für die Anwendung der Bewertungsfunktion keinerlei Annahmen über im Protein-Ligand-Komplex auftretende *Protonierungszustände* notwendig.

Die Untersuchung der Eignung dieser Funktion zur *Identifikation nativ-ähnlicher Protein-Ligand-Anordnungen* (Kap. 5.4) in einer Menge davon abweichender Alternativen erfolgte an Datensätzen von 91 bzw. 68 Komplexen, für die mit dem Programm FlexX bis zu 500 Protein-Ligand-Konfigurationen erzeugt wurden sowie an 100 Komplexen, bei denen die Konfigurationserzeugung mit dem Programm DOCK durchgeführt wurde. Im Fall der FlexX-generierten Datensätze ergibt sich dabei eine Erkennungsrate für am besten bewertete Ligandanordnungen mit einem *rmsd*-Wert < 2.0 Å von der Kristallstruktur in 80 % der möglichen Fälle. Bezogen auf den 91 Komplexe umfassenden Datensatz bedeutet dies eine Verbesserung um 35 % gegenüber den mit FlexX erzielten Ergebnissen. Im Fall der mit DOCK unter Verwendung der „Chemischen Bewertungsfunktion“ erzeugten Rezeptor-Ligand-Anordnungen beträgt der Anteil korrekt erkannter nativ-ähnlicher Konfigurationen 70 %, was sogar eine 46 %ige Verbesserung gegenüber den mit DOCK erzielten Ergebnissen bedeutet. Eine solch zuverlässige Erkennung nativ-ähnlicher gedockter Protein-Ligand-Anordnungen, die als notwendige Voraussetzung für eine verläßliche Vorhersage von Bindungsaffinitäten im Rahmen des strukturbasierten Designs von Liganden angesehen werden kann, ist bislang noch von keinem anderen Verfahren berichtet worden.

Die Validierung der entwickelten Bewertungsfunktion zur *Vorhersage von Bindungsaffinitäten* (Kap. 5.5) erfolgte an sechs Datensätzen mit kristallographisch bestimmten Protein-Ligand-Komplexen sowie drei Datensätzen, für die Protein-Ligand-Anordnungen mit FlexX erzeugt wurden. Im Fall von 16 Serinprotease-Inhibitor-Komplexen mit experimentell bestimmter Struktur wird dabei ein $R^2$-Wert von 0.86 und eine Standardabweichung von 0.95 logarithmischen Einheiten verglichen mit den experimentell bestimmten Affinitäten erzielt. Für einen Satz von 64 in das jeweilige Protein gedockten Thrombin- und Trypsininhibitoren ergibt sich ein $R^2$-Wert von 0.48 und eine Standardabweichung von 0.71 logarithmischen Einheiten. Ein Vergleich der erzielten Ergebnisse mit denen von vier der Literatur entnommenen Verfahren zeigt, daß insbesondere bei der Priorisierung *verschiedener* Liganden gegenüber *einem* Protein die hier entwickelte Methode deutlich geringere Standardabweichungen zwischen experimentellen und berechneten Bindungsaffinitäten ergibt. Die Eignung des Verfahrens, im Rahmen von *virtual screening*-Ansätzen „aktive“ Verbindungen in einer Menge von „inaktiven“ erkennen zu können, wird unter Verwendung von 31 Thrombin- und Tryp-

sinliganden mit bekannter Affinität sowie gegen 800 aus dem *Available Chemicals Directory* extrahierten Verbindungen getestet, die in die beiden Proteine mit FlexX gedockt wurden. Hierbei kann eine 5- bzw. 23-fache Steigerung des Anreicherungsfaktors bei Betrachtung der ersten 10 % der mit der entwickelten Funktion bewerteten Moleküle gegenüber den mit FlexX erzielten Ergebnissen erreicht werden.

Unter Verwendung der distanzabhängigen Paarpräferenzen gelingt die *Identifikation von für einen Ligandatomtyp günstigen Bereichen* (*hot spots*) (Kap. 4.7 und 5.6) in Proteinbindetaschen durch Berechnung von Potentialwerten für an Gitterpunkten lokalisierte Sondenatome und Visualisierung der Ergebnisse in Form von Isokonturoberflächen. Bei Verwendung von 159 Protein-Ligand-Komplexen sowie fünf Sondenatomtypen kann zudem für Lösemittelunzugängliche Ligandatome in 74 % der Fälle der korrekte Atomtyp und in sogar 85 % der Fälle ein für die korrekte Wechselwirkung passender Atomtyp vorhergesagt werden. Die erzielten Ergebnisse demonstrieren nicht nur, daß die zusammengesetzte Repräsentation der (an sich kugelsymmetrischen) Paarpräferenzen in einer Bindetasche auch gerichtete Wechselwirkungen korrekt beschreiben können, sondern sind auch mit den Resultaten anderer Methoden vergleichbar, die auf 3D-Verteilungen beruhende Wahrscheinlichkeitsdichten für Protein-Ligand-Wechselwirkungen verwenden.

Der Frage, inwieweit vorhandene strukturelle und energetische Informationen für *ein* betrachtetes Protein zur *Steigerung der Vorhersagegenauigkeit von Bindungsaffinitäten* neuer Verbindungen herangezogen werden können (Kap. 4.8 und 5.7), wurde im letzten Teil der Arbeit anhand eines Datensatzes aus insgesamt 76 Thermolysininhibitoren nachgegangen. Hierzu werden ebenfalls an Gitterpunkten in der Bindetasche von Thermolysin berechnete Potentialwerte mit Hilfe einer *Partial-Least-Squares*-Analyse sowie unter Verwendung der bekannten Anordnungen der 61 Verbindungen des Trainingsdatensatzes relativ zum Protein an experimentell bestimmte Affinitätswerte angepaßt. Für 15 Verbindungen eines Testdatensatzes kann so eine bemerkenswerte Erhöhung der Vorhersagegenauigkeit erzielt wird, verglichen mit den Affinitätsvorhersagen basierend auf den ursprünglichen Paarpräferenzen. Unter Verwendung von Teilmengen der Trainingsdaten kann zudem demonstriert werden, daß schon mit einem Zehntel der ursprünglichen Verbindungsanzahl von 61 eine signifikante Steigerung der Verläßlichkeit der berechneten Werte erhalten wird. Die entwickelte Methode ist daher auch in der frühen Phase eines Ligandfindungsprozesses bei nur beschränkt vorliegender Zusatzinformation einsetzbar.

## *6.2 Ausblick*

Die für die Ableitung der statistischen Präferenzen in dieser Arbeit zugrundegelegte Wissensbasis bestand aus Protein-Ligand-Komplexen der PDB. Obwohl damit die für die Problemstellung – die Vorhersage von Struktur und Affinität von Ligand-Rezeptor-Komplexen – naheliegenste Informationsquelle verwendet wurde, führt sie doch zwangsläufig zu Einschränkungen. Die immer noch stark limitierte Anzahl verfügbarer Komplexe erlaubt nur die Definition einer begrenzten Anzahl von Ligandatomtypen, und für viele der so erhaltenen Subsysteme stehen auch dann noch keine ausreichenden Anzahlen an Beobachtungen zur Verfügung. Ein möglicher Ausweg eröffnet sich bei der Verwendung von Informationen über nichtbindende Wechselwirkungen in Kristallpackungen von kleinen organischen Molekülen, wie sie aus der CSD (Allen *et al.*, 1991) extrahiert werden können. Hierbei ist jedoch insbesondere zu untersuchen, welchen Einfluß z.B. unterschiedliche Kristallisationsbedingungen auf den Verlauf der erhaltenen Präferenzen ausüben; eine notwendige Skalierung der erhaltenen Daten hinsichtlich der aus der PDB gewonnenen Verteilungen kann dabei nicht ausgeschlossen werden.

Die Verwendung von ausschließlich Nichtwasserstoffatomen im Rahmen dieser Arbeit erweist sich – nach den erzielten Ergebnissen zu urteilen – in der überwiegenden Mehrzahl aller Fälle als nicht nachteilhaft. Dennoch wäre zu untersuchen, ob sich durch Hinzunahme von Informationen über Wasserstoffatome insbesondere an polaren Atomen die Vorhersagegenauigkeiten bezüglich Struktur und Affinität noch weiter steigern lassen. Insbesondere durch die Klassifizierung von Atomtypen mit unterschiedlichen Wechselwirkungscharakteristika unter einen Obertyp auftretende Mehrdeutigkeiten (wie z.B. bei der Zusammenfassung von Sauerstoffatomen in Ether- bzw. Hydroxylgruppen) sollten hierbei wegfallen. Auch hierfür wären insbesondere aus der CSD (Allen *et al.*, 1991) gewonnene Moleküldaten geeignet. Allerdings tritt bei der Anwendung solcher Präferenzen das bislang nur unvollständig gelöste Problem auf, Protonierungszustände ionisierbarer Gruppen von Proteinen und Liganden im gebundenen Zustand korrekt (und mit ausreichender Geschwindigkeit) vorhersagen zu müssen.

Obwohl durch den gewählten wissensbasierten Ansatz durch das umgebende Lösemittel bedingte Beiträge in den abgeleiteten Präferenzen mitberücksichtigt sind, stellt sich die Frage, inwiefern das auch für Wassermoleküle gilt, die unmittelbar Wechselwirkungs-vermittelnde Funktionen zwischen Protein und Ligand ausüben. Im Rahmen des Protein-Ligand-Dockings bedeutet dieses, daß ein Algorithmus zur Plazierung von Wassermolekülen mit einer zuver-

lässigen Bewertung der so erzeugten Wasseranordnungen gekoppelt werden müßte. In Vorversuchen konnten unter Verwendung des oben vorgestellten Formalismus zur Ableitung distanzabhängiger Paarpräferenzen und einem zusätzlich eingeführten Atomtyp „Wasser“ Potentiale abgeleitet werden, die eine Identifizierung von günstigen Wasserpositionen im Bindungsepitop ermöglichen.

Die implizite Berücksichtigung von gerichteten Wechselwirkungen durch eine zusammengesetzte Repräsentation der Paarpräferenzen in der Bindetasche führt zu der Frage, ob durch eine mit diesem „Kraftfeld“ erfolgende „Energie“-Minimierung die aus Docking-Programmen erhaltenen Protein-Ligand-Anordnungen weiter strukturell optimiert werden können. Zusätzlich wäre zu hinterfragen, ob damit die Identifikation von nativ-ähnlichen Protein-Ligand-Konfigurationen gegenüber solchen mit einer vom Experiment stark abweichenden Anordnung verläßlicher würde. Bei diesbezüglichen Vorversuchen, in denen unter Verwendung ebenfalls empirisch bestimmter Torsionswinkelpotentiale (Klebe & Mietzner, 1994) sowie intramolekularer van-der-Waals-Potentiale eine zusätzliche Flexibilität des Liganden im Torsionswinkelraum ermöglicht wurde, konnte solches für einzelne Protein-Ligand-Kombinationen bestätigt werden.

Das hohe Maß an Übereinstimmung zwischen vorhergesagten und im Experiment tatsächlich gefundenen, für Atome eines speziellen Typs günstigen Bereichen in der Bindetasche wirft die Frage auf, ob die erhaltene Funktion nicht nur zur *Bewertung*, sondern – etwa in Form der Potentialfelder – auch schon zur *Plazierung* von Liganden im Rahmen eines Dokking-Verfahrens verwendet werden kann. Dieses könnte im Sinne einer Überlagerung der Potentialfelder in der Bindetasche mit einer adäquaten Feldrepräsentation des (flexibel gehaltenen) Liganden geschehen. Für eine Untersuchung, inwieweit die erhaltene Funktion schon einen Einfluß auf die Ligandplazierung alleine durch die Bewertung von Teillösungen im Verlauf eines inkrementellen Aufbaualgorithmus‘ nimmt, wird die Berechnung der statistischen Präferenzen derzeit in FlexX implementiert.

# Anhang

## *Programmiertechnische Hilfsmittel*

Die Programmteile zur Ableitung der Potentiale sowie zur Berechnung der Bewertungen für gegebene Protein-Ligand-Konfigurationen, zur Vorhersage „ausgezeichneter“ Bereiche in einer Proteinbindetasche sowie zur proteinspezifischen Anpassung der Bewertungsfunktion wurden in der Programmiersprache C++ erstellt. Aufbauend auf einer hierarchischen Klassenstruktur wurden die einzelnen Aufgabenstellungen in eigenständigen Einheiten organisiert (Abb. 45).

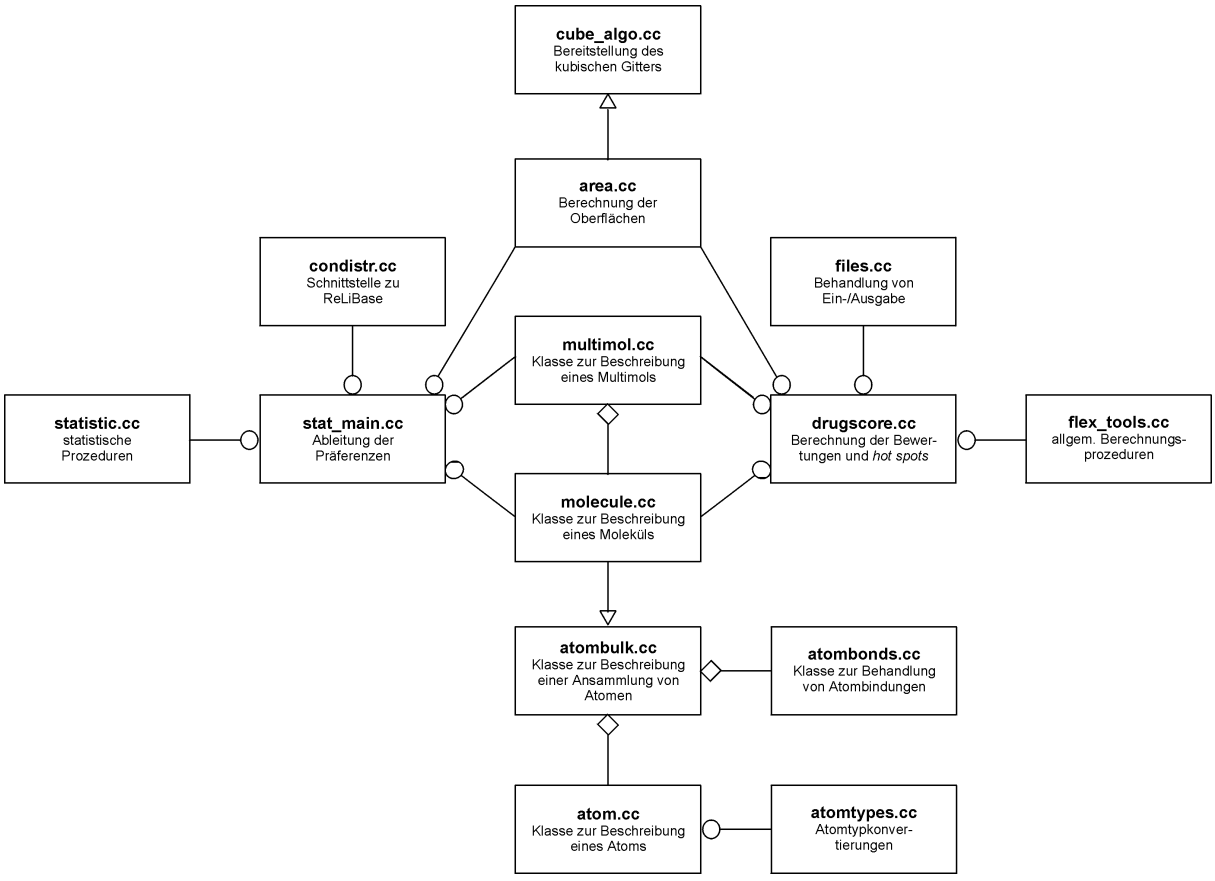



**Abb. 45:** Klassen- und Programmstruktur zur Ableitung der statistischen Präferenzen und zur Berechnung der Bewertungen und *hot spots*. Die Notation der Beziehungen zwischen den Untereinheiten folgt der von Booch (Booch, 1994) eingeführten (○——: Assoziationsbeziehung (‚*uses*‘); ◇——: Aggregationsbeziehung (‚*has*’); ——▷: Vererbungsbeziehung (‚*is*’)).

Die Entwicklung erfolgte mit dem MIPSpro C++-Compiler von SGI (Version 7.3.1.1) unter dem Betriebssytem IRIX auf einer SGI O2 (R5000-Prozessor, 195 MHz) sowie mit dem GNU C/C++-Compiler (Version 2.91.66) unter Linux auf einem Pentium-II-PC (450 MHz). Auf ersterer Maschine wurde auch die Ableitung der statistischen Präferenzen durchgeführt, auf den letzteren die Berechnungen der Bewertungen, *hot spots* und PLS-Analysen.

***Verwendete Programme***

- Die Ableitung der Potentiale erfolgte unter Verwendung der Datenbank ReLiBase (Version II.1998) (Hemm *et al.*, 1995; Hendlich, 1998).
- Das Docking der Liganden erfolgte mit den Programmen FlexX (Version 1.65) (Rarey *et al.*, 1995; Rarey *et al.*, 1996a) sowie DOCK (Version 4.0.1) (Kuntz *et al.*, 1982; Makino & Kuntz, 1997; Meng *et al.*, 1992).
- Für die Aufbereitung der Datensätze und die Visualisierung der Ergebnisse wurde das Programmpaket SYBYL (Version 6.5) (SYBYL) verwendet. Die Suche des Datensatzes der „Inaktiven“ im Rahmen des *virtual screening*-Ansatzes wurde mit dem Programm UNITY (Version 4.1) (UNITY) durchgeführt.
- Für die Minimierung der Thermolysin-Inhibitoren zur proteinspezifischen Anpassung der Bewertungsfunktion wurde das im Programm MOLOC implementierte Kraftfeld MAB (Gerber & Müller, 1995) verwendet.

## *Ergebnistabellen*

**Tab. 36:** Docking-Ergebnisse für 100 Protein-Ligand-Komplexe (DOCK_DS) unter Verwendung des rigiden Dockings in DOCK.

| PDB-Code | rig_cnt | | rig_min_cnt | | rig_nrg | | rig_min_nrg | | rig_chm | | rig_min_chm | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] |
| 1abe | 0.80 | 1.88 | 0.80 | 3.25 | 0.80 | 1.49 | 0.70 | 1.71 | 0.80 | 1.49 | 0.53 | 1.71 |
| 1abf | 0.92 | 2.20 | 1.01 | 4.49 | 0.92 | 2.17 | 0.69 | 3.97 | 0.92 | 2.17 | 0.74 | 3.97 |
| 1acj | 0.83 | 5.14 | 0.78 | 4.79 | 0.83 | 1.94 | 0.25 | 0.57 | 0.83 | 1.94 | 0.35 | 2.17 |
| 1ack | 0.32 | 4.24 | 0.39 | 4.56 | 0.32 | 0.32 | 0.16 | 0.16 | 0.32 | 0.32 | 0.22 | 3.75 |
| 1ase | 1.64 | 6.35 | 1.76 | 4.06 | 1.64 | 6.35 | 1.62 | 3.94 | 1.64 | 6.35 | 1.90 | 3.36 |
| 1azm | 0.81 | 1.57 | 0.77 | 2.54 | 0.81 | 2.54 | 0.93 | 1.56 | 0.81 | 5.65 | 1.08 | 5.82 |
| 1blh | 1.37 | 6.14 | 1.55 | 3.84 | 1.37 | 3.45 | 2.16 | 4.23 | 1.37 | 3.45 | 2.03 | 3.37 |
| 1cbx | 0.97 | 5.80 | 0.89 | 1.03 | 0.97 | 5.80 | 0.63 | 5.95 | 0.97 | 8.79 | 0.46 | 7.55 |
| 1cde | 1.72 | 8.25 | 0.83 | 7.38 | 1.72 | 6.88 | 0.46 | 3.94 | 1.72 | 6.88 | 0.50 | 11.08 |
| 1cil | 1.24 | 6.89 | 1.16 | 1.16 | 1.24 | 6.89 | 1.60 | 1.94 | 1.24 | 7.94 | 0.73 | 6.55 |
| 1cps | 1.04 | 1.29 | 0.34 | 3.20 | 1.04 | 1.29 | 0.33 | 0.77 | 1.04 | 9.61 | 0.46 | 0.91 |
| 1ctr | 1.55 | 7.75 | 1.39 | 8.31 | 1.55 | 7.75 | 0.98 | 7.16 | 1.55 | 8.68 | 1.27 | 8.40 |
| 1dbm | 1.46 | 9.17 | 0.69 | 6.04 | 1.46 | 9.17 | 0.45 | 0.79 | 1.46 | 9.17 | 0.32 | 1.56 |
| 1did | 1.60 | 4.33 | 1.54 | 5.33 | 1.60 | 4.33 | 1.04 | 5.29 | 1.60 | 8.70 | 1.24 | 5.69 |
| 1die | 1.16 | 4.42 | 0.83 | 4.35 | 1.16 | 2.20 | 0.65 | 2.53 | 1.16 | 6.80 | 0.67 | 4.91 |
| 1dr1 | 0.96 | 0.96 | 0.88 | 3.24 | 0.96 | 0.96 | 0.61 | 0.65 | 0.96 | 2.68 | 0.51 | 6.54 |
| 1dwd | 1.40 | 10.08 | 0.77 | 4.37 | 1.40 | 10.08 | 1.01 | 14.75 | 1.40 | 10.08 | 1.15 | 14.35 |
| 1ela | 1.34 | 2.49 | 1.80 | 2.31 | 1.34 | 2.49 | 0.91 | 0.91 | 1.34 | 2.49 | 1.66 | 9.34 |
| 1frp | 1.44 | 14.05 | 0.90 | 4.14 | 1.44 | 8.83 | 1.29 | 8.23 | 1.44 | 10.48 | 1.51 | 7.66 |
| 1ghb | 1.37 | 6.00 | 1.39 | 6.24 | 1.37 | 4.56 | 0.87 | 5.91 | 1.37 | 4.56 | 0.74 | 5.95 |
| 1hfc | 1.56 | 7.58 | 1.31 | 1.50 | 1.56 | 1.56 | 0.62 | 0.64 | 1.56 | 7.58 | 0.50 | 1.24 |
| 1hgi | 1.11 | 6.80 | 0.63 | 1.35 | 1.11 | 9.11 | 0.80 | 1.92 | 1.11 | 9.11 | 1.15 | 1.98 |
| 1hgj | 1.68 | 7.94 | 1.14 | 7.82 | 1.68 | 7.94 | 1.23 | 1.50 | 1.68 | 12.68 | 2.00 | 3.21 |
| 1hsl | 0.47 | 1.87 | 0.47 | 1.59 | 0.47 | 1.63 | 0.36 | 0.49 | 0.47 | 1.63 | 0.43 | 0.81 |
| 1hti | 0.96 | 4.48 | 0.63 | 3.54 | 0.96 | 4.45 | 0.52 | 4.48 | 0.96 | 3.62 | 0.46 | 5.14 |
| 1hyt | 1.43 | 5.27 | 1.37 | 9.21 | 1.43 | 8.03 | 0.29 | 5.31 | 1.43 | 8.03 | 0.56 | 8.89 |
| 1icn | 1.42 | 2.68 | 1.18 | 1.60 | 1.42 | 2.08 | 1.01 | 1.70 | 1.42 | 2.08 | 1.00 | 1.57 |
| 1imb | 1.51 | 4.20 | 0.87 | 4.47 | 1.51 | 5.13 | 1.27 | 5.63 | 1.51 | 5.68 | 1.30 | 5.27 |
| 1ivc | 1.70 | 4.18 | 1.38 | 5.48 | 1.70 | 2.46 | 0.52 | 2.15 | 1.70 | 3.97 | 1.17 | 2.50 |
| 1ivd | 0.37 | 5.26 | 1.45 | 4.62 | 0.37 | 5.55 | 0.61 | 3.24 | 0.37 | 2.91 | 0.86 | 3.42 |
| 1ive | 1.42 | 3.86 | 1.46 | 6.64 | 1.42 | 3.47 | 1.17 | 16.94 | 1.42 | 2.69 | 1.10 | 19.49 |
| 1ivf | 1.48 | 4.27 | 1.21 | 4.79 | 1.48 | 4.27 | 1.38 | 4.06 | 1.48 | 6.02 | 0.57 | 4.09 |
| 1lah | 0.39 | 0.41 | 0.27 | 0.34 | 0.39 | 0.41 | 0.45 | 0.47 | 0.39 | 0.41 | 0.48 | 0.48 |
| 1lcp | 0.58 | 3.60 | 1.00 | 5.18 | 0.58 | 3.00 | 0.43 | 2.69 | 0.58 | 3.03 | 1.20 | 11.02 |
| 1lic | 1.19 | 11.19 | 0.70 | 3.29 | 1.19 | 11.19 | 0.60 | 8.79 | 1.19 | 12.79 | 0.74 | 9.38 |
| 1lna | 1.47 | 5.74 | 1.09 | 3.39 | 1.47 | 7.77 | 0.59 | 7.98 | 1.47 | 9.70 | 0.55 | 8.59 |
| 1lst | 0.73 | 1.00 | 0.42 | 0.42 | 0.73 | 0.90 | 0.40 | 0.57 | 0.73 | 0.90 | 0.50 | 0.58 |
| 1mld | 1.08 | 4.06 | 1.16 | 6.12 | 1.08 | 3.89 | 1.32 | 3.88 | 1.08 | 3.89 | 1.24 | 3.88 |
| 1mrg | 0.43 | 3.75 | 0.24 | 0.72 | 0.43 | 3.15 | 0.50 | 0.54 | 0.43 | 4.69 | 0.49 | 5.49 |
| 1mrk | 0.64 | 3.10 | 0.31 | 1.67 | 0.64 | 2.93 | 0.62 | 1.21 | 0.64 | 3.17 | 0.82 | 1.26 |
| 1mup | 0.69 | 3.51 | 1.13 | 4.36 | 0.69 | 4.20 | 0.51 | 4.16 | 0.69 | 4.22 | 0.33 | 4.45 |
| 1nis | 0.73 | 3.82 | 0.68 | 3.80 | 0.73 | 3.82 | 0.41 | 3.84 | 0.73 | 3.82 | 0.57 | 5.43 |
| 1nsc | 1.24 | 4.57 | 0.89 | 4.27 | 1.24 | 5.74 | 1.03 | 7.74 | 1.24 | 5.74 | 0.60 | 7.11 |
| 1pbd | 0.50 | 1.95 | 0.31 | 5.03 | 0.50 | 0.50 | 0.25 | 0.62 | 0.50 | 0.50 | 0.18 | 0.54 |
| 1phf | 0.71 | 2.51 | 1.16 | 9.94 | 0.71 | 5.82 | 0.98 | 9.45 | 0.71 | 1.81 | 0.87 | 7.56 |

Fortsetzung von Tab. 36:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1poc | 1.57 | 11.64 | 1.11 | 6.40 | 1.57 | 13.17 | 0.89 | 0.89 | 1.57 | 13.17 | 1.12 | 11.34 |
| 1pph | 1.82 | 8.59 | 1.85 | 6.25 | 1.82 | 8.59 | 1.30 | 1.48 | 1.82 | 8.59 | 1.85 | 9.15 |
| 1ppl | 0.48 | 0.48 | 0.32 | 0.91 | 0.48 | 0.48 | 0.36 | 0.40 | 0.48 | 0.48 | 0.44 | 0.44 |
| 1pso | 1.78 | 11.98 | 0.45 | 0.45 | 1.78 | 15.19 | 0.25 | 0.25 | 1.78 | 15.19 | 0.46 | 13.34 |
| 1rds | 1.81 | 10.33 | 2.07 | 6.87 | 1.81 | 3.00 | 1.06 | 1.06 | 1.81 | 12.55 | 0.54 | 12.01 |
| 1rnt | 1.31 | 4.39 | 0.39 | 0.62 | 1.31 | 6.83 | 0.77 | 11.10 | 1.31 | 9.92 | 1.65 | 12.65 |
| 1rob | 1.79 | 6.08 | 1.96 | 7.14 | 1.79 | 10.59 | 0.69 | 0.69 | 1.79 | 10.59 | 0.81 | 0.81 |
| 1snc | 1.41 | 7.96 | 1.46 | 1.46 | 1.41 | 1.41 | 0.57 | 0.57 | 1.41 | 1.41 | 0.64 | 0.64 |
| 1srj | 0.81 | 10.23 | 0.72 | 0.72 | 0.81 | 3.50 | 0.56 | 1.20 | 0.81 | 7.42 | 0.69 | 7.72 |
| 1tdb | 1.63 | 8.38 | 1.57 | 7.70 | 1.63 | 2.04 | 1.04 | 1.04 | 1.63 | 2.49 | 1.72 | 8.76 |
| 1thy | 2.50 | 11.97 | 1.79 | 1.94 | 2.50 | 5.17 | 1.57 | 4.00 | 2.50 | 11.97 | 1.93 | 3.31 |
| 1tlp | 1.76 | 14.52 | 0.77 | 7.52 | 1.76 | 14.52 | 0.60 | 0.87 | 1.76 | 14.52 | 0.54 | 7.08 |
| 1tng | 0.47 | 2.86 | 0.97 | 2.08 | 0.47 | 0.89 | 0.39 | 0.50 | 0.47 | 1.62 | 0.33 | 0.67 |
| 1tnh | 0.97 | 2.49 | 0.78 | 3.90 | 0.97 | 2.49 | 0.38 | 0.48 | 0.97 | 2.48 | 0.42 | 0.44 |
| 1tni | 1.17 | 2.08 | 1.26 | 2.34 | 1.17 | 2.13 | 1.34 | 2.14 | 1.17 | 2.13 | 1.30 | 1.91 |
| 1tpp | 1.41 | 5.99 | 1.50 | 2.95 | 1.41 | 5.77 | 2.17 | 16.33 | 1.41 | 6.92 | 2.20 | 14.27 |
| 1ukz | 0.96 | 11.91 | 0.88 | 6.73 | 0.96 | 11.89 | 0.73 | 11.61 | 0.96 | 11.89 | 0.63 | 12.02 |
| 1ulb | 0.92 | 4.45 | 0.61 | 4.28 | 0.92 | 6.35 | 0.77 | 7.47 | 0.92 | 7.07 | 0.88 | 7.18 |
| 1wap | 0.79 | 10.95 | 0.79 | 2.86 | 0.79 | 10.75 | 0.45 | 0.66 | 0.79 | 10.75 | 0.37 | 0.57 |
| 1xid | 1.05 | 4.27 | 0.81 | 4.43 | 1.05 | 4.03 | 0.71 | 2.84 | 1.05 | 3.64 | 0.93 | 2.03 |
| 1xie | 0.99 | 3.23 | 0.33 | 5.09 | 0.99 | 3.09 | 0.49 | 2.75 | 0.99 | 5.02 | 0.71 | 6.40 |
| 2ada | 0.82 | 11.42 | 0.48 | 0.48 | 0.82 | 9.04 | 0.48 | 0.70 | 0.82 | 9.04 | 0.46 | 7.31 |
| 2ak3 | 1.01 | 1.44 | 0.90 | 12.21 | 1.01 | 1.44 | 0.64 | 1.31 | 1.01 | 11.24 | 0.95 | 11.13 |
| 2cgr | 4.55 | 11.50 | 4.14 | 6.22 | 4.55 | 11.50 | 4.51 | 5.58 | 4.55 | 11.50 | 4.73 | 8.40 |
| 2cht | 0.86 | 0.86 | 0.38 | 3.99 | 0.86 | 1.94 | 0.59 | 0.77 | 0.86 | 1.94 | 0.79 | 0.81 |
| 2cmd | 0.49 | 0.49 | 0.35 | 7.61 | 0.49 | 3.17 | 0.47 | 0.47 | 0.49 | 3.17 | 0.49 | 0.49 |
| 2cpp | 1.57 | 5.91 | 1.44 | 6.00 | 1.57 | 5.16 | 1.20 | 6.08 | 1.57 | 2.81 | 0.97 | 2.88 |
| 2gbp | 0.97 | 12.07 | 0.13 | 3.34 | 0.97 | 13.28 | 0.41 | 3.83 | 0.97 | 13.28 | 0.16 | 13.56 |
| 2lgs | 1.21 | 4.53 | 1.20 | 3.89 | 1.21 | 5.03 | 0.79 | 4.82 | 1.21 | 5.03 | 0.75 | 4.82 |
| 2mcp | 1.38 | 5.11 | 0.90 | 5.52 | 1.38 | 3.28 | 0.53 | 2.37 | 1.38 | 4.09 | 0.69 | 2.02 |
| 2mth | 1.33 | 12.04 | 1.25 | 13.75 | 1.33 | 5.04 | 0.70 | 1.06 | 1.33 | 1.93 | 0.53 | 0.59 |
| 2pk4 | 1.32 | 5.47 | 1.09 | 1.60 | 1.32 | 1.96 | 0.85 | 0.91 | 1.32 | 2.83 | 1.06 | 1.21 |
| 2r04 | 0.61 | 18.44 | 0.82 | 12.45 | 0.61 | 21.88 | 0.34 | 12.83 | 0.61 | 21.88 | 0.45 | 16.40 |
| 2r07 | 1.16 | 13.00 | 0.30 | 12.10 | 1.16 | 13.00 | 0.55 | 12.53 | 1.16 | 13.00 | 0.52 | 11.16 |
| 2sim | 1.39 | 5.73 | 0.52 | 0.52 | 1.39 | 6.50 | 0.51 | 0.51 | 1.39 | 5.93 | 0.82 | 1.02 |
| 2tmn | 0.80 | 9.70 | 0.46 | 3.62 | 0.80 | 7.96 | 0.63 | 4.84 | 0.80 | 6.98 | 0.41 | 6.22 |
| 2xis | 1.06 | 4.87 | 0.70 | 2.57 | 1.06 | 4.99 | 0.69 | 2.38 | 1.06 | 4.99 | 0.62 | 5.08 |
| 2yhx | 1.87 | 6.42 | 1.13 | 1.22 | 1.87 | 6.42 | 0.77 | 5.68 | 1.87 | 5.34 | 1.41 | 3.41 |
| 2ypi | 0.58 | 4.48 | 0.55 | 2.84 | 0.58 | 4.47 | 0.65 | 4.86 | 0.58 | 6.33 | 0.45 | 4.27 |
| 3aah | 1.03 | 1.32 | 0.51 | 0.70 | 1.03 | 1.03 | 0.38 | 6.97 | 1.03 | 1.03 | 0.30 | 14.56 |
| 3cpa | 5.89 | 8.23 | 5.97 | 7.99 | 5.89 | 8.85 | 7.25 | 8.26 | 5.89 | 9.77 | 6.75 | 11.96 |
| 4cts | 0.76 | 3.70 | 0.58 | 5.42 | 0.76 | 3.81 | 0.35 | 0.83 | 0.76 | 2.68 | 0.80 | 0.96 |
| 4est | 1.53 | 12.12 | 1.28 | 7.64 | 1.53 | 12.12 | 2.86 | 18.98 | 1.53 | 12.12 | 2.30 | 11.27 |
| 4fab | 1.39 | 9.17 | 1.12 | 5.98 | 1.39 | 6.62 | 0.85 | 4.78 | 1.39 | 13.20 | 0.97 | 4.22 |
| 4fbp | 1.46 | 9.82 | 0.94 | 7.29 | 1.46 | 5.73 | 0.96 | 1.17 | 1.46 | 4.74 | 0.85 | 10.75 |
| 4phv | 3.51 | 14.17 | 3.32 | 13.99 | 3.51 | 15.03 | 3.91 | 15.34 | 3.51 | 15.03 | 5.86 | 16.99 |
| 4tim | 1.44 | 4.26 | 0.68 | 4.50 | 1.44 | 1.84 | 0.68 | 1.44 | 1.44 | 1.84 | 0.60 | 1.34 |
| 4tln | 1.17 | 3.19 | 0.97 | 5.12 | 1.17 | 3.97 | 0.89 | 4.21 | 1.17 | 10.26 | 0.94 | 10.15 |
| 5abp | 0.50 | 2.72 | 0.22 | 2.67 | 0.50 | 2.71 | 0.41 | 3.78 | 0.50 | 2.71 | 0.25 | 3.83 |
| 5p2p | 1.29 | 14.73 | 0.60 | 7.05 | 1.29 | 14.73 | 1.05 | 7.26 | 1.29 | 14.73 | 0.88 | 13.90 |
| 6abp | 0.80 | 3.21 | 0.54 | 3.72 | 0.80 | 1.67 | 0.39 | 1.92 | 0.80 | 1.67 | 0.57 | 1.79 |
| 6rnt | 1.89 | 5.77 | 2.39 | 4.12 | 1.89 | 5.46 | 1.50 | 6.30 | 1.89 | 8.06 | 1.87 | 5.26 |
| 6tmn | 1.59 | 10.32 | 1.31 | 12.43 | 1.59 | 9.72 | 1.48 | 10.07 | 1.59 | 10.32 | 1.48 | 6.63 |
| 7tim | 0.74 | 3.74 | 0.56 | 2.06 | 0.74 | 3.80 | 0.62 | 6.00 | 0.74 | 6.16 | 0.50 | 5.79 |
| 8atc | 1.14 | 4.85 | 0.97 | 4.69 | 1.14 | 4.85 | 0.45 | 1.12 | 1.14 | 4.85 | 0.60 | 1.19 |

a) Bei der Strukturgenerierung überhaupt gefundener kleinster *rmsd*-Wert (in Å) bezüglich der Eingabestruktur. b) *rmsd*-Wert (in Å) bezüglich der Eingabestruktur für die Protein-Ligand-Anordnung, die bei Bewertung mit der jeweiligen Funktion auf Rang 1 gefunden wurde.

**Tab. 37:** Docking-Ergebnisse für 100 Protein-Ligand-Komplexe (DOCK_DS) unter Verwendung des flexiblen Dockings in DOCK.

| PDB-Code | flex_cnt | | flex_min_cnt | | flex_nrg | | flex_min_nrg | | flex_chm | | flex_min_chm | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] | D[a)] | R[b)] |
| 1abe | 1.24 | 1.88 | 1.30 | 3.00 | 1.49 | 1.88 | 0.90 | 1.81 | 1.49 | 1.57 | 0.56 | 1.66 |
| 1abf | 2.20 | 2.20 | 2.16 | 4.51 | 0.92 | 2.17 | 0.73 | 2.22 | 0.92 | 2.17 | 0.85 | 1.36 |
| 1acj | 0.83 | 5.14 | 0.95 | 4.79 | 0.83 | 1.94 | 0.25 | 0.57 | 0.83 | 1.94 | 0.66 | 2.17 |
| 1ack | 0.58 | 0.60 | 1.07 | 4.53 | 0.58 | 0.60 | 0.21 | 0.78 | 0.58 | 0.60 | 0.66 | 3.78 |
| 1ase | 2.50 | 2.50 | 2.50 | 2.50 | 2.55 | 3.00 | 1.92 | 4.42 | 2.55 | 3.00 | 3.75 | 4.03 |
| 1azm | 1.52 | 1.56 | 1.83 | 2.69 | 1.05 | 1.94 | 1.37 | 2.25 | 1.05 | 1.05 | 2.02 | 2.15 |
| 1blh | 4.33 | 6.23 | 3.22 | 6.65 | 4.33 | 5.05 | 3.68 | 5.79 | 3.90 | 4.18 | 3.08 | 4.41 |
| 1cbx | 2.07 | 7.93 | 2.75 | 4.28 | 0.52 | 10.13 | 0.98 | 1.26 | 0.57 | 12.16 | 1.36 | 10.30 |
| 1cde | 0.97 | 7.52 | 2.10 | 5.93 | 0.85 | 7.52 | 2.55 | 7.04 | 0.85 | 7.52 | 5.54 | 6.49 |
| 1cil | 3.16 | 3.16 | 1.29 | 3.25 | 3.07 | 3.97 | 1.18 | 2.64 | 3.16 | 3.62 | 0.84 | 4.82 |
| 1cps | 1.62 | 6.43 | 2.83 | 8.37 | 4.82 | 4.82 | 0.88 | 1.49 | 4.91 | 9.03 | 0.79 | 1.88 |
| 1ctr | 2.84 | 10.53 | 5.66 | 6.70 | 2.84 | 4.38 | 2.40 | 7.07 | 3.18 | 3.27 | 2.65 | 8.78 |
| 1dbm | 2.72 | 8.72 | 1.20 | 1.59 | 1.45 | 7.83 | 0.92 | 7.22 | 2.72 | 7.83 | 1.57 | 1.95 |
| 1did | 2.77 | 4.36 | 2.01 | 10.63 | 2.77 | 4.06 | 1.26 | 5.11 | 2.77 | 5.58 | 4.42 | 4.76 |
| 1die | 2.33 | 4.11 | 2.49 | 9.85 | 2.20 | 4.11 | 2.19 | 3.05 | 2.20 | 3.90 | 3.15 | 3.30 |
| 1dr1 | 1.15 | 2.66 | 1.49 | 5.27 | 0.72 | 10.20 | 1.16 | 2.65 | 2.29 | 5.16 | 4.31 | 7.04 |
| 1dwd | 2.21 | 11.04 | 3.87 | 9.10 | 2.70 | 6.65 | 12.15 | 14.48 | 2.70 | 3.27 | 9.36 | 14.79 |
| 1ela | 6.43 | 7.33 | 4.96 | 5.41 | 6.43 | 8.92 | 4.74 | 4.95 | 5.29 | 6.65 | 5.44 | 5.95 |
| 1frp | 1.02 | 9.80 | 1.76 | 5.58 | 4.02 | 7.98 | 1.45 | 4.05 | 5.35 | 7.98 | 4.12 | 8.35 |
| 1ghb | 1.46 | 4.43 | 1.28 | 3.93 | 1.34 | 3.75 | 0.97 | 1.00 | 1.34 | 3.75 | 1.00 | 1.00 |
| 1hfc | 4.85 | 9.36 | 5.57 | 7.27 | 4.08 | 8.72 | 4.16 | 6.58 | 4.83 | 8.90 | 5.66 | 16.04 |
| 1hgi | 2.04 | 5.41 | 2.47 | 10.92 | 2.97 | 8.29 | 2.31 | 4.92 | 2.97 | 4.07 | 2.43 | 5.40 |
| 1hgj | 3.32 | 5.74 | 2.05 | 10.59 | 2.74 | 5.74 | 1.47 | 1.50 | 3.32 | 6.37 | 2.04 | 3.23 |
| 1hsl | 0.44 | 2.97 | 1.85 | 3.80 | 0.32 | 3.80 | 1.17 | 1.59 | 0.32 | 3.28 | 1.00 | 1.90 |
| 1hti | 3.23 | 4.01 | 1.89 | 3.82 | 3.36 | 4.05 | 1.61 | 1.61 | 3.96 | 6.38 | 3.76 | 5.65 |
| 1hyt | 1.24 | 13.01 | 2.50 | 9.10 | 1.18 | 2.13 | 1.34 | 5.19 | 1.05 | 13.78 | 1.70 | 7.21 |
| 1icn | 2.11 | 4.95 | 3.82 | 8.52 | 2.35 | 3.80 | 2.07 | 3.87 | 2.17 | 10.54 | 2.14 | 8.36 |
| 1imb | 3.07 | 6.38 | 3.14 | 3.57 | 3.03 | 5.00 | 2.91 | 5.78 | 3.03 | 5.00 | 3.83 | 5.64 |
| 1ivc | 2.39 | 4.68 | 1.48 | 6.50 | 1.80 | 1.80 | 1.20 | 1.48 | 1.80 | 1.80 | 2.49 | 4.69 |
| 1ivd | 0.37 | 5.82 | 1.95 | 5.14 | 0.50 | 0.50 | 1.01 | 3.14 | 2.78 | 2.78 | 3.11 | 3.91 |
| 1ive | 2.71 | 4.38 | 3.15 | 6.82 | 2.91 | 9.33 | 2.10 | 6.78 | 2.82 | 9.33 | 2.44 | 16.73 |
| 1ivf | 0.84 | 5.14 | 1.84 | 4.69 | 2.48 | 3.31 | 0.95 | 1.49 | 2.48 | 9.61 | 2.04 | 4.17 |
| 1lah | 1.08 | 3.76 | 1.22 | 1.53 | 1.33 | 2.14 | 0.89 | 1.33 | 1.33 | 2.49 | 0.78 | 1.26 |
| 1lcp | 0.54 | 3.40 | 1.66 | 4.84 | 0.54 | 3.40 | 2.08 | 2.35 | 2.09 | 3.40 | 5.54 | 10.24 |
| 1lic | 2.43 | 8.04 | 7.32 | 7.49 | 2.67 | 2.78 | 2.16 | 2.18 | 2.67 | 2.89 | 3.26 | 4.77 |
| 1lna | 0.59 | 6.65 | 2.79 | 9.90 | 0.59 | 6.48 | 3.12 | 5.04 | 0.59 | 12.08 | 1.77 | 11.05 |
| 1lst | 2.52 | 6.35 | 1.14 | 6.13 | 1.34 | 2.26 | 0.45 | 1.10 | 1.34 | 2.26 | 0.35 | 1.22 |
| 1mld | 1.55 | 2.44 | 2.26 | 6.62 | 1.38 | 1.38 | 2.17 | 3.39 | 1.38 | 1.38 | 2.09 | 3.49 |
| 1mrg | 1.51 | 3.75 | 0.24 | 3.01 | 0.66 | 3.15 | 0.50 | 0.54 | 0.66 | 4.69 | 4.38 | 5.49 |
| 1mrk | 2.36 | 2.65 | 1.37 | 2.96 | 2.36 | 2.93 | 1.19 | 2.54 | 2.36 | 2.62 | 1.79 | 2.12 |
| 1mup | 1.13 | 3.35 | 2.08 | 5.31 | 0.68 | 1.82 | 1.62 | 5.09 | 0.77 | 1.82 | 0.53 | 1.76 |
| 1nis | 1.69 | 4.75 | 1.83 | 3.04 | 2.23 | 4.37 | 2.96 | 3.70 | 2.23 | 4.37 | 2.93 | 2.93 |
| 1nsc | 2.91 | 2.91 | 2.55 | 6.09 | 2.91 | 6.78 | 7.33 | 8.47 | 2.91 | 6.78 | 7.12 | 8.98 |
| 1pbd | 0.57 | 3.78 | 0.87 | 3.82 | 0.43 | 1.76 | 0.39 | 0.39 | 0.43 | 1.76 | 0.30 | 0.30 |
| 1phf | 1.81 | 2.51 | 2.14 | 9.94 | 1.81 | 5.82 | 3.58 | 9.45 | 0.90 | 1.81 | 1.03 | 7.56 |
| 1poc | 3.43 | 9.88 | 5.77 | 8.67 | 3.48 | 4.06 | 4.20 | 8.38 | 4.06 | 4.06 | 4.95 | 8.59 |
| 1pph | 3.99 | 7.15 | 3.26 | 6.56 | 3.99 | 7.61 | 4.06 | 4.08 | 5.84 | 7.61 | 5.86 | 6.51 |
| 1ppl | 8.54 | 13.36 | 6.55 | 6.99 | 5.37 | 5.37 | 6.18 | 6.18 | 7.92 | 13.57 | 10.61 | 12.03 |
| 1pso | 7.51 | 21.78 | 10.48 | 10.70 | 8.16 | 21.73 | 8.62 | 12.56 | 8.16 | 18.91 | 12.11 | 12.40 |
| 1rds | 2.09 | 3.31 | 5.40 | 6.14 | 2.72 | 3.43 | 1.59 | 1.61 | 3.43 | 5.34 | 7.12 | 9.13 |
| 1rnt | 2.20 | 4.18 | 1.21 | 3.78 | 2.20 | 7.66 | 5.31 | 13.26 | 4.00 | 7.66 | 6.42 | 11.89 |

Fortsetzung von Tab. 37:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1rob | 2.97 | 10.79 | 3.81 | 10.46 | 1.60 | 8.37 | 1.05 | 1.18 | 1.61 | 6.58 | 2.87 | 5.97 |
| 1snc | 1.24 | 8.23 | 2.55 | 4.41 | 0.45 | 2.06 | 1.38 | 5.74 | 2.06 | 2.06 | 1.46 | 1.61 |
| 1srj | 1.17 | 1.17 | 0.85 | 1.23 | 1.17 | 3.20 | 1.00 | 1.01 | 3.20 | 4.84 | 1.92 | 7.14 |
| 1tdb | 2.23 | 3.63 | 3.38 | 9.26 | 2.30 | 2.58 | 0.66 | 2.73 | 1.79 | 2.71 | 1.60 | 1.98 |
| 1thy | 2.17 | 4.66 | 2.40 | 6.43 | 2.07 | 2.07 | 1.94 | 2.31 | 2.07 | 12.06 | 1.98 | 2.20 |
| 1tlp | 3.46 | 9.81 | 5.58 | 11.22 | 3.46 | 4.45 | 5.52 | 6.16 | 3.69 | 9.83 | 6.60 | 9.52 |
| 1tng | 0.76 | 2.22 | 1.14 | 1.98 | 0.48 | 1.04 | 0.59 | 1.10 | 0.48 | 2.08 | 0.20 | 0.36 |
| 1tnh | 1.09 | 4.21 | 2.08 | 4.51 | 0.76 | 1.79 | 0.76 | 0.78 | 0.76 | 2.03 | 0.37 | 0.67 |
| 1tni | 3.20 | 4.82 | 2.08 | 5.47 | 3.41 | 3.88 | 1.16 | 2.23 | 2.09 | 3.88 | 1.75 | 2.18 |
| 1tpp | 1.92 | 2.09 | 2.46 | 3.11 | 1.17 | 1.40 | 2.93 | 9.97 | 3.65 | 3.65 | 7.09 | 12.62 |
| 1ukz | 4.38 | 11.95 | 1.30 | 10.64 | 5.66 | 11.95 | 1.08 | 1.85 | 5.60 | 11.77 | 2.67 | 13.19 |
| 1ulb | 1.29 | 4.45 | 0.61 | 4.28 | 1.29 | 6.35 | 2.95 | 7.47 | 1.28 | 7.07 | 2.90 | 7.18 |
| 1wap | 1.07 | 11.32 | 1.89 | 3.00 | 1.07 | 11.15 | 0.55 | 0.66 | 2.44 | 11.15 | 0.86 | 2.03 |
| 1xid | 0.78 | 3.72 | 1.92 | 3.38 | 0.78 | 3.72 | 2.63 | 2.93 | 0.78 | 2.06 | 3.13 | 4.66 |
| 1xie | 1.63 | 3.47 | 0.97 | 3.24 | 2.54 | 3.53 | 2.68 | 4.58 | 1.84 | 3.53 | 6.57 | 8.44 |
| 2ada | 0.90 | 6.02 | 0.86 | 0.86 | 1.92 | 8.57 | 0.64 | 0.90 | 1.92 | 9.13 | 6.50 | 6.59 |
| 2ak3 | 2.54 | 4.44 | 2.98 | 3.99 | 4.44 | 8.44 | 1.33 | 3.52 | 4.92 | 8.23 | 3.22 | 7.62 |
| 2cgr | 5.72 | 8.35 | 6.70 | 7.12 | 5.72 | 8.26 | 6.48 | 7.28 | 2.48 | 2.53 | 5.29 | 7.32 |
| 2cht | 0.91 | 0.91 | 1.12 | 3.31 | 0.91 | 0.91 | 1.42 | 1.42 | 0.91 | 4.53 | 1.49 | 1.49 |
| 2cmd | 1.29 | 2.68 | 1.69 | 3.39 | 1.08 | 2.24 | 1.09 | 1.49 | 1.08 | 2.24 | 1.13 | 1.41 |
| 2cpp | 1.86 | 5.54 | 2.47 | 6.00 | 1.57 | 5.16 | 2.53 | 6.08 | 1.57 | 2.81 | 0.97 | 2.88 |
| 2gbp | 1.56 | 3.24 | 1.00 | 4.29 | 1.78 | 3.24 | 0.83 | 3.03 | 1.78 | 13.55 | 2.93 | 14.09 |
| 2lgs | 1.68 | 2.73 | 1.78 | 5.11 | 1.01 | 1.01 | 1.64 | 4.99 | 1.01 | 1.01 | 2.66 | 6.39 |
| 2mcp | 4.12 | 5.82 | 3.88 | 4.87 | 3.55 | 5.40 | 1.03 | 1.85 | 3.55 | 3.83 | 1.24 | 1.91 |
| 2mth | 1.50 | 3.45 | 2.15 | 5.66 | 1.18 | 2.77 | 1.56 | 4.89 | 1.50 | 13.69 | 1.34 | 1.98 |
| 2pk4 | 4.07 | 7.47 | 4.61 | 5.16 | 1.79 | 1.79 | 1.04 | 1.72 | 1.79 | 1.79 | 1.21 | 1.60 |
| 2r04 | 1.69 | 10.26 | 1.75 | 9.54 | 0.98 | 0.98 | 1.39 | 1.54 | 0.98 | 10.06 | 1.50 | 12.50 |
| 2r07 | 1.16 | 1.16 | 2.06 | 11.81 | 1.16 | 1.16 | 0.89 | 1.17 | 3.10 | 10.88 | 1.25 | 12.32 |
| 2sim | 3.43 | 8.35 | 4.32 | 5.35 | 3.59 | 7.98 | 1.54 | 1.57 | 4.40 | 6.02 | 4.04 | 5.14 |
| 2tmn | 5.47 | 7.98 | 2.96 | 10.49 | 2.42 | 7.67 | 3.43 | 4.67 | 2.42 | 7.54 | 4.79 | 6.58 |
| 2xis | 1.46 | 1.60 | 1.92 | 9.50 | 1.75 | 2.65 | 2.02 | 3.17 | 1.75 | 8.18 | 2.80 | 4.55 |
| 2yhx | 0.90 | 0.90 | 2.47 | 4.78 | 0.90 | 0.90 | 1.79 | 2.70 | 2.56 | 6.16 | 3.12 | 3.73 |
| 2ypi | 3.38 | 3.38 | 1.96 | 4.19 | 3.38 | 6.48 | 1.06 | 4.73 | 2.94 | 6.48 | 1.64 | 4.55 |
| 3aah | 1.67 | 5.65 | 0.64 | 0.68 | 1.65 | 5.49 | 0.80 | 0.80 | 1.65 | 5.49 | 0.54 | 13.79 |
| 3cpa | 9.16 | 10.64 | 7.86 | 11.57 | 9.16 | 10.36 | 8.25 | 13.20 | 10.13 | 10.72 | 11.33 | 11.94 |
| 4cts | 1.64 | 5.15 | 2.59 | 3.65 | 1.59 | 1.82 | 0.67 | 3.59 | 1.64 | 1.84 | 1.23 | 3.65 |
| 4est | 3.21 | 10.82 | 6.48 | 7.59 | 1.97 | 2.74 | 5.78 | 13.68 | 2.74 | 3.05 | 9.04 | 14.70 |
| 4fab | 2.23 | 12.19 | 1.78 | 5.14 | 1.02 | 4.74 | 0.93 | 5.45 | 1.02 | 12.19 | 1.02 | 1.12 |
| 4fbp | 2.24 | 9.42 | 2.19 | 6.31 | 2.24 | 9.42 | 1.08 | 2.22 | 4.23 | 14.43 | 4.38 | 4.93 |
| 4phv | 2.40 | 11.80 | 7.66 | 11.34 | 2.40 | 7.33 | 5.56 | 5.88 | 3.94 | 14.22 | 6.83 | 11.74 |
| 4tim | 3.89 | 4.95 | 2.04 | 4.11 | 3.32 | 3.94 | 1.39 | 1.56 | 3.32 | 4.71 | 1.40 | 6.37 |
| 4tln | 2.26 | 3.43 | 3.57 | 7.83 | 2.78 | 9.20 | 2.31 | 4.33 | 2.78 | 8.80 | 2.40 | 4.50 |
| 5abp | 0.77 | 0.77 | 1.36 | 4.09 | 0.99 | 2.77 | 0.57 | 1.16 | 0.99 | 2.77 | 0.60 | 3.82 |
| 5p2p | 4.07 | 9.11 | 5.43 | 8.72 | 7.00 | 10.03 | 6.96 | 9.85 | 5.43 | 8.52 | 7.06 | 7.08 |
| 6abp | 1.15 | 3.21 | 2.04 | 2.38 | 0.80 | 3.26 | 0.74 | 1.82 | 0.80 | 1.67 | 0.62 | 1.79 |
| 6rnt | 1.31 | 5.30 | 2.04 | 7.45 | 1.31 | 7.20 | 1.30 | 6.86 | 1.19 | 6.34 | 1.59 | 5.10 |
| 6tmn | 2.28 | 6.27 | 3.11 | 8.74 | 4.63 | 6.25 | 3.78 | 3.95 | 2.18 | 11.85 | 3.37 | 3.68 |
| 7tim | 2.98 | 5.46 | 1.71 | 4.74 | 2.91 | 5.89 | 1.39 | 1.85 | 2.91 | 6.34 | 3.30 | 6.15 |
| 8atc | 2.53 | 8.03 | 2.48 | 4.88 | 2.53 | 2.53 | 2.31 | 4.86 | 2.53 | 2.53 | 1.81 | 5.26 |

a) Bei der Strukturgenerierung überhaupt gefundener kleinster *rmsd*-Wert (in Å) bezüglich der Eingabestruktur. b) *rmsd*-Wert (in Å) bezüglich der Eingabestruktur für die Protein-Ligand-Anordnung, die bei Bewertung mit der jeweiligen Funktion auf Rang 1 gefunden wurde.

**Tab. 38:** Docking-Ergebnisse für 91 Protein-Ligand-Komplexe (FlexX_DS1) unter Verwendung von FlexX.

| PDB-Code | D[a)] | R[b)] | PDB-Code | D[a)] | R[b)] |
|---|---|---|---|---|---|
| 1abe | 0.40 | 1.80 | 1rbp | 0.74 | 0.89 |
| 1abf | 0.45 | 3.05 | 1rds | 1.18 | 4.92 |
| 1atl | 0.67 | 1.30 | 1rnt | 0.96 | 1.87 |
| 1azm | 1.06 | 2.95 | 1rob | 1.57 | 7.95 |
| 1bbp | 1.97 | 2.28 | 1slt | 1.42 | 1.61 |
| 1cbx | 0.89 | 5.21 | 1snc | 2.60 | 6.70 |
| 1cde | 1.13 | 1.53 | 1srj | 2.90 | 8.03 |
| 1cil | 1.53 | 3.97 | 1tlp | 2.41 | 2.46 |
| 1com | 1.00 | 1.23 | 1tng | 0.46 | 2.02 |
| 1cps | 0.64 | 4.85 | 1tnh | 0.54 | 1.48 |
| 1ctr | 1.69 | 3.16 | 1tni | 0.47 | 2.70 |
| 1did | 2.69 | 3.55 | 1tpp | 1.44 | 2.66 |
| 1die | 2.21 | 4.59 | 1ukz | 0.96 | 10.84 |
| 1dr1 | 0.85 | 1.48 | 1wap | 0.21 | 0.57 |
| 1dwc | 0.99 | 8.86 | 1xid | 0.52 | 4.14 |
| 1dwd | 1.00 | 1.20 | 1xie | 0.51 | 3.74 |
| 1ela | 1.56 | 1.63 | 2ada | 0.58 | 0.60 |
| 1epb | 1.78 | 3.06 | 2ak3 | 0.87 | 2.52 |
| 1frp | 0.60 | 1.91 | 2cgr | 0.92 | 0.99 |
| 1ghb | 1.32 | 11.88 | 2cht | 1.49 | 4.57 |
| 1hfc | 0.62 | 2.11 | 2cmd | 0.58 | 1.79 |
| 1hgj | 2.19 | 3.74 | 2cpp | 0.39 | 2.91 |
| 1hsl | 0.60 | 0.60 | 2gbp | 0.37 | 0.97 |
| 1hyt | 0.56 | 1.45 | 2mth | 0.94 | 4.28 |
| 1icn | 2.19 | 10.19 | 2pk4 | 0.84 | 1.37 |
| 1imb | 0.63 | 5.09 | 2sim | 0.52 | 1.24 |
| 1ivc | 1.12 | 1.78 | 2tmn | 0.81 | 5.26 |
| 1ivd | 0.87 | 5.59 | 2xis | 0.96 | 4.51 |
| 1ive | 1.21 | 2.56 | 2ypi | 0.66 | 1.34 |
| 1ivf | 0.89 | 8.20 | 3aah | 0.57 | 1.04 |
| 1lah | 0.29 | 0.29 | 3cpa | 0.71 | 2.53 |
| 1lcp | 0.80 | 4.86 | 3hvt | 3.53 | 10.20 |
| 1lic | 2.68 | 6.01 | 4fbp | 1.57 | 1.77 |
| 1lna | 0.81 | 1.14 | 4hmg | 0.80 | 0.89 |
| 1lst | 0.44 | 0.82 | 4phv | 2.48 | 4.25 |
| 1mld | 0.47 | 1.71 | 4tim | 0.83 | 3.99 |
| 1mrg | 0.40 | 1.10 | 4tln | 0.92 | 3.67 |
| 1mrk | 2.85 | 2.89 | 4ts1 | 0.66 | 1.45 |
| 1nis | 0.89 | 1.41 | 5abp | 0.35 | 0.79 |
| 1nsc | 0.85 | 1.30 | 5p2p | 0.72 | 1.01 |
| 1pbd | 0.45 | 0.48 | 6abp | 0.39 | 1.69 |
| 1phf | 0.79 | 4.23 | 6rnt | 2.31 | 6.79 |
| 1poc | 2.08 | 3.52 | 6tmn | 2.27 | 4.46 |
| 1pph | 2.44 | 4.42 | 7tim | 0.53 | 1.47 |
| 1ppl | 2.30 | 5.71 | 8atc | 0.58 | 0.97 |
| 1pso | 1.74 | 1.98 | - | - | - |

a) Bei der Strukturgenerierung überhaupt gefundener kleinster *rmsd*-Wert (in Å) bezüglich der Kristallstruktur. b) *rmsd*-Wert (in Å) bezüglich der Kristallstruktur für die Protein-Ligand-Anordnung, die bei Bewertung mit der jeweiligen Funktion auf Rang 1 gefunden wurde.

**Tab. 39:** Docking-Ergebnisse für 68 Protein-Ligand-Komplexe (FlexX_DS2) unter Verwendung von FlexX.

| PDB-Code | D[a)] | R[b)] | PDB-Code | D[a)] | R[b)] |
|---|---|---|---|---|---|
| 121p | 0.91 | 0.92 | 1lpm | 6.09 | 6.75 |
| 1aaq | 0.81 | 1.71 | 1mbi | 0.28 | 0.28 |
| 1acm | 0.81 | 0.90 | 1mdr | 0.56 | 1.05 |
| 1aco | 0.35 | 0.75 | 1mmq | 4.18 | 11.80 |
| 1aec | 6.98 | 8.52 | 1nco | 6.30 | 9.14 |
| 1aha | 0.51 | 0.56 | 1phd | 0.33 | 0.77 |
| 1ake | 1.21 | 1.60 | 1phg | 4.49 | 4.60 |
| 1apt | 5.22 | 6.46 | 1ppc | 0.70 | 2.78 |
| 1avd | 0.73 | 1.30 | 1ppi | 6.51 | 6.75 |
| 1bma | 10.52 | 14.72 | 1ppk | 1.16 | 1.54 |
| 1byb | 1.09 | 1.57 | 1ppm | 5.39 | 5.97 |
| 1cbs | 1.23 | 1.39 | 1rne | 3.54 | 12.01 |
| 1cdg | 3.76 | 6.00 | 1tnk | 0.64 | 1.63 |
| 1coy | 0.93 | 1.18 | 1tnl | 0.65 | 0.71 |
| 1dbb | 0.81 | 0.81 | 1tph | 0.52 | 1.48 |
| 1eap | 3.65 | 3.95 | 1trk | 0.74 | 1.53 |
| 1eed | 10.76 | 12.94 | 2ctc | 0.63 | 1.98 |
| 1elb | 3.26 | 7.19 | 2er6 | 9.63 | 11.26 |
| 1elc | 4.02 | 4.74 | 3cla | 5.19 | 6.45 |
| 1eld | 5.15 | 6.98 | 3gch | 1.61 | 1.97 |
| 1ele | 4.35 | 10.72 | 3ptb | 0.56 | 0.60 |
| 1etr | 8.02 | 8.46 | 4dfr | 0.52 | 1.08 |
| 1fen | 1.39 | 1.39 | 4fxn | 0.41 | 0.41 |
| 1fkg | 5.06 | 5.91 | 4hvp | 13.03 | 13.47 |
| 1glp | 0.45 | 0.45 | 4phv | 2.48 | 4.25 |
| 1glq | 5.40 | 6.16 | 4tmn | 2.00 | 8.34 |
| 1hdc | 10.21 | 13.84 | 5cts | 6.47 | 6.99 |
| 1hef | 14.76 | 15.09 | 5tim | 0.86 | 1.99 |
| 1hvr | 3.00 | 10.15 | 5tmn | 4.54 | 4.76 |
| 1ida | 8.71 | 11.53 | 6cpa | 5.45 | 7.00 |
| 1igj | 4.65 | 7.18 | 6tim | 1.22 | 1.61 |
| 1ivb | 0.61 | 0.61 | 7cpa | 8.82 | 9.30 |
| 1ldm | 0.50 | 0.69 | 8gch | 6.04 | 7.41 |
| 1lmo | 4.76 | 4.91 | 9hvp | 13.57 | 13.90 |

a) Bei der Strukturgenerierung überhaupt gefundener kleinster *rmsd*-Wert (in Å) bezüglich der Kristallstruktur. b) *rmsd*-Wert (in Å) bezüglich der Kristallstruktur für die Protein-Ligand-Anordnung, die bei Bewertung mit der jeweiligen Funktion auf Rang 1 gefunden wurde.

**Tab. 40:** *rmsd*-Wert (in Å) bezüglich der Kristallstruktur der bei Bewertung mit der hier entwickelten Funktion auf dem ersten Rang erhaltenen Protein-Ligand-Anordnungen des Datensatzes DOCK_DS.

| PDB-Code | flex_min_nrg | | flex_min_chm | |
|---|---|---|---|---|
| | Ohne Kristallstruktur | Mit[a)] | Ohne Kristallstruktur | Mit[a)] |
| 1abe | 1.57 | 0.00 | 0.56 | 0.00 |
| 1abf | 3.26 | 0.00 | 1.00 | 0.00 |
| 1acj | 0.62 | 0.62 | 4.76 | 4.76 |
| 1ack | 3.83 | 3.83 | 0.71 | 0.71 |
| 1ase | 5.21 | 5.21 | 3.79 | 0.00 |
| 1azm | 5.86 | 5.86 | 5.87 | 5.87 |
| 1blh | 6.63 | 6.60 | 3.54 | 0.00 |
| 1cbx | 0.99 | 0.99 | 1.40 | 0.00 |
| 1cde | 7.12 | 7.12 | 9.32 | 0.00 |
| 1cil | 2.93 | 2.93 | 4.82 | 0.00 |
| 1cps | 1.45 | 1.45 | 1.64 | 0.00 |
| 1ctr | 7.17 | 7.17 | 2.65 | 0.00 |
| 1dbm | 1.33 | 0.00 | 1.59 | 0.00 |
| 1did | 10.67 | 0.00 | 10.74 | 0.00 |
| 1die | 10.32 | 0.00 | 10.66 | 10.66 |
| 1dr1 | 7.95 | 7.95 | 7.11 | 7.11 |
| 1dwd | 27.63 | 0.00 | 9.36 | 0.00 |
| 1ela | 5.00 | 0.00 | 10.10 | 0.00 |
| 1frp | 1.70 | 0.00 | 4.12 | 0.00 |
| 1ghb | 1.05 | 1.05 | 1.15 | 1.15 |
| 1hfc | 6.29 | 0.00 | 7.50 | 0.00 |
| 1hgi | 5.19 | 0.00 | 5.64 | 0.00 |
| 1hgj | 1.47 | 0.0 | 5.30 | 0.00 |
| 1hsl | 1.59 | 0.00 | 1.63 | 0.00 |
| 1hti | 4.46 | 4.40 | 4.43 | 0.00 |
| 1hyt | 1.88 | 0.00 | 1.70 | 0.00 |
| 1icn | 10.57 | 0.00 | 2.14 | 2.14 |
| 1imb | 7.49 | 7.49 | 6.91 | 0.00 |
| 1ivc | 4.89 | 4.89 | 2.72 | 2.72 |
| 1ivd | 2.66 | 2.66 | 3.44 | 0.00 |
| 1ive | 6.40 | 6.40 | 8.01 | 8.01 |
| 1ivf | 0.95 | 0.95 | 4.18 | 0.00 |
| 1lah | 1.20 | 0.00 | 1.26 | 0.00 |
| 1lcp | 11.10 | 0.00 | 9.05 | 0.00 |
| 1lic | 7.70 | 0.00 | 5.83 | 0.00 |
| 1lna | 10.20 | 10.20 | 4.86 | 4.86 |
| 1lst | 0.45 | 0.00 | 0.35 | 0.00 |
| 1mld | 3.22 | 0.00 | 2.18 | 0.00 |
| 1mrg | 0.54 | 0.00 | 5.67 | 0.00 |
| 1mrk | 2.48 | 2.48 | 2.25 | 2.25 |
| 1mup | 5.16 | 5.16 | 4.30 | 4.30 |
| 1nis | 4.19 | 0.00 | 3.05 | 0.00 |
| 1nsc | 8.50 | 0.00 | 7.92 | 0.00 |
| 1pbd | 0.39 | 0.00 | 1.70 | 0.00 |
| 1phf | 5.08 | 0.00 | 1.73 | 1.73 |
| 1poc | 8.40 | 0.00 | 8.48 | 0.00 |
| 1pph | 6.24 | 0.00 | 6.46 | 0.00 |
| 1ppl | 7.46 | 0.00 | 12.07 | 0.00 |

Fortsetzung von Tab. 40:

| | | | | |
|---|---|---|---|---|
| 1pso | 12.56 | 0.00 | 12.58 | 0.00 |
| 1rds | 1.61 | 0.00 | 12.48 | 0.00 |
| 1rnt | 7.25 | 0.00 | 7.05 | 0.00 |
| 1rob | 1.18 | 0.00 | 5.22 | 0.00 |
| 1snc | 5.74 | 5.74 | 8.39 | 0.00 |
| 1srj | 2.56 | 2.56 | 3.42 | 3.42 |
| 1tdb | 8.35 | 8.35 | 2.38 | 0.00 |
| 1thy | 2.54 | 0.00 | 1.98 | 0.00 |
| 1tlp | 7.71 | 0.00 | 8.12 | 0.00 |
| 1tng | 1.26 | 1.26 | 0.59 | 0.59 |
| 1tnh | 3.75 | 3.75 | 1.72 | 0.00 |
| 1tni | 3.34 | 3.34 | 2.12 | 2.12 |
| 1tpp | 3.39 | 0.00 | 7.94 | 0.00 |
| 1ukz | 5.88 | 0.00 | 6.76 | 0.00 |
| 1ulb | 7.17 | 0.00 | 2.90 | 0.00 |
| 1wap | 0.98 | 0.98 | 0.99 | 0.00 |
| 1xid | 2.70 | 2.70 | 10.79 | 10.79 |
| 1xie | 4.58 | 4.58 | 8.49 | 0.00 |
| 2ada | 0.91 | 0.00 | 6.64 | 0.00 |
| 2ak3 | 1.54 | 0.00 | 3.79 | 0.00 |
| 2cgr | 6.85 | 6.85 | 7.28 | 0.00 |
| 2cht | 3.14 | 0.00 | 1.49 | 0.00 |
| 2cmd | 1.32 | 0.00 | 1.27 | 0.00 |
| 2cpp | 2.53 | 0.00 | 3.21 | 3.21 |
| 2gbp | 3.79 | 0.00 | 3.85 | 0.00 |
| 2lgs | 4.50 | 0.00 | 4.81 | 0.00 |
| 2mcp | 1.84 | 1.84 | 1.91 | 1.91 |
| 2mth | 4.24 | 4.24 | 1.98 | 0.00 |
| 2pk4 | 1.41 | 1.41 | 1.40 | 1.40 |
| 2r04 | 1.60 | 1.60 | 1.50 | 1.50 |
| 2r07 | 9.45 | 0.00 | 8.68 | 0.00 |
| 2sim | 1.66 | 0.00 | 4.63 | 0.00 |
| 2tmn | 4.65 | 0.00 | 4.79 | 0.00 |
| 2xis | 2.79 | 0.00 | 10.65 | 0.00 |
| 2yhx | 8.89 | 8.89 | 7.34 | 0.00 |
| 2ypi | 1.44 | 1.44 | 1.64 | 1.64 |
| 3aah | 0.80 | 0.00 | 5.89 | 0.00 |
| 3cpa | 12.38 | 0.00 | 12.40 | 0.00 |
| 4cts | 3.59 | 3.59 | 3.66 | 0.00 |
| 4est | 10.34 | 10.34 | 17.69 | 0.00 |
| 4fab | 5.47 | 5.47 | 1.05 | 1.05 |
| 4fbp | 6.44 | 6.44 | 11.26 | 0.00 |
| 4phv | 6.17 | 0.00 | 9.24 | 0.00 |
| 4tim | 1.49 | 1.49 | 1.60 | 1.60 |
| 4tln | 4.36 | 0.00 | 2.40 | 0.00 |
| 5abp | 0.57 | 0.00 | 0.60 | 0.00 |
| 5p2p | 9.34 | 0.00 | 7.63 | 0.00 |
| 6abp | 3.47 | 0.00 | 0.62 | 0.00 |
| 6rnt | 6.58 | 0.00 | 1.63 | 0.00 |
| 6tmn | 3.78 | 0.00 | 4.20 | 0.00 |
| 7tim | 1.39 | 1.39 | 3.30 | 0.00 |
| 8atc | 2.41 | 0.00 | 1.81 | 0.00 |

a) Zu der von DOCK generierten Lösungsmenge wurde die Kristallstruktur hinzugefügt.

**Tab. 41:** *rmsd*-Wert (in Å) bezüglich der Kristallstruktur der bei Bewertung mit der hier entwickelten Funktion auf dem ersten Rang erhaltenen Protein-Ligand-Anordnungen des Datensatzes FlexX_DS1.

| PDB-Code | Ohne Kristallstruktur | Mit[a)] Kristallstruktur | PDB-Code | Ohne Kristallstruktur | Mit[a)] Kristallstruktur |
|---|---|---|---|---|---|
| 1abe | 0.40 | 0.00 | 1rbp | 0.74 | 0.00 |
| 1abf | 0.40 | 0.00 | 1rds | 1.29 | 0.00 |
| 1atl | 2.35 | 2.36 | 1rnt | 1.73 | 1.73 |
| 1azm | 1.47 | 1.40 | 1rob | 1.63 | 0.00 |
| 1bbp | 3.66 | 0.00 | 1slt | 1.68 | 0.00 |
| 1cbx | 0.81 | 0.81 | 1snc | 5.22 | 5.22 |
| 1cde | 7.45 | 7.45 | 1srj | 6.65 | 0.00 |
| 1cil | 2.04 | 2.06 | 1tlp | 3.84 | 3.84 |
| 1com | 1.09 | 1.05 | 1tng | 1.98 | 1.98 |
| 1cps | 0.66 | 0.64 | 1tnh | 0.55 | 0.00 |
| 1ctr | 7.47 | 7.45 | 1tni | 3.35 | 3.35 |
| 1did | 3.66 | 0.00 | 1tpp | 2.97 | 0.00 |
| 1die | 4.45 | 4.43 | 1ukz | 4.07 | 4.07 |
| 1dr1 | 1.48 | 0.00 | 1wap | 0.21 | 0.00 |
| 1dwc | 1.05 | 0.00 | 1xid | 1.95 | 1.95 |
| 1dwd | 1.65 | 0.00 | 1xie | 4.45 | 4.45 |
| 1ela | 11.66 | 0.00 | 2ada | 0.69 | 0.00 |
| 1epb | 2.76 | 2.76 | 2ak3 | 2.15 | 2.15 |
| 1frp | 5.41 | 0.00 | 2cgr | 0.92 | 0.00 |
| 1ghb | 1.39 | 0.00 | 2cht | 8.95 | 0.00 |
| 1hfc | 2.11 | 0.00 | 2cmd | 0.75 | 0.00 |
| 1hgj | 2.10 | 0.00 | 2cpp | 2.27 | 2.27 |
| 1hsl | 0.66 | 0.00 | 2gbp | 0.85 | 0.00 |
| 1hyt | 1.42 | 1.44 | 2mth | 1.59 | 1.59 |
| 1icn | 11.30 | 11.30 | 2pk4 | 1.31 | 1.31 |
| 1imb | 5.07 | 5.07 | 2sim | 1.33 | 1.33 |
| 1ivc | 2.07 | 2.07 | 2tmn | 1.45 | 1.45 |
| 1ivd | 5.43 | 5.43 | 2xis | 1.35 | 1.35 |
| 1ive | 1.46 | 1.46 | 2ypi | 1.05 | 1.05 |
| 1ivf | 1.85 | 1.85 | 3aah | 1.05 | 0.00 |
| 1lah | 1.09 | 0.00 | 3cpa | 1.17 | 0.00 |
| 1lcp | 2.58 | 2.58 | 3hvt | 3.53 | 0.00 |
| 1lic | 6.02 | 6.02 | 4fbp | 1.77 | 0.00 |
| 1lna | 0.87 | 0.87 | 4hmg | 0.89 | 0.00 |
| 1lst | 0.54 | 0.00 | 4phv | 4.25 | 0.00 |
| 1mld | 1.31 | 0.00 | 4tim | 1.18 | 1.18 |
| 1mrg | 0.40 | 0.00 | 4tln | 1.03 | 0.00 |
| 1mrk | 6.06 | 6.06 | 4ts1 | 1.69 | 1.69 |
| 1nis | 4.30 | 4.30 | 5abp | 0.51 | 0.00 |
| 1nsc2 | 1.28 | 1.28 | 5p2p | 0.87 | 0.00 |
| 1pbd | 0.45 | 0.00 | 6abp | 0.54 | 0.54 |
| 1phf | 1.39 | 0.00 | 6rnt | 8.31 | 0.00 |
| 1poc | 2.08 | 0.00 | 6tmn | 2.31 | 2.31 |
| 1pph | 4.69 | 0.00 | 7tim | 0.61 | 0.61 |
| 1ppl | 5.70 | 0.00 | 8atc | 0.99 | 0.00 |
| 1pso | 1.87 | 0.00 | - | - | - |

a) Zu der von FlexX generierten Lösungsmenge wurde die Kristallstruktur hinzugefügt.

**Tab. 42:** *rmsd*-Wert (in Å) bezüglich der Kristallstruktur der bei Bewertung mit der hier entwickelten Funktion auf dem ersten Rang erhaltenen Protein-Ligand-Anordnungen des Datensatzes FlexX_DS2.

| PDB-Code | Ohne Kristallstruktur | Mit[a)] Kristallstruktur | PDB-Code | Ohne Kristallstruktur | Mit[a)] Kristallstruktur |
|---|---|---|---|---|---|
| 121p | 1.64 | 0.00 | 1lpm | 7.05 | 0.00 |
| 1aaq | 0.82 | 0.82 | 1mbi | 0.46 | 0.46 |
| 1acm | 1.66 | 0.00 | 1mdr | 0.94 | 0.00 |
| 1aco | 1.58 | 1.58 | 1mmq | 6.31 | 0.00 |
| 1aec | 8.52 | 8.52 | 1nco | 6.37 | 0.00 |
| 1aha | 2.99 | 0.00 | 1phd | 0.64 | 0.64 |
| 1ake | 1.81 | 0.00 | 1phg | 4.60 | 0.00 |
| 1apt | 5.96 | 0.00 | 1ppc | 1.71 | 0.00 |
| 1avd | 1.19 | 1.19 | 1ppi | 6.95 | 0.00 |
| 1bma | 14.50 | 0.00 | 1ppk | 1.50 | 0.00 |
| 1byb | 1.13 | 0.00 | 1ppm | 6.11 | 0.00 |
| 1cbs | 1.23 | 0.00 | 1rne | 3.58 | 0.00 |
| 1cdg | 4.54 | 4.54 | 1tnk | 1.04 | 0.00 |
| 1coy | 1.06 | 0.00 | 1tnl | 0.67 | 0.00 |
| 1dbb | 0.81 | 0.81 | 1tph | 1.79 | 0.00 |
| 1eap | 3.95 | 0.00 | 1trk | 1.29 | 0.00 |
| 1eed | 12.94 | 0.00 | 2ctc | 0.63 | 0.00 |
| 1elb | 7.10 | 7.10 | 2er6 | 11.29 | 11.29 |
| 1elc | 4.02 | 0.00 | 3cla | 5.20 | 5.20 |
| 1eld | 6.46 | 0.00 | 3gch | 6.80 | 6.80 |
| 1ele | 11.24 | 0.00 | 3ptb | 0.60 | 0.00 |
| 1etr | 9.75 | 0.00 | 4dfr | 1.43 | 1.43 |
| 1fen | 1.47 | 0.00 | 4fxn | 0.41 | 0.00 |
| 1fkg | 5.98 | 5.98 | 4hvp | 13.48 | 0.00 |
| 1glp | 6.69 | 6.69 | 4phv | 4.25 | 0.00 |
| 1glq | 10.52 | 0.00 | 4tmn | 5.09 | 0.00 |
| 1hdc | 11.34 | 0.00 | 5cts | 7.11 | 7.11 |
| 1hef | 14.93 | 0.00 | 5tim | 1.72 | 1.72 |
| 1hvr | 3.65 | 0.00 | 5tmn | 5.31 | 5.31 |
| 1ida | 11.52 | 0.00 | 6cpa | 9.08 | 9.08 |
| 1igj | 7.01 | 0.00 | 6tim | 1.61 | 1.61 |
| 1ivb | 1.24 | 1.24 | 7cpa | 9.27 | 0.00 |
| 1ldm | 5.67 | 0.00 | 8gch | 6.04 | 0.00 |
| 1lmo | 4.91 | 4.91 | 9hvp | 13.90 | 0.00 |

a) Zu der von FlexX generierten Lösungsmenge wurde die Kristallstruktur hinzugefügt.

**Tab. 43:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 16 Serinproteasekomplexe (s.a. Tab. 6, S. 91).

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1bra | 3.21 | 1tmt | 7.65 |
| 1dwb | 3.34 | 1tng | 3.12 |
| 1dwd | 7.70 | 1tnh | 3.35 |
| 1etr | 6.54 | 1tni | 3.25 |
| 1ets | 8.38 | 1tnj | 3.56 |
| 1ett | 5.98 | 1tnk | 3.73 |
| 1ppc | 7.45 | 1tnl | 3.75 |
| 1pph | 6.25 | 3ptb | 4.04 |

**Tab. 44:** Nach Gl. 36 ($\gamma$= 0.09) und Gl. 42 berechnete Bindungsaffinitäten für 15 Metalloprotease-komplexe (s.a. Tab. 7, S. 91).

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1cbx | 6.63 | 4tmn | 10.81 |
| 1mnc | 8.34 | 5tln | 7.13 |
| 1tlp | 8.45 | 5tmn | 8.56 |
| 1tmn | 9.03 | 6cpa | 9.21 |
| 2tmn | 4.98 | 6tmn | 8.46 |
| 3cpa | 5.82 | 7cpa | 11.61 |
| 3tmn | 5.64 | 8cpa | 7.84 |
| 4tln | 4.14 | - | - |

**Tab. 45:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 11 Endothiapepsinkomplexe (s.a. Tab. 8, S. 91)

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1eed | 6.22 | 2er9 | 7.63 |
| 1epo | 6.63 | 3er3 | 7.30 |
| 1epp | 6.55 | 4er1 | 7.79 |
| 2er0 | 7.24 | 4er4 | 7.63 |
| 2er6 | 7.63 | 5er2 | 8.00 |
| 2er7 | 8.67 | - | - |

**Tab. 46:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 9 Arabinose-bindende Proteine enthaltende Komplexe (s.a. Tab. 9, S. 92).

| PDB-Code | $pK_i$ [a)] | PDB-Code | $pK_i$ [a)] |
|---|---|---|---|
| 1abe | 6.45 (6.39) | 6abp | 6.35 (6.26) |
| 1abf | 6.98 (7.06) | 7abp | 7.11 (7.06) |
| 1apb | 6.95 (7.04) | 8abp | 7.62 (7.66) |
| 1bap | 6.49 (6.36) | 9abp | 7.72 (7.73) |
| 5abp | 7.62 (7.60) | - | - |

a) Der Wert für das in der Kristallstruktur jeweils als zweites aufgeführte Kohlenhydratepimer ist in Klammern angegeben.

**Tab. 47:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 17 Komplexe aus dem kombinierten Trainings- und Testdatensatz der Arbeit von Böhm (Böhm, 1994), die nicht in Tab. 6 – Tab. 9 (S. 91 – 92) enthalten sind und eine Auflösung besser als 2.5 Å besitzen (s.a. Tab. 10, S. 92).

| PDB-Code | $pK_i$ | PDB-Code | $pK_i$ |
|---|---|---|---|
| 1fkf | 6.18 | 2tsc | 5.66 |
| 1mbi | 0.89 | 2xis | 2.91 |
| 1phf | 2.09 | 2ypi | 2.63 |
| 1phg | 3.12 | 4dfr | 6.72 |
| 1rbp | 6.11 | 4hvp | 10.05 |
| 1rne | 10.98 | 4phv | 12.19 |
| 2cpp | 1.96 | 5cna | 3.31 |
| 2gbp | 4.78 | 5cpp | 2.18 |
| 2ifb | 5.17 | - | - |

**Tab. 48:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 71 Komplexe aus dem kombinierten Trainings- und Testdatensatz der Arbeit von Böhm (Böhm, 1998), die auch in der PDB enthalten sind (s.a. Tab. 11, S. 92).

| **PDB-Code** | $pK_i$ | **PDB-Code** | $pK_i$ |
|---|---|---|---|
| 1acj | 5.04 | 2gbp | 5.51 |
| 1add | 5.96 | 2gpb | 4.10 |
| 1bzm | 2.64 | 2ifb | 5.96 |
| 1cbx | 4.97 | 2phh | 3.27 |
| 1cil | 3.73 | 2r04 | 7.77 |
| 1cps | 4.16 | 2tmn | 4.00 |
| 1ctt | 4.36 | 2tsc | 6.53 |
| 1dwb | 3.86 | 2xis | 3.36 |
| 1dwc | 7.14 | 2ypi | 3.03 |
| 1ela | 7.43 | 3cpa | 4.88 |
| 1elc | 8.16 | 3dfr | 8.44 |
| 1fkf | 7.13 | 3ptb | 4.67 |
| 1hpv | 9.24 | 3tpi | 6.75 |
| 1hvr | 13.18 | 4dfr | 7.75 |
| 1l83 | 2.31 | 4er4 | 12.39 |
| 1ldm | 0.78 | 4fab | 6.00 |
| 1mbi | 1.02 | 4gr1 | 1.45 |
| 1phe | 3.01 | 4hmg | 4.22 |
| 1phf | 2.41 | 4hvp | 11.60 |
| 1phg | 3.60 | 4phv | 14.06 |
| 1ppc | 8.61 | 4tln | 3.42 |
| 1pph | 7.22 | 4tmn | 8.83 |
| 1pso | 11.33 | 4ts1 | 3.74 |
| 1r09[a)] | 5.54 | 5cna | 3.81 |
| 1rbp | 7.05 | 5cpp | 2.51 |
| 1rne | 12.67 | 5tim | 0.69 |
| 1sbp | 1.78 | 5tln | 5.70 |
| 1sre | 5.91 | 5tmn | 6.83 |
| 1stp | 5.39 | 6acn | 1.90 |
| 1tlp | 6.32 | 6cpa | 6.70 |
| 1tmn | 7.28 | 6rsa | 4.48 |
| 1tnk | 4.31 | 7cpa | 8.83 |
| 1ulb | 3.55 | 7cpp | 2.05 |
| 2cpp | 2.26 | 9aat | 5.83 |
| 2ctc | 3.76 | 9hvp | 12.08 |
| 2er6 | 12.50 | - | - |

**Tab. 49:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten gegenüber Thrombin und Trypsin für 32 Inhibitoren aus der Arbeit von Obst *et al.* (Obst, 1997; Obst *et al.*, 1997) (s.a. Tab. 12, S. 94).

| **Bezeichnung** | $pK_i$ **(Thrombin)** | $pK_i$ **(Trypsin)** |
|---|---|---|
| UO54 | 5.30 | 4.82 |
| UO62a | 5.68 | 5.08 |
| UO62b | 5.20 | 4.21 |
| UO62c | 4.83 | 4.64 |
| UO62d | 4.71 | 4.32 |
| UO62e | 4.79 | 3.92 |
| UO62f | 5.97 | 5.44 |
| UO62g | 5.98 | 5.27 |
| UO62h | 5.93 | 4.84 |
| UO62i | 6.44 | 5.19 |
| UO62j | 5.62 | 4.37 |
| UO62k | 6.42 | 5.15 |
| UO62l | 5.90 | 4.63 |
| UO63i | 5.31 | 4.49 |
| UO63k | 5.23 | 4.30 |
| UO63l | 5.38 | 4.34 |
| UO67 | 6.73 | 5.15 |
| UO68 | 6.12 | 4.93 |
| UO71 | 6.58 | 5.53 |
| UO75 | 6.23 | 5.16 |
| UO89 | 6.14 | 5.16 |
| UO90 | 6.07 | 5.09 |
| UO95 | 6.05 | 5.11 |
| UO109 | 6.41 | 5.27 |
| UO110 | 6.51 | 5.36 |
| UO111 | 6.69 | 5.31 |
| UO112 | 6.30 | 5.56 |
| UO128 | 6.79 | 5.66 |
| UO129 | 6.82 | 5.76 |
| UO130 | 7.10 | 5.70 |
| UO131 | 6.86 | 6.08 |
| UO132 | 6.49 | 5.71 |

**Tab. 50:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 61 Thermolysininhibitoren aus dem Trainingsdatensatz von Klebe *et al.* (Klebe *et al.*, 1994) (s.a. Tab. 13, S. 95).

| **Bezeichnung** | $pK_i$ | **Bezeichnung** | $pK_i$ |
|---|---|---|---|
| ACE_OHLEU_AGNH2 | 4.24 | PO3_FAGNH2 | 5.03 |
| BZSAG | 4.65 | PPHEOH | 3.97 |
| C6PCLTNME | 5.66 | R_THIORPHAN | 4.01 |
| C6PLTNME | 6.25 | S02P_FAGNH2 | 4.73 |
| C6POLTNME | 6.26 | S_THIORPHAN | 3.45 |
| CBZPHE | 4.78 | SO3_FAGNH2 | 4.73 |
| CH3COCH2CO_FAGNH2 | 5.08 | Z_D_APOLA | 5.91 |
| CH3O2S_FAGNH2 | 4.73 | Z_D_FPLA | 6.37 |
| CHO_OHLEU_AGNH2 | 4.22 | Z_D_FPOLA | 6.27 |
| CLTZNCRYS | 6.55 | Z_D_LPOLA | 6.35 |
| DAH50 | 6.88 | Z_NH_GLNH2 | 4.88 |
| DAH51 | 6.09 | Z_NH_GLNHOH | 4.31 |
| DAH52 | 5.19 | ZALA | 7.55 |
| DAH53 | 8.58 | ZAPOLA | 5.43 |
| DAH54 | 6.46 | ZFPLAZNCRYS | 6.72 |
| DAH55 | 5.51 | ZFPOLA | 6.74 |
| HOCH2CO_FAGNH2 | 4.86 | ZG_D_LNHOH | 4.71 |
| NHOHBZMAGNA | 5.76 | ZGG_D_LNHOH | 5.42 |
| NHOHBZMAGNH2 | 4.82 | ZGGLNHOH | 5.55 |
| NHOHBZMAGOH | 4.84 | ZGGNHOH | 4.10 |
| NHOHBZMOET | 3.74 | ZGLNH2 | 5.20 |
| NHOHIBMAGNH2 | 4.46 | ZGLNHOH | 4.71 |
| NHOHLEU | 2.63 | ZGLNMEOH | 4.51 |
| NHOHMALAGNH2 | 4.00 | ZGLY | 7.40 |
| OHBZMAGNH2 | 5.10 | ZGPCLLZNCRYS | 6.38 |
| P_ILE_AOH | 3.54 | ZGPLA | 6.08 |
| P_OPHE_OME_LEUNH2 | 4.02 | ZGPLLZNCRYS | 6.53 |
| PAAOH | 3.28 | ZGPOLA | 5.64 |
| PHOSPHORAMIDON | 6.79 | ZGPOLLZNCRYS | 6.12 |
| PLEUNH2 | 3.20 | ZLPOLA | 5.96 |
| PNHET | 1.91 | - | - |

**Tab. 51:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 15 Thermolysininhibitoren aus dem Testdatensatz von Klebe *et al.* (Klebe *et al.*, 1994) (s.a. Tab. 14, S. 95).

| **Bezeichnung** | $pK_i$ | **Bezeichnung** | $pK_i$ |
|---|---|---|---|
| PLFOH | 5.11 | ZGPLG | 6.16 |
| PPPHE | 5.18 | ZGPLNH2 | 4.50 |
| ZFGNH2 | 4.87 | ZGPOLF | 5.51 |
| ZGPCLA | 5.77 | ZGPOLG | 5.32 |
| ZGPCLF | 6.42 | ZGPOLNH2 | 4.69 |
| ZGPCLG | 5.65 | ZLGNH2 | 4.67 |
| ZGPCLNH2 | 5.06 | ZYGNH2 | 5.01 |
| ZGPLF | 6.69 | - | - |

**Tab. 52:** Nach Gl. 36 und Gl. 42 berechnete Bindungsaffinitäten für 53 Protein-Ligand-Komplexe, die aus den Datensätzen FlexX_DS1 und FlexX_DS2 extrahiert wurden und für die eine Bindungsaffinität in Eldrige *et al.* (Eldridge *et al.*, 1997), Head *et al.* (Head *et al.*, 1996) bzw. Böhm (Böhm, 1994; Böhm, 1998) berichtet wurde (s.a. Tab. 15, S. 96).

| **PDB-Code** | $pK_i$ | **PDB-Code** | $pK_i$ |
|---|---|---|---|
| 1aaq | 11.52 | 1tni | 4.81 |
| 1abe | 3.94 | 1tnk | 4.53 |
| 1abf | 4.20 | 1tnl | 4.21 |
| 1apt | 7.60 | 2cgr | 6.95 |
| 1cbx | 5.27 | 2cpp | 2.63 |
| 1cps | 4.43 | 2ctc | 4.10 |
| 1dwd | 8.73 | 2er6 | 2.78 |
| 1eed | 7.91 | 2gbp | 5.72 |
| 1ela | 6.57 | 2tmn | 4.18 |
| 1elc | 7.25 | 2xis | 4.03 |
| 1etr | 5.27 | 2ypi | 3.55 |
| 1hsl | 4.62 | 3cpa | 5.20 |
| 1hvr | 10.76 | 3ptb | 4.65 |
| 1ldm | 2.04 | 4dfr | 8.04 |
| 1mbi | 1.26 | 4hmg | 8.04 |
| 1nsc | 5.40 | 4hvp | 4.15 |
| 1phf | 2.47 | 4phv | 8.45 |
| 1phg | 3.71 | 4tln | 12.41 |
| 1ppc | 9.10 | 4tmn | 3.66 |
| 1pph | 6.86 | 4ts1 | 8.02 |
| 1ppk | 8.28 | 5abp | 4.57 |
| 1pso | 11.45 | 5tmn | 4.47 |
| 1rbp | 7.13 | 6abp | 8.15 |
| 1rne | 11.69 | 6cpa | 3.87 |
| 1tlp | 8.18 | 6tmn | 7.56 |
| 1tng | 3.81 | 7cpa | 7.72 |
| 1tnh | 4.05 | - | - |

**Tab. 53:** Nach Gl. 54 unter Anwendung einer nicht-kreuzvalidierten PLS-Analyse erhaltene Bindungsaffinitäten für 61 Thermolysininhibitoren aus dem Trainingsdatensatz von Klebe *et al.* (Klebe *et al.*, 1994) (s.a. Tab. 13, S. 95).

| **Bezeichnung** | $pK_i$ | **Bezeichnung** | $pK_i$ |
|---|---|---|---|
| ACE_OHLEU_AGNH2 | 2.29 | PO3_FAGNH2 | 5.30 |
| BZSAG | 6.34 | PPHEOH | 4.45 |
| C6PCLTNME | 7.43 | R_THIORPHAN | 5.71 |
| C6PLTNME | 8.30 | S02P_FAGNH2 | 5.43 |
| C6POLTNME | 5.97 | S_THIORPHAN | 6.23 |
| CBZPHE | 2.82 | SO3_FAGNH2 | 1.67 |
| CH3COCH2CO_FAGNH2 | 2.34 | Z_D_APOLA | 4.67 |
| CH3O2S_FAGNH2 | 0.54 | Z_D_FPLA | 6.50 |
| CHO_OHLEU_AGNH2 | 2.23 | Z_D_FPOLA | 4.06 |
| CLTZNCRYS | 7.80 | Z_D_LPOLA | 4.42 |
| DAH50 | 7.78 | Z_NH_GLNH2 | 3.00 |
| DAH51 | 6.31 | Z_NH_GLNHOH | 5.22 |
| DAH52 | 4.72 | ZALA | 6.15 |
| DAH53 | 6.68 | ZAPOLA | 5.90 |
| DAH54 | 5.50 | ZFPLAZNCRYS | 10.06 |
| DAH55 | 3.26 | ZFPOLA | 7.44 |
| HOCH2CO_FAGNH2 | 2.83 | ZG_D_LNHOH | 5.09 |
| NHOHBZMAGNA | 6.51 | ZGG_D_LNHOH | 3.55 |
| NHOHBZMAGNH2 | 5.81 | ZGGLNHOH | 4.50 |
| NHOHBZMAGOH | 5.68 | ZGGNHOH | 2.54 |
| NHOHBZMOET | 4.86 | ZGLNH2 | 1.71 |
| NHOHIBMAGNH2 | 6.40 | ZGLNHOH | 5.05 |
| NHOHLEU | 3.62 | ZGLNMEOH | 2.66 |
| NHOHMALAGNH2 | 3.56 | ZGLY | 6.49 |
| OHBZMAGNH2 | 3.43 | ZGPCLLZNCRYS | 6.62 |
| P_ILE_AOH | 6.43 | ZGPLA | 7.60 |
| P_OPHE_OME_LEUNH2 | 1.37 | ZGPLLZNCRYS | 8.13 |
| PAAOH | 3.95 | ZGPOLA | 4.64 |
| PHOSPHORAMIDON | 7.46 | ZGPOLLZNCRYS | 5.23 |
| PLEUNH2 | 4.56 | ZLPOLA | 5.91 |
| PNHET | 0.59 | - | - |

**Tab. 54:** Nach Gl. 62 ($\theta = 1$) unter Verwendung des aus der PLS-Analyse erhaltenen Modells vorhergesagte Bindungsaffinitäten für 15 Thermolysininhibitoren aus dem Testdatensatz von Klebe *et al.* (Klebe *et al.*, 1994) (s.a. Tab. 14, S. 95).

| **Bezeichnung** | $pK_i$ | **Bezeichnung** | $pK_i$ |
|---|---|---|---|
| PLFOH | 6.19 | ZGPLG | 6.88 |
| PPPHE | 2.93 | ZGPLNH2 | 6.14 |
| ZFGNH2 | 0.26 | ZGPOLF | 4.69 |
| ZGPCLA | 5.75 | ZGPOLG | 3.84 |
| ZGPCLF | 5.93 | ZGPOLNH2 | 3.59 |
| ZGPCLG | 5.40 | ZLGNH2 | 0.07 |
| ZGPCLNH2 | 4.92 | ZYGNH2 | -0.17 |
| ZGPLF | 7.04 | - | - |

# Literaturverzeichnis

Ackers, G. F. & Smith, F. R. (1985). Effects of site-specific amino acid modification on protein interactions and biological function. *Annu Rev Biochem* **54**, 597-629.

Ackers, G. K., Doyle, M. L., Myers, D. & Daugherty, M. A. (1992). Molecular code for cooperativity in hemoglobin. *Science* **255**, 54-63.

Ajay & Murcko, M. A. (1995). Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes. *J Med Chem* **38**, 4953-4967.

Allen, F. H. (1998). The Development, Status and Scientific Impact of Crystallographic Databases. *Acta Cryst* **A54**, 758-771.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). The development of version-3 and version-4 of the Cambridge Structural Database system. *J Chem Inf Comput Sci* **31**, 187-204.

Andrews, P. R., Craik, D. J. & Martin, J. L. (1984). Functional Group Contributions to Drug-Receptor Interactions. *J Med Chem* **27**, 1648-1657.

Andrus, M. B. & Schreiber, S. L. (1993). Structure-Based Design of an Acyclic Ligand that Bridges FKBP12 and Calcineurin. *J Am Chem Soc* **115**, 10420-10421.

Antosiewicz, J., McCammon, J. A. & Gilson, M. K. (1996). The Determinants of $pK_a$s in Proteins. *Biochemistry* **35**, 7819-7833.

Aquist, J., Medina, C. & Samuelsson, J. E. (1994). New method for predicting binding-affinity in computer-aided drug design. *Protein Eng* **7**, 385-391.

Astley, T., Birch, G. G., Drew, M. G. B., Rodger, P. M. & Wilden, G. R. H. (1998). Effect of available volumes on radial distribution fucntions. *J Comput Chem* **19**, 363-367.

Atkins, P. W. (1990). *Physikalische Chemie*, VCH Verlagsgesellschaft mbH, Weinheim.

Babine, R. E. & Bender, S. L. (1997). Molecular Recognition of Protein-Ligand Complexes: Applications to Drug Design. *Chem Rev* **97**, 1359-1472.

Bahar, I. & Jernigan, R. L. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol* **266**, 195-214.

Balkenhohl, F., von dem Bussche-Hünnefeld, C., Lansky, A. & Zechel, C. (1996). Kombinatorische Synthese niedermolekularer organischer Verbindungen. *Angew Chem* **108**, 2436-2488.

Bamborough, P. & Cohen, F. E. (1996). Modeling protein-ligand interactions. *Curr Opin Strruct Biol* **6**, 236-241.

Barril, X., Aleman, C., Orozco, M. & Luque, F. J. (1998). Salt Bridge Interactions: Stability of the Ionic and Neutral Complexes in the Gas Phase, in Solution, and in Proteins. *Proteins* **32**, 67-79.

Bartlett, P. A. & Marlowe, C. K. (1987). Evaluation of intrinsic binding energy from a hydrogen bonding group in an enzyme inhibitor. *Science* **235**, 569-71.

Bash, P. A., Singh, U. C., Brown, F. K., Langridge, R. & Kollman, P. A. (1987). Calculation of the Relative Change in Binding Free Energy of a Protein-Inhibitor Complex. *Science* **235**, 574-576.

Bashford, D. & Karplus, M. (1990). pKa's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model. *Biochemistry* **29**, 10219-10225.

Beauchamp, J. C. & Isaacs, N. W. (1999). Methods for X-ray diffraction analysis of macromolecular structures. *Curr Opin Chem Biol* **3**, 525-529.

Beeson, C., Pham, N., Shipps, G. & Dix, T. A. (1993). A Comprehensive Description of the Free Energy of an Intramolecular Hydrogen Bond as a Function of Solvation: NMR Study. *J Am Chem Soc* **115**, 6803-6812.

Ben-Naim, A. (1980). *Hydrophobic Interactions*, Plenum Press, New York.

Ben-Naim, A. (1987). *Solvation Thermodynamics*, Plenum Press, New York.

Ben-Naim, A. (1992). *Statistical Thermodynamics for Chemists and Biochemists*, Plenum Press, New York.

Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: Are these meaningful potentials? *J Chem Phys* **107**, 3698-3706.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**, 535-42.

Bernstein, J. (1989). Polymorphism in drug design and delivery. *Prog Clin Biol Res* **289**, 203-215.

Bernstein, P. R., Andisik, D., Bradley, P. K., Bryant, C. B., Ceccarelli, C., Damewood, J. R., Earley, R., Edwards, P. D., Feeney, S., Gomes, B. C., Kosmider, B. J., Steelman, G. B., Thomas, R. M., Vacek, E. P., Veale, C. A., Williams, J. C., Wolanin, D. J. & Woolson, S. A. (1994). Nonpeptidic inhibitors of human leukocyte elastase. 3. Design, synthesis, X-ray crystallographic analysis, and structure-activity relationships for a se-

ries of orally active 3-amino-6-phenylpyridin-2-one trifluoromethyl ketones. *J Med Chem* **37**, 3313-3326.

Beveridge, D. L. & DiCapua, F. M. (1989). Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu Rev Biophys Biophys Chem* **18**, 431-92.

Bibel, W., Hölldobler, S. & Schaub, T. (1993). *Wissensrepräsentation und Inferenz: Eine grundlegende Einführung*, Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig.

Blake, C. C. F., Cassels, R., Dobson, C. M., Poulsen, F. M., Williams, R. J. P. & Wilson, K. S. (1981). Structure and binding properties of hen lysozyme modified at tryptophan 62. *J Mol Biol* **147**, 73-95.

Blaney, J. M. & Dixon, J. S. (1993). A good ligand is hard to find: Automated docking methods. *Persp Drug Discov Design* **1**, 301-319.

Blokzijl, W. & Engberts, J. B. F. N. (1993). Hydrophobe Effekte - Ansichten und Tatsachen. *Angew Chem* **105**, 1610-1648.

Bode, W., Mayr, I., Baumann, U., Huber, R., Stone, S. R. & Hofsteenge, J. (1989). The refined 1.9 A crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J* **8**, 3467-3475.

Bohacek, R. S. & McMartin, C. (1994). Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a *de Novo* Design Method Incorporating Combinatorial Growth. *J Am Chem Soc* **116**, 5560-5571.

Böhm, H. J. (1992). The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* **6**, 61-78.

Böhm, H. J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three- dimensional structure. *J Comput Aided Mol Des* **8**, 243-56.

Böhm, H. J. (1998). Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* **12**, 309-23.

Böhm, H.-J. & Klebe, G. (1996). What Can We Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs? *Angew Chem Int Ed Engl* **35**, 2566-2587.

Böhm, H.-J., Klebe, G. & Kubinyi, H. (1996). *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg.

Böhm, M., Stürzebecher, J. & Klebe, G. (1999). Three-Dimensional Quantitative Structure-Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indics Analysis To Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J Med Chem* **42**, 458-477.

Boobbyer, D. N. A., Goodford, P. J., McWhinnie, P. M. & Wade, R. C. (1989). New Hydrogen-Bond Potentials for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. *J Med Chem* **32**, 1083-1094.

Booch, G. (1994). *Object-Oriented Analysis and Design*, Benjamin-Cummings, New York.

Borchardt-Ott, W. (1993). *Kristallographie - Eine Einführung für Naturwissenschaftler*, Springer, Berlin.

Boresch, S. & Karplus, M. (1995). The meaning of component analysis - decomposition of free-energy in terms of specific interactions. *J Mol Biol* **254**, 801-807.

Bostrom, J., Norrby, P. O. & Liljefors, T. (1998). Conformational energy penalties of protein-bound ligands. *J Comput Aided Mol Des* **12**, 383-96.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three- dimensional structure. *Science* **253**, 164-70.

Boyd, D. B. (1990). Successes of Computer-Assisted Molecular Design. In *Reviews of Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., eds.), Vol. 1, S. 355-371. VCH Publishers, New York.

Boyd, D. B. (1998). Rational drug design: controlling the size of the haystack. *Modern Drug Discov* **6**, 41-48.

Brady, G. P. & Sharp, K. A. (1995). Decomposition of Interaction Free Energies in Proteins and Other Complex Systems. *J Mol Biol* **254**, 77-85.

Brady, G. P. & Sharp, K. A. (1997a). Energetics of cyclic dipeptide crytsal packing and solvation. *Biophys J* **72**, 913-927.

Brady, G. P. & Sharp, K. A. (1997b). Entropy in protein folding and in protein-protein interactions. *Curr Opin Struct Biol* **7**, 215-221.

Brooks, R. R., Bruccoleri, B. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* **4**, 187-217.

Brown, R. D. & Martin, Y. C. (1996). Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* **36**, 572-584.

Brünger, A. T. (1992). The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-474.

Brünger, A. T. (1997). X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nature Struct Biol (NMR supplement)* **4**, 862-865.

Bruno, I. J., Cole, J. C., Lommerse, J. P., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). IsoStar: a library of information about nonbonded interactions. *J Comput Aided Mol Des* **11**, 525-37.

Bures, M. G. (1997). Recent techniques and applications in pharmacophore mapping. In *Practical application of computer-aided drug design* (Charifson, P. S., ed.), S. 39-72. Marcel Dekker, New York.

Bürgi, H. B. & Dunitz, J. D. (1988). Can Statistical Analysis of Structural Parameters from Different Crystal Environments Lead to Quantitative Energy Relationships. *Acta Cryst* **B44**, 445-448.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). Structural genomics: beyond the human genome project. *Nat Genet* **23**, 151-157.

Bush, B. L. & Nachbar, R. B. (1993). Sample-distance Partial Least Squares: PLS optimized for many variables, with application to CoMFA. *J Comput Aided Mol Des* **7**, 587-619.

Buzbee, B. (1993). Workstation clusters rise and shine (Computing in Science: Perspective). *Science* **261**, 852-853.

Carell, T., Wintner, E. A. & Rebeck Jr., J. (1994). Screeningverfahren in Lösung zur Isolierung biologisch aktiver Verbindungen aus einer Molekülbibliothek. *Angew Chem* **106**, 2162-2164.

Carlson, H. A. & Jorgensen, W. L. (1995). An extended linear-response method for determining free-energies of hydration. *J Phys Chem* **99**, 10667-10673.

Carrupt, P., Testa, B. & Gaillard, P. (1997). Computational Approaches to Lipophilicity: Methods and Applications. In *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 11. Wiley-VCH, New York.

Carugo, O. & Bordo, D. (1999). How many water molecules can be detected by protein crystallography. *Acta Cryst D* **55**, 479-483.

Caspar, D. L. D., Clarage, J., Salunke, D. M. & Clarage, M. (1988). Liquid-like movements in crystalline insulin. *Nature* **332**, 659-662.

Chan, O. H. & Stewart, B. H. (1996). Physicochemical and drug-delivery considerations for oral drug bioavailibilty. *Drug Discov Today* **1**, 461-473.

Charifson, P. S., Corkerey, J. J., Murcko, M. A. & Walters, W. P. (1999). Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J Med Chem* **42**, 5100-5109.

Chayen, N. E., Boggon, T. J., Casetta, A., Deacon, A., Gleichmann, T., Habash, J., Harrop, S. J., Helliwell, J. R., Nieh, Y. P., Peterson, M. R., Raftery, J., Snell, E. H., Hädener, A., Niemann, A. C., Siddons, D. P., Stojanoff, V., Thompson, A. W., Ursby, T. & Wulff, M. (1996). Trends and Challenges in Experimental Macromolecular Crystallography. *Quart Rev Biophys* **29**, 227-278.

Checa, A., Ortiz, A. R., de Pascual-Teresa, B. & Gago, F. (1997). Assessment of Solvation Effects on Calculated Binding Affinity Differences: Trypsin Inhibition by Flavonoids as a Model System for Congeneric Series. *J Med Chem* **40**, 4136-4145.

Cheng, H. M., Keitz, P. & Jones, J. B. (1994). Design an d Synthesis of a Conformationally Restricted Cysteine Protease Inhibitor. *J Org Chem* **59**, 7671-7676.

Chervenak, M. C. & Toone, E. J. (1994). A Direct Measure of the Contribution of Solvent Reorganization to the Enthalpy of Ligand Binding. *J Am Chem Soc* **116**, 10533-10539.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338-339.

Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature* **256**, 705-708.

Clark, D. E., Murray, C. W. & Li, J. (1997). Current Issues in De Novo Molecular Design. In *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 11, S. 67-125. Wiley-VCH, New York.

Clark, D. E. & Pickett, S. D. (2000). Computational methods for the prediction of 'drug-likeness'. *Drug Discov Today* **5**, 49-58.

Clark, M., Cramer, R. D., III & Van Opdenbosch, N. (1989). Validation of the General Purpose Tripos 5.2 Force Field. *J Comp Chem* **10**, 982-1012.

Clore, G. M. & Gronenborn, A. M. (1991). Structures of Larger Proteins in Solution: Three- and Four-Dimensional Heteronuclear NMR Spectroscopy. *Science* **252**, 1390-1399.

Cole, J. C., Taylor, R. & Verdonk, M. L. (1998). Directional Preferences of Intermolecular Contacts to Hydrophobic Groups. *Acta Cryst* **D54**, 1183-1193.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., Fearon, E., Hartwelt, L., Langley, C. H., Mathies, R. A., Olson, M., Pawson, A. J., Pollard, T., Williamson, A., Wold, B., Buetow, K., Branscomb, E., Capecchi, M., Church, G., Garner, H., Gibbs, R. A., Hawkins, T., Hodgson, K., Knotek, M., Meisler, M., Rubin, G. M., Smith, L. M., Westerfield, M., Clayton, E. W., Fisher, N. L., Lerman, C. E.,

McInerney, J. D., Nebo, W., Press, N. & Valle, D. (1998). New Goals for the US Human Genom Project: 1998-2000. *Science* **282**, 682-689.

Connelly, P. R., Ed. (1997). Structural-based Drug Design: thermodynamics, modeling and strategy. Edited by Ladbury, J. E. & Connelly, P. R. Austin (Texas): R. G. Landes.

Connelly, P. R., Aldape, R. A., Bruzzese, F. J., Chambers, S. P., Fitzgibbon, M. J., Fleming, M. A., Itoh, S., Livingston, D. J., Navia, M. A., Thomson, J. A. & Wilson, K. P. (1994). Enthalpy of hydrogen bond formation in a protein-ligand binding raction. *Proc Natl Acad Sci USA* **91**, 1964-1968.

Connolly, M. L. (1983). Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science* **221**, 709-713.

Couzin, J. (1998). Supercomputing - computer experts urge new federal initiative. *Science* **281**, 762.

Covell, D. G. & Wallquist, A. (1997). Analysis of Protein-Protein Interactions and the Effects of Amino Acid Mutations on Their Energetics. The Importance of Water Molecules in the Binding Epitope. *J Mol Biol* **269**, 281-297.

Cramer III, R. D. & Bunce, J. D. (1987). The DYLOMMS Method: Initial Results from a Comparative Study of Approaches to 3D QSAR. In *QSAR in Drug Design and Toxicology* (Hadzi, D. & Jerman-Blasiz, B., Ed.). Elsevier, Amsterdam.

Cramer III, R. D., DePriest, S. A., Patterson, D. E. & Hecht, P. (1993). The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design. Theory, Methods and Applications* (Kubinyi, H., ed.). ESCOM, Leiden.

Cramer III, R. D., Patterson, D. E. & Bunce, J. D. (1988). Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Proteins. *J Am Chem Soc* **110**, 5959.

Crippen, G. M. & Snow, M. E. (1990). A 1.8 Å resolution potential function for protein folding. *Biopolymers* **29**, 1479-1489.

Damewood, J. R., Edwards, P. D., Feeney, S., Gomes, B. C., Steelman, G. B., Tuthill, P. A., Williams, J. C., Warner, P., Woolson, S. A., Wolanin, D. J. & Veale, C. A. (1994). Nonpeptidic inhibitors of human leukocyte elastase. 2. Design, synthesis, and in vitro activity of a series of 3-amino-6-arylopyridin- 2-one trifluoromethyl ketones. *J Med Chem* **37**, 3303-3312.

Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M. & Hagler, A. T. (1988). Structure and Energetics of ligand binding to proteins: Escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system. *Proteins* **4**, 31-47.

Davis, A. M. & Teague, S. J. (1999). Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew Chem Int Ed Engl* **38**, 736-749.

Dawson, R. M. C., Elliott, D. C., Elliot, W. H. & Jones, K. M. (1969). *Data for Biochemical Research*. 2nd edit, Oxford University Press, Oxford.

de Namor, A. F. D., Ritt, M.-C., Schwing-Weill, M.-J., Arnaud-Neu, F. & Lewis, D. F. V. (1991). Solution Thermodynamics of Amino Acid-18-Crown-6 and Amino Acid-Cryptand 222 Complexes in Methanol and Ethanol. *J Chem Soc Faraday Trans* **87**, 3231-3239.

De Priest, S. A., Mayer, D., Naylor, C. B. & Marshall, G. R. (1993). 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J Am Chem Soc* **115**, 5372-5384.

Dean, P. M. (1987). *Molecular foundations of drug-receptor interaction*, Cambridge University Press, Cambridge.

Dean, P. M. (1995). Defining molecular similarity and complementarity for drug design. In *Molecular Similarity in Drug Design* (Dean, P. M., ed.), S. 1-23. Blackie Academic & Professional, London.

Delarue, M. & Koehl, P. (1995). Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J Mol Biol* **249**, 675-90.

DelPhi/Solvation. (1995). Electrostatic Potential and Solvation Energy Software, Molecular Simulations Inc., San Diego, CA.

Desiraju, G. R. (1996). The C-H...O Hydrogen Bond: Structural Implications and Supramolecular Design. *Acc Chem Res* **29**, 441-449.

DeWitte, R. S. & Shaknovich, E. I. (1996). SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J Am Chem Soc* **118**, 11733-11744.

di Cera, E. (1995). *Thermodynamic theory of site-specific binding processes in biological macromolecules*, Cambridge University Press, Cambridge.

Dill, K. A. (1990). Dominant Forces in Protein Folding. *Biochemistry* **29**, 7133-7155.

Dill, K. A. (1997). Additivity Principles in Biochemistry. *J Biol Chem* **272**, 701-4.

Dixon, J. S. (1997). Evaluation of the CASP2 Docking Section. *Proteins* **Suppl.1**, 198-204.

Doig, A. J. & Sternberg, M. J. E. (1995). Side-chain conformational entropy in protein folding. *Protein Sci* **4**, 2247-2251.

Doig, A. J. & Williams, D. H. (1992). Binding Energy of an Amide-Amide Hydrogen Bond in Aqueous and Nonpolar Solvents. *J Am Chem Soc* **114**, 338-343.

Doscher, M. S. & Richards, F. M. (1963). The Activity of an Enzyme in the Crystalline State: Ribonuclease S. *J Biol Chem* **238**, 2399-2406.

Doucet, J. & Benoit, J. P. (1987). Molecular dynamics studied by analysis of the X-ray diffuse scattering from lysozyme crystals. *Nature* **325**, 643.

Dougherty, D. A. (1996). Cation-π Interactions in Chemistry and Biology: A New View of Benzene, Phe, Tyr, and Trp. *Science* **271**, 163-168.

Doweyko, A. (1988). The Hypothetical Active Site Lattice. An Approach to Modelling Active Sites from Data on Inhibitor Molecules. *J Med Chem* **31**, 1396.

Dransfeld, K., Kienle, P. & Vonach, H. (1992). *Physik I - Newtonsche und relativistische Mechanik*. 6 edit, R. Oldenbourg Verlag, München.

Drenth, J. (1999). *Principles of Protein X-ray Crystallography*. Springer Advanced Texts in Chemistry (Cantor, C. R., Ed.), Springer, New York.

Dunitz, J. D. (1994). The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **264**, 670.

Dunitz, J. D. (1995). Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem Biol* **2**, 709-712.

Eads, J., Sacchettini, J. C., Kromminga, A. & Gordon, J. I. (1993). Escherichia coli-derived rat intestinal fatty acid binding protein with bound myristate at 1.5 Å resolution and I-FABPArg106-->Gln with bound oleate at 1.74 Å resolution. *J Biol Chem* **268**, 26375-85.

Ehrlich, P. (1913). Address in Pathology on Chemotherapeutics: Scientific Principles, Methods and Results. *Lancet* **II**, 445-451.

Eisenberg, D. & Kauzmann, W. (1969). *The structure and properties of water*, Oxford University Press, Oxford.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199-203.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **11**, 425-45.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst* **A47**, 392-400.

Ewing, T., Ed. (1997). DOCK Version 4.0 Manual. Regents of the University of California, San Francisco, CA.

Ewing, T. J. A. & Kuntz, I. D. (1997). Critical evaluation of search algorithms for automated molecular docking and database screening. *J Comput Chem* **18**, 1175-1189.

Fernandez-Recio, J., Romero, A. & Sancho, J. (1999). Energetics of a Hydrogen Bond (Charged and Neutral) and of a Cation-π–Interaction in Apoflavodoxin. *J Mol Biol* **290**, 319-330.

Fersht, A. (1985). *Enzyme Structure and Mechanism*, W. H. Freeman and Company, New York.

Fersht, A. R. (1987). The hydrogen bond in molecular recognition. *Trends Biochem Sci* **12**, 301-304.

Fersht, A. R., Shi, J. P., Knill-Jones, J., Lowe, D. M., Wilkinson, A. J., Blow, D. M., Brick, P., Carter, P., Waye, M. M. Y. & Winter, G. (1985). Hydrogen bonding and biological specifity analysed by protein engineering. *Nature* **314**, 235-238.

Finkelstein, A. V. (1997). Protein structure: what is possible to predict now? *Curr Opin Struct Biol* **7**, 60-71.

Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Y. (1995). Perfect temperature for protein structure prediction and folding. *Proteins* **23**, 151-62.

Finkelstein, A. V. & Janin, J. (1989). The price of lost freedom: entropy of bimolecular complex formation. *Protein Eng* **3**, 1-3.

Fischer, E. (1894). Einfluß der Konfiguration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges* **27**, 2985-2993.

Folkers, G. & Merz, A. (1996). Hydrophobic Fields in Quantitative Structure-Activity Relationships. In *Lipophilicity ind Drug Action and Toxicology* (Pliska, V., Testa, B. & van de Waterbeemd, H., Ed.), Vol. 4, S. 219-232. VCH Publishers, Weinheim.

Folkers, G., Merz, A. & Rognan, D. (1993). CoMFA: Scope and Limitations. In *3D QSAR in Drug Design* (Kubinyi, H., ed.), S. 583-618. ESCOM, Leiden.

Forman-Kay, J. D. (1999). The 'dynamics' in the thermodynamics of binding. *Nat Struct Biol* **6**, 1086-1087.

Frank, H. S. & Evans, M. W. (1945). Free Volume and Entropy in Condensed Systems. *J Chem Phys* **13**, 507.

Franks, F., Ed. (1972-1982). Water, a Comprehensive Treatise. Vol. 1-7. New York: Plenum Press.

Free, S. M. & Wilson, J. W. (1964). A Mathematical Contribution to Structure-Activity Studies. *J Med Chem* **7**, 395-399.

Froloff, N., Windemuth, A. & Honig, B. (1997). On the calculation of binding free energies using continuum methods: Application to MHC class I protein-peptide interactions. *Prot Sci* **6**, 1293-1301.

Furuichi, E. & Koehl, P. (1998). Influence of Protein Structure Databases on the Predictive Power of Statistical Pair Potentials. *Proteins* **31**, 139-149.

Gao, J., Kuczera, K., Tidor, B. & Karplus, M. (1989). Hidden thermodynamics of mutant proteins - a molecular-dynamics analysis. *Science* **244**, 1069-1072.

Gasteiger, J. & Marsili, M. (1980). Iterative Partial Equalization of Orbital Electronegativity - A rapid Access to Atomic Charges. *Tetrahedron* **36**, 3219-3228.

Gasteiger, J., Rudolph, C. & Sadowski, J. (1990). Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comp Method* **3**, 537-547.

Gasteiger, J. & Zupan, J. (1993). Neural Networks in Chemistry. *Angew Chem Int Ed Engl* **105**, 510-536.

Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B. & Free, S. T. (1995). Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease. *Chem Biol* **2**, 317-324.

Geladi, P. & Kowalski, B. R. (1986). Partial Least Squares Regression: A Tutorial. *Analyt Chim Acta* **185**, 1-17.

Gerber, P. R. (1998). Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. *J Comput Aided Mol Des* **12**, 37-51.

Gerber, P. R., Mark, A. E. & van Gunsteren, W. F. (1993). An approximate but efficient method to calculate free energy trends by computer simulation: Application to dihydrofolate reductase-inhibitor complexes. *J Comput Aided Mol Des* **7**, 305-323.

Gerber, P. R. & Müller, K. (1995). MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J Comput Aided Mol Des* **9**, 251-268.

Gilli, P., Ferretti, V., Gilli, G. & Borea, P. A. (1994). Enthalpiy-Entropy Compensation in Drug-Receptor Binding. *J Phys Chem* **98**, 1515-1518.

Gilson, M. K. (1995). Theory of electrostatic interactions in macromolecules. *Curr Opin Struct Biol* **5**, 216-223.

Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. (1997). The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys J* **72**, 1047-1069.

Glusker, J. P., Lewis, M. & Rossi, M. (1994). *Crystal Structure Analysis for Chemists and Biologists*, VCH, Weinheim.

Godzik, A., Kolinski, A. & Skolnick, J. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci* **4**, 2107-17.

Golender, V. E. & Vorpagel, E. R. (1993). Computer-Assisted Pharmacophore Identification. In *3D-QSAR in Drug Design: Theory, Methods and Applications* (Kubinyi, H., ed.), S. 137-149. ESCOM, Leiden.

Goodford, P. J. (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J Am Chem Soc* **28**, 849.

Gordon, E. M., Gallop, M. A. & Patel, D. V. (1996). Strategy and Tactics in Combinatorial Organic Syntheses. Applications to Drug Discovery. *Acc Chem Res* **29**, 144-154.

Gordon, M. S. & Jensen, J. H. (1996). Understanding the Hydrogen Bond Using Quantum Chemistry. *Acc Chem Res* **29**, 536-543.

Greco, G., Novellino, E. & Martin, Y. C. (1998). 3D-QSAR Methods. In *Designing Bioactive Molecules* (Martin, Y. C., Willett, P. & Heller, S. R., Ed.), S. 219-251. American Chemical Society, Washington, DC.

Greer, J., Erickson, J. W., Baldwin, J. J. & Varney, M. D. (1994). Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J Med Chem* **37**, 1035-1054.

Grobelny, D., Goli, U. B. & Galardy, R. E. (1989). Binding energetics of phosphorus-containing inhibitors of thermolysin. *Biochemistry* **28**, 4948-51.

Grootenhuis, P. D. J. & van Galen, P. J. M. (1995). Correlation of binding affinities with non-bonded interaction energies of thrombin-inhibitor complexes. *Acta Cryst* **D51**, 560-566.

Grootenhuis, P. D. J. & van Helden, S. P. (1994). Rational approaches towards protease inhibition: predicting the binding of thrombin inhibitors. In *Computational Approaches in Supramolecular Chemistry* (Wipff, G., ed.), S. 137-149. Kluwer Academic Press, Dordrecht.

Grunwald, E. & Steel, C. (1995). Solvent Reorganization and Thermodynamic Enthalpy-Entropy Compensation. *J Am Chem Soc* **117**, 5687-5692.

Gschwend, D. A. & Kuntz, I. D. (1996). Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal. *J Comput Aided Mol Des* **10**, 123-32.

Guo, Z., Brooks, C. L. & Kong, X. (1998). Efficient and flexible algorithm for free energy calculations using the lambda-dynamics approach. *J Phys Chem* **B 102**, 2032-2036.

Habermann, S. M. & Murphy, K. P. (1996). Energetics of hydrogen bonding in protein: A model compound study. *Protein Sci* **5**, 1229-1239.

Hansch, C. & Fujita, T. (1964). p-σ–π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc* **86**, 1616-1626.

Hansch, C. & Leo, A. (1979). *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley and Sons, New York.

Hansson, T., Marelius, J. & Aquist, J. (1998). Ligand binding affinity prediction by linear interaction energy methods. *J Comput Aided Mol Des* **12**, 27-35.

Hardcastle, I. R., Rowlands, M. G., Houghton, J., Parr, I. B., Potter, G. A., Jarmann, M., Edwards, K. J., Laughton, C. A., Trent, J. O. & Neidle, S. (1995). Rationally designed analogues of tamoxifen with improved calmodulin antagonism. *J Med Chem* **38**, 241-248.

Harder, S. (1999). Can C-H...C(π) Bonding Be Classified as Hydrogen Bonding? A Systematic Investigation of C-H...C(π) Bonding to Cyclopentadienyl Anions. *Chem Eur J* **5**, 1852-1861.

Harding, M. M. (1999). The geometry of metal-ligand interactions relevant to proteins. *Acta Cryst* **D55**, 1432-1443.

Harnett, D. L. & Murphy, J. L. (1975). *Introductory Statistical Analysis*, Addison-Wesley Publishing Co., Philippines.

Harris, J. W. & Stocker, H. (1998). *Handbook of Mathematics and Computational Science*, Springer, New York.

Hartree, D. R. (1925). The atomic structure factor in the intensity of reflexion of X-rays by crystals. *Phil Mag* **50**, 289-306.

Head, R. D., Smythe, M. L., Oprea, T. I., Waller, C. L., Green, S. M. & Marshall, G. R. (1996). VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J Am Chem Soc* **118**, 3959-3969.

Hemm, K., Aberer, K. & Hendlich, M. (1995). Constituting a Receptor-Ligand Information Base from Quality-Enriched Data. *ISMB* **10**, 170-179.

Hendlich, M. (1998). Databases for protein-ligand complexes. *Acta Cryst* **D54**, 1178-1182.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J Mol Biol* **216**, 167-80.

Hermann, R. B. (1972). Theory of Hydrophobic Bonding. II. The Correlation of Hydrocarbon Solubility in Water with Solvent Cavity Surface Area. *J Phys Chem* **76**, 2754.

Hill, T. L. (1956). *Statistical Mechanics*, McGraw-Hill, New York.

Hilpert, K., Ackermann, J., Banner, D. W., Gast, A., Gubernator, K., Hadvary, P., Labler, L., Müller, K., Schmid, G., Tschopp, T. B. & van de Waterbeemd, H. (1994). Desing and synthesis of potent and highly selective thrombin inhibitors. *J Med Chem* **37**, 3889-3901.

Hirst, J. D. (1998). Predicting ligand binding energies. *Curr Opin Drug Discov Development* **1**, 28-33.

Hitzemann, R. (1988). Molecular dynamics studied by analysis of the X-ray diffuse scattering from lysozyme crystals. *Trends Pharmacol Sci* **9**, 408.

Hodgkin, E. E. & Richards, W. G. (1987). Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int J Quant Chem: Quant Biol Symp* **14**, 105-110.

Hoffmann, D., Kramer, B., Washio, T., Steinmetzer, T., Rarey, M. & Lengauer, T. (1999). Two-Stage Method for Protein-Ligand Docking. *J Med Chem* **42**, 4422-4433.

Holloway, M. K., Wai, J. M., Halgren, T. A., Fitzgerald, P. M., Vacca, J. P., Dorsey, B. D., Levin, R. B., Thompson, W. J., Chen, L. J., de-Solms, S. J., Gaffin, N., Ghosh, A. K., Giuliani, E. A., Graham, S. L., Guare, J. P., Hungate, R. W., Lyle, T. A., Sanders, W. M., Tucker, T. J., Wiggins, M., Wiscount, C. M., Woltersdorf, O. W., Young, S. D., Darke, P. L. & Zugay, J. A. (1995). *A priori* prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J Med Chem* **38**, 305-317.

Höltje, H.-D. & Kier, L. B. (1974). Sweet taste receptor studies using model interaction energy calculations. *J Pharm Sci* **63**, 1722-1725.

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science* **268**, 1144-9.

Horovitz, A. (1987). Non-additivity in protein-protein interactions. *J Mol Biol* **196**, 733-735.

Hossain, M. A. & Schneider, H.-J. (1999). Flexibility, Association Constant, and Salt Effects in Organic Ion Pairs: How Single Bonds Affect Molecular Recognition. *Chem Eur J* **5**, 1284-1290.

Hosur, M. V., Bhat, T. N., Kempf, D. J., Baldwin, E. T., Liu, B., Gulnik, S., Wideburg, N. E., Norbeck, D. W., Appelt, K. & Erickson, J. W. (1994). Influence of Stereochemistry on Activity and Binding Modes for $C_2$ Symmetry-Based Diol Inhibitors of HIV-1 Protase. *J Am Chem Soc* **116**, 847-855.

Houston, J. G. & Banks, M. (1997). The chemical-biological interface: developments in automated and miniaturised screening technology. *Curr Opin Biotechnol* **8**, 734-740.

Humblet, C. & Dunbar, J. B. (1993). 3D database searching and docking strategies. In *Annual Reports in Medicinal Chemistry* (Venuti, M. C., ed.), Vol. 28, S. 275-284. Academic Press, London.

Hunter, C. A. & Sanders, J. K. M. (1990). The Nature of π-π Interactions. *J Am Chem Soc* **112**, 5525-5534.

Hunter, C. A., Singh, J. & Thornton, J. M. (1991). π-π Interactions: the Geometry and Energetics of Phenylalanine-Phenylalanine Interactions in Proteins. *J Mol Biol* **218**, 837-846.

Israelachvili, J. & Wennerstrom, H. (1996). Role of hydration and water structure in biological and colloidal interactions. *Nature* **379**, 219-225.

Jain, A. N. (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J Comput Aided Mol Des* **10**, 427-40.

Jain, A. N., Koile, K. & Chapman, D. (1994). Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J Med Chem* **37**, 2315.

James, R. W. (1954). *Optical Principles of the Diffraction of X-rays*, Bell, London.

Janin, J. (1995). Elusive affinities. *Proteins* **21**, 30-39.

Janin, J. (1996). For Guldberg and Waage, with love and cratic entropy. *Proteins* **24**, R1-R2.

Jeffrey, G. A. (1997). *An Introduction to Hydrogen Bonding*, Oxford University Press, New York.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in Biological Structures*, Springer-Verlag, Berlin.

Jencks, W. P. (1981). On the attribution and additivity of binding energies. *Proc Natl Acad Sci USA* **78**, 4046-4050.

Jensen, F. V. (1996). *Introduction to Bayesian networks*, UCL Press, London.

Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr Opin Struct Biol* **6**, 195-209.

Johnson, M. A. & Maggiora, G. M., Eds. (1990). Concepts and Applications of Molecular Similarity. New York: John Wiley & Sons.

Jones, D. A. & Fitzpatrick, F. A. (1999). Genomics and the discovery of new drug targets. *Curr Opin Chem Biol* **3**, 71-76.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-9.

Jones, D. T. & Thornton, J. M. (1996). Potential Energy Functions for Threading. *Curr Opin Struct Biol* **6**, 210-216.

Jones, G., Willett, P. & Glen, R. C. (1995). Molecular Recognition of Receptor Sites using a Genetic Algorithm with a Description of Desolvation. *J Mol Biol* **245**, 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727-48.

Jorgensen, W. L. (1989). Free Energy Calculations: A Breaktrough for Modeling Organic Chemistry in Solution. *Acc Chem Res* **22**, 184-189.

Jorgensen, W. L. & Ravimohan, C. (1985). Monte Carlo simulation of differences in free energies of hydration. *J Chem Phys* **83**, 3050.

Joseph-McCarthy. (1999). Computational approaches to structure-based ligand design. *Pharm Therap* **84**, 179-191.

Joseph-McCarthy, D., Hogle, J. M. & Karplus, M. (1997). Use of the multiple copy simultaneous search (MCSS) method to design a new class of picornavirus capsid binding drugs. *Proteins* **29**, 32-58.

Karplus, P. A. & Faerman, C. (1994). Ordered water in macromolecular structure. *Curr Opin Struct Biol* **4**, 770-776.

Katz, B. A., Johnson, C. R. & Cass, R. T. (1995). Structure-Based Design of High Affinity Streptavidin Binding Cyclic Peptide Ligands Containing Thioether Cross-Links. *J Am Chem Soc* **117**, 8541-8547.

Kellis, J. T., Nyberg, K. & Fersht, A. R. (1989). Energetics of complementary side-chain pakking in a protein hydrophobic core. *Biochemistry* **28**, 4914.

Kellog, G. E. & Abraham, D. J. (1991). HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J Comput Aided Mol Des* **5**, 545.

Kim, K. H. (1993). Comparison of Classical and 3D QSAR. In *3D QSAR in Drug Design: Theory, Methods and Applications* (Kubinyi, H., ed.), S. 619-642. ESCOM, Leiden.

Kim, K. H., Greco, G. & Novellino, E. (1998). A critical review of recent CoMFA applications. *Persp Drug Discov Design* **12**, 257-315.

Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *J Chem Phys* **3**, 300-313.

Klebe, G. (1993). Structural alignment of molecules. In *3D QSAR in Drug Desin: Theory, Methods and Applications* (Kubinyi, H., ed.), S. 173-199. ESCOM, Leiden.

Klebe, G. (1994). The Use of Composite Crystal-field Environments in Molecular Recognition and the *de Novo* Design of Protein Ligands. *J Mol Biol* **237**, 212-235.

Klebe, G. (1998a). Molecular similarity - a guideline for the design of new protein ligands. In *Rational Molecular Design in Drug Research* (Liljefors, T., Jorgensen, F. S. & Krogsgaard-Larsen, P., Ed.), S. 151-160. Munksgaard, Kopenhagen.

Klebe, G. (1998b). Success Stories in Structure-based Drug Design. *Period biol* **100(Supplement 2)**, 93-98.

Klebe, G. & Abraham, U. (1999). Comparative Molecular Similarity Index Analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J Comput Aided Mol Des* **13**, 1-10.

Klebe, G., Abraham, U. & Mietzner, T. (1994). Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J Med Chem* **37**, 4130.

Klebe, G. & Böhm, H.-J. (1998). Energetic and Entropic Factors Determining Binding Affinity in Protein-Ligand Complexes. *Period biol* **100 (Supplement 2)**, 77-83.

Klebe, G. & Mietzner, T. (1994). A fast and efficient method to generate biologically relevant conformations. *J Comput Aided Mol Des* **8**, 583-606.

Klebe, G., Mietzner, T. & Weber, F. (1999). Methodological developments and strategies for a fast flexible superposition of drug-size molecules. *J Comput Aided Mol Des* **13**, 35-49.

Knegtel, R. M., Bayada, D. M., Engh, R. A., von der Saal, W., van Geerestein, V. J. & Grootenhuis, P. D. (1999). Comparison of two implementations of the incremental construction algorithm in flexible docking of thrombin inhibitors. *J Comput Aided Mol Des* **13**, 167-83.

Knegtel, R. M. A. & Grootenhuis, P. D. J. (1998). Binding-affinities and non-bonded interaction energies. In *3D QSAR in drug design: ligand protein interactions and molecular similarity* (Kubinyi, H., Folkers, G. & Martin, Y. C., Ed.), S. 99-114. Kluwer/Escom, Dordrecht.

Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol* **235**, 1598-613.

Koehl, P. & Delarue, M. (1994). Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins* **20**, 264-78.

Kogan, T. P., Dupré, B., Keller, K. M., Scott, I. L., Bui, H., Market, R. V., Beck, P. J., Voytus, J. A., Revelle, B. M. & Scott, D. (1995). Rational design and synthesis of small molecule, non-oligosaccharide selectin inhibitors: (alpha-D-mannopyranosyloxy)biphenyl-substituted carboxylic acids. *J Med Chem* **38**, 4976-4984.

Kollman, P. (1993). Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem Rev* **93**, 2395-2417.

Kollman, P. A. (1994). Theory of macromolecule-ligand interactions. *Curr Opin Struct Biol* **4**, 240-245.

Kollman, P. A. (1996). Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Acc Chem Res* **29**, 461-469.

Kollman, P. A. & Merz Jr., K. M. (1990). Computer Modeling of the Interactions of Complex Molecules. *Acc Chem Res* **23**, 246-252.

Koppensteiner, W. A. & Sippl, M. J. (1998). Knowledge-based potentials - back to the roots. *Biochemistry (Mosc)* **63**, 247-52.

Koshland, D. E. (1994). Das Schlüssel-Schloß-Prinzip und die Induced-fit-Theorie. *Angew Chem* **106**, 2468-2472.

Kossiakoff, A. A., Randal, M., Guenot, J. & Eigenbrot, C. (1992). Variability of conformations at crystal contacts in BPTI represent true low-energy structures: correspondence among lattice packing and molecular dynamics structures. *Proteins* **14**, 65-74.

Kramer, B., Rarey, M. & Lengauer, T. (1999). Evaluation of the FlexX Incremental Construction Algorithm for Protein-Ligand Docking. *Proteins* **37**, 145-156.

Kroon, J., Kanters, J. A., van Duijneveldt-van de Rijdt, J. G. C. M., van Duijneveldt, F. B. & Vliegenhardt, J. A. (1975). O-H...O Hydrogen Bonds in Molecular Crystals. A Statistical and Quantum-Chemical Analysis. *J Mol Struct* **24**, 109-129.

Krystek, S., Stouch, T. & Novotny, J. (1993). Affinity and Specifity of Serine Endopeptidase-Protein Inhibitor Interactions. Empirical Free Energy Calculations Based on X-ray Crystallograhic Structures. *J Mol Biol* **234**, 661-679.

Kubinyi, H., Ed. (1993). 3D-QSAR in Drug Design: Theory, Methods and Applications. Leiden: ESCOM.

Kubinyi, H. (1997). QSAR and 3D-QSAR in Drug Design. Part 1: Methodology. *Drug Discov Today* **2**, 457-467.

Kubinyi, H. (1998). Structure-based design of enzyme inhibitors and receptor ligands. *Curr Opin Drug Discov Develop* **1**, 4-15.

Kubinyi, H. & Abraham, U. (1993). Practical Problems in PLS Analyses. In *3D QSAR in Drug Design. Theory, Methodes and Applications.* (Kubinyi, H., ed.), S. 717-728. ESCOM, Leiden.

Kubinyi, H., Folkers, G. & Martin, Y. C., Eds. (1997). 3D-QSAR in Drug Design: Recent Advances. Leiden: ESCOM/Kluwer.

Kühlbrandt, W. & Williams, K. A. (1999). Analysis of macromolecular structure and dynamics by electron cryo-microscopy. *Curr Opin Chem Biol* **3**, 537-543.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Trainer, J. A. (1992). The Interdependence of Protein Surface Topography and Bound Water Molecules Revealed by Surface Accessibility and Fractal Density Measures. *J Mol Biol* **228**, 13-22.

Kuntz, I. D. (1992). Structure-based strategies for drug design and discovery. *Science* **257**, 1078-82.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A Geometric Approach to Macromolecule-Ligand Interactions. *J Mol Biol* **161**, 269-288.

Kuntz, I. D., Meng, E. C. & Shoichet, B. K. (1994). Structure-Based Molecular Design. *Acc Chem Res* **27**, 117-123.

Kurinov, I. V. & Harrison, R. W. (1994). Prediction of new serine proteinase inhibitors. *Nat Struct Biol* **1**, 735-743.

Ladbury, J. E. (1996). Just add water! The effect of water on the specifity of protein-ligand binding sites and its potential application to drug design. *Chem Biol* **3**, 973-980.

Lahana, R. (1999). How many leads from HTS. *Drug Discov Today* **4**, 447-448.

Laidig, K. E. & Daggett, V. (1996). Testing the Modified Hydration-Shell Hydrogen-Bond Model of Hydrophobic Effects Using Molecular Dynamics Simulation. *J Phys Chem* **100**, 5616-5619.

Laskowski, R. A., Thornton, J. M., Humblet, C. & Singh, J. (1996). X-SITE: Use of Empirically Derived Atomic Packing Preferences to Identify Favourable Interaction Regions in the Binding Sites of Proteins. *J Mol Biol* **259**, 175-201.

Lau, W. F. & Pettitt, B. M. (1989). Selective Elimination of Interactions: A Method for Assessing Thermodynmic Contributions to Ligand Binding with Application to Rhinovirus Antivirals. *J Med Chem* **32**, 2542-2547.

Lazaridis, T. & Karplus, M. (1998). Discrimination of the Native from Misfolded Protein Models with an Energy Function Including Implicit Solvation. *J Mol Biol* **288**, 477-487.

Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133-152.

Lehn, J. M. (1988). Supramolekulare Chemie - Moleküle, Übermoleküle und molekulare Funktionseinheiten (Nobel-Vortrag). *Angew Chem* **100**, 91-116.

Lemieux, R. U. (1996). How Water Provides the Impetus for Molecular Recognition in Aqeous Solution. *Acc Chem Res* **29**, 373-380.

Lemmen, C. & Lengauer, T. (2000). Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des* **14**, 215-232.

Lengauer, T. & Rarey, M. (1996). Computational methods for biomolecular docking. *Curr Opin Struct Biol* **6**, 402-6.

Levitt, M. & Park, B. H. (1993). Water: now you see it, now you don't. *Structure* **1**, 223-226.

Li, A.-J. & Nussinov, R. (1998). A Set of van der Waals and Coulombic Radii of Protein Atoms for Molecular and Solvent-Accessible Surface Calculation, Packing Evaluation, and Docking. *Proteins* **32**, 111-127.

Liljefors, T. (1998). Progress in Force-Field Calculations of Molecular Interaction Fields and Intermolecular Interactions. In *3D QSAR in Drug Design* (Kubinyi, H., Folkers, G. & Martin, Y., Ed.), Vol. 2, S. 3-17. Kluwer Academic Publishers, Dordrecht.

Lim, M. S. L., Johnston, E. R. & Kettner, C. A. (1993). The Solution Conformation of D-Phe-Pro-Containing Peptides: Implications on the Activity of Ac-D-Phe-Pro-boroArg-OH, a Potent Thrombin Inhibitor. *J Med Chem* **36**, 1831-1838.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* **23**, 3-25.

Lottspeich, F. (1999). Proteomanalyse - ein Weg zur Funktionsanalyse von Proteinen. *Angew Chem* **111**, 2630-2647.

Makino, S. & Kuntz, I. D. (1997). Automated Flexible Ligand Docking Method and Its Application for Database Search. *J Comput Chem* **18**, 1812-1825.

Mancera, R. L. (1996). Towards an understanding of the molecular basis of hydrophobicity. *J Comput Aided Mol Des* **10**, 321-326.

Marcotte, E. M., Pellegrini, M., Ng, H., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting Protein Function and Protein-Protein Interactions form Genome Sequences. *Science* **285**, 751-753.

Mark, A. E. & van Gunsteren, W. F. (1994). Decomposition of the Free Energy of a System in Terms of Specific Interactions. Implications for Theoretical and Experimental Studies. *J Mol Biol* **240**, 167-176.

Märki, H. P., Fischli, W., Binggeli, A., Breu, V., Bur, D., Clozel, J. P., D'Arcy, A., Grüninger, F., Güller, R., Hirth, G., Lave, T., Mathews, S., Müller, M., Oefner, C., Stadler, H., Vieira, E., Wilhelm, M. & Wostl, W. (1997). *9th RSC-SCI Medicinal Chemistry Symposium, Cambridge, UK*.

Marriot, D. P., Dougall, I. A., Meghani, P., Liu, Y. & Flower, D. R. (1999). Lead Generation Using Pharmacophore Mapping and Three-Dimensional Database Searching: Application to Muscarinic $M_3$ Receptor Antagonists. *J Med Chem* **42**, 3210-3216.

Marshall, G. R., Barry, C. D., Bosshard, H. E., Dammkoehler, R. A. & Dunn, D. A. (1979). The Conformational Parameter in Drug Design: The Active Analog Approach. In *Computer-Assisted Drug Design* (Olson, E. C. & Christoffersen, R. E., Ed.), S. 205-226. American Chemical Society, Washington DC.

Martin, Y. C. (1999). Pharmacophore Mapping. In *Designing Bioactive Molecules* (Martin, Y. C., Willett, P. & Heller, S. R., Ed.). Americal Chemical Society, Washington, DC.

Martin, Y. C., Bures, M. G., Danaher, E. A. & DeLazzer, J. (1993). New Strategies That Improve the Efficiency of the 3D Design of Bioactive Molecules. In *Trends in QSAr and Molecular Modelling 92* (Wermuth, C.-G., ed.), S. 20-27. ESCOM, Leiden.

Martin, Y. C., Willet, P. & Heller, S. R., Eds. (1999). Designing Bioactive Molecules. Washington, DC: American Chemical Society.

Massova, I. & Kollman, P. A. (1999). Computational Alanine Scanning To Probe Protein-Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J Am Chem Soc* **121**, 8133-8143.

Matsumura, M., Becktel, W. J. & Matthews, B. W. (1988). Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature* **334**, 406.

McCammon, J. A. (1998). Theory of biomolecular recognition. *Curr Opin Struct Biol* **8**, 245-249.

McCarrick, M. A. & Kollman, P. A. (1999). Predicting relative binding affinities of non-peptide HIV protease inhibitors with free energy perturbation calculations. *J Comput Aided Mol Des* **13**, 109-121.

McDonald, I. K. & Thornton, J. M. (1994). Satisfying Hydrogen Bonding Potential in Proteins. *J Mol Biol* **238**, 777-793.

McMartin, C. & Bohacek, R. S. (1997). QXP: Powerful, rapid computer algorithms for structure-based drug design. *J Comput Aided Mol Des* **11**, 333-344.

MDL. Information Systems Inc., San Leandro, CA.

Meirovitch, H. (1998). Calculation of the Free Energy and the Entropy of Macromolecular Systems by Computer Simulation. In *Reviews in Computional Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 12, S. 1-74. Wiley-VCH, New York.

Meng, E. C., Gschwend, D. A., Blaney, J. M. & Kuntz, I. D. (1993). Orientational sampling and rigid-body minimization in molecular docking. *Proteins* **17**, 266-78.

Meng, E. C., Shoichet, B. K. & Kuntz, I. D. (1992). Automated Docking with Grid-Based Energy Evaluation. *J Comput Chem* **13**, 505-524.

Merz, K. M. & Kollman, P. A. (1989). Free Energy Perturbation Simulations of the Inhibition of Thermolysin: Prediction of the Free Energy of Binding of a New Inhibitor. *J Am Chem Soc* **111**, 5649-5658.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *J Chem Phys* **21**, 1087.

Miranker, A. & Karplus, M. (1991). Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **11**, 29-34.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., Forster, M. J. & Thornton, J. M. (1999a). BLEEP-potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J Comput Chem* **20**, 1177-1185.

Mitchell, J. B. O., Laskowski, R. A., Alex, A. & Thornton, J. M. (1999b). BLEEP-potential of mean force describing protein-ligand interactions: I. Generating potential. *J Comput Chem* **20**, 1165-1176.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* **18**, 534-552.

Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**, 623-44.

Miyazawa, S. & Jernigan, R. L. (1999). Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* **34**, 49-68.

Montgomery, J. A., Niwas, S., Rose, J. D., Secrist III, J. A., Babu, Y. S., Bugg, C. E., Erion, M. D., Guida, W. C. & Ealick, S. E. (1993). Structure-based design of inhibitors of purine nucleoside phosphorylase. 1. 9-(arylmethyl) derivatives of 9-deazaguanine. *J Med Chem* **36**, 55-69.

Morgan, B. P., Holland, D. R., Matthews, B. W. & Bartlett, P. A. (1994). Structure-Based Design of an Inhibitor of the Zinc Peptidase Thermolysin. *J Am Chem Soc* **116**, 3251-3260.

Morris, G. M., Goodsell, D. S., Huey, R. & Olson, A. J. (1996). Distributed automated dokking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J Comput Aided Mol Des* **10**, 293-304.

Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Curr Opin Struct Biol* **7**, 194-199.

Muegge, I. & Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **42**, 791-804.

Muegge, I., Martin, Y. C., Hajduk, P. J. & Fesik, S. W. (1999). Evaluation of PMF Scoring in Docking Weak Ligands to the FK506 Binding Protein. *J Med Chem* **42**, 2498-2503.

Muller, N. (1990). Search for a Realistic View of Hydrophobic Effects. *Acc Chem Res* **23**, 23-28.

Murcko, M. A. (1997). Recent Advances in Ligand Design Methods. In *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 11, S. 1-66. Wiley-VCH, New York.

Murphy, K. P. & Gill, S. J. (1991). Solid model compounds and the thermodynamics of protein unfolding. *J Mol Biol* **222**, 699-706.

Murphy, K. P., Privalov, P. L. & Gill, S. J. (1990). Common features of protein unfolding and dissolution of hydrophobic compounds. *Science* **247**, 559.

Murphy, K. P., Xie, D., Thompson, K., Arnzel, M. & Freire, E. (1994). Entropy loss in biological processes: estimate of translational entropy loss. *Proteins* **18**, 63-67.

Murray, C. W., Auton, T. R. & Eldridge, M. D. (1998). Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J Comput Aided Mol Des* **12**, 503-19.

Murray-Rust, P. & Glusker, J. P. (1984). Directional Hydrogen Bonding to $sp^2$- and $sp^3$-Hybridized Oxygen Atoms and Its Relevance to Ligand-Macromolecule Interactions. *J Am Chem Soc* **106**, 1018-1025.

Nakamura, H. (1996). Roles of electrostatic interaction in proteins. *Quart Rev Biophys* **29**, 1-90.

Navia, M. A. & Chaturvedi, P. R. (1996). Design principles for orally bioavailable drugs. *Drug Discov Today* **1**, 179-189.

Nemethy, G. & Scheraga, H. A. (1962). Structure of Water and Hydrophobic Bonding in Proteins. II. Model for the Thermodynamic Properties of Aqueous Solutions of Hydrocarbons. *J Chem Phys* **36**, 3401.

Ng, K.-C., Meath, W. J. & Allnatt, A. R. (1979). A reliable semi-empirical approach for evaluating the isotropic intermolecular forces between closed-shell systems. An application to the He-He, Ne-Ne, Ar-Ar, Kr-Kr and $H_2$-$H_2$ interactions. *Molec Phys* **37**, 237.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281.

Nicklaus, M. C., Wang, S., Driscoll, J. S. & Milne, G. W. A. (1995). Conformational Changes of Small Molecules Binding to Proteins. *Bioorg Med Chem* **3**, 411-428.

Nissink, J. W. M., Verdonk, M. L. & Klebe, G. (2000). Simple knowledge-based descriptors to predict protein-ligand interactions. Methodology and validation. *in press*.

Novotny, J., Bruccoleri, R. E. & Saul, F. A. (1989). On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry* **28**, 4735-4749.

Obst, U. (1997). Dissertation, ETH Zurich.

Obst, U., Banner, D. W., Weber, L. & Diederich, F. (1997). Molecular recognition at the thrombin active site: structure-based design and synthesis of potent and selective thrombin inhibitors and the X-ray crystal structures of two thrombin-inhibitor complexes. *Chem Biol* **4**, 287-95.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics*, 2B, Wiley & Sons Inc., New York.

Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci USA* **84**, 3086-3090.

Oprea, T. I. & Marshall, G. R. (1998). Receptor-based prediction of binding activities. In *3D Qsar in drug design: ligand protein interactions and molecular similarity* (Kubinyi, H., Folkers, G. & Martin, Y. C., Ed.), S. 3-17. Kluwer/Escom, Dordrecht.

Oprea, T. I. & Waller, C. L. (1997). Theoretical and Practical Aspects of Three-Dimensional Quantitative Structure-Activity Relationships. In *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 11. Wiley-VCH, New York.

Oshiro, C. M. & Kuntz, I. D. (1998). Characterization of receptors with a new negative image: use in molecular docking and lead optimization. *Proteins* **30**, 321-36.

Ota, N. & Brunger, A. T. (1997). Overcoming barriers in macromolecular simulations: non-Boltzmann thermodynamic integration. *Theor Chem Acc* **98**, 407-435.

Ota, N., Stroupe, C., Ferreira-da-Silva, J. M. S., Shah, S. A., Mares-Guia, M. & Brunger, A. T. (1999). Non-Boltzmann Thermodynamic Integration (NBTI) for Macromolecular Systems: Relative Free Energy of Binding of Trypsin to Benzamidine and Benzylamine. *Proteins* **37**, 641-653.

Otzen, D. E. & Fersht, A. R. (1999). Analysis of protein-protein interactions by mutagenesis: direct versus indirect effects. *Protein Eng* **12**, 41-45.

Pace, C. N. (1992). Contribution of the Hydrophobic Effect to Globular Protein Stability. *J Mol Biol* **226**, 29-35.

Page, M. (1977). Entropy, Binding Energy, and Enzymic Catalysis. *Angew Chem Int Ed Engl* **16**, 449-459.

Page, M. I. (1973). The Energetics of Neighbouring Group Participation. *Chem Soc Rev* **2**, 295-323.

Page, M. I. & Jencks, W. P. (1971). Entropic Contributions to Rate Accelerations in Enzymic and Intramolecular Reactions and the Chelat Effect. *Proc Natl Acad Sci USA* **68**, 1678-1683.

Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol* **258**, 367-92.

Pauling, L. & Corey, R. B. (1951). Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc Natl Acad Sci USA* **37**, 729-740.

Pearlman, D. A. (1994). A comparison of alternative approaches to free energy calculations. *J Phys Chem* **98**, 1487-1493.

Pearlman, D. A. (1999). Free Energy Grids: A Practical Qualitative Application of Free Energy Perturbation to Ligand Design Using the OWFEG Method. *J Med Chem* **42**, 4313-4324.

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* **132**, 185-219.

Pertsemlidis, A., Saxena, A. M., Soper, A. K., Head-Gordon, T. & Glaeser, R. M. (1996). Direct evidence for modified solvent structure within the hydration shell of a hydrophobic amino acid. *Proc Natl Acad Sci USA* **93**, 10769.

Petsco, G. A. (1996). For medicinal purposes. *Nature* **384**, 7-9.

Pfeifer, S., Pflegel, P. & Borchert, H.-H. (1995). *Biopharmazie - Pharmakokinetik, Bioverfügbarkeit, Biotransformation*, Ullstein/Mosby GmbH & Co.KG, Berlin.

Pickett, S. D. & Sternberg, M. J. (1993). Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol* **231**, 825-39.

Plummer, M. S., Shahripour, A., Kaltenbronn, J. S., Lunney, E. A., Steinbaugh, B. A., Hamby, J. M., Hamilton, H. W., Sawyer, T. K., Humblet, C., Doherty, A. M., Taylor, M. D., Hingorani, G., Batley, B. L. & Rapundalo, S. T. (1995). Design and synthesis of renin inhibitors: incorporation of transition-state isostere side chains that span from the S1 to the S3 binding pockets and examination of P3-modified renin inhibitors. *J Med Chem* **38**, 2893-2905.

Polticelli, F., Ascenzi, P., Bolognesi, M. & Honig, B. (1999). Structural determinants of trypsin affinity and specifity for cationic inhibitors. *Prot Sci* **8**, 2621-2629.

Poornima, C. S. & Dean, P. M. (1995a). Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J Comput Aided Mol Des* **9**, 500-512.

Poornima, C. S. & Dean, P. M. (1995b). Hydration in drug design. 2. Influence of local site surface shape on water binding. *J Comput Aided Mol Des* **9**, 513-520.

Poornima, C. S. & Dean, P. M. (1995c). Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *J Comput Aided Mol Des* **9**, 521-531.

Postma, J. P. M., Berendsen, H. J. C. & Haak, J. R. (1981). Thermodynamics of Cavity Formation in Water. *Faraday Symp Chem Soc* **17**, 55-67.

Quiocho, F. A., Wilson, D. K. & Vyas, N. K. (1989). Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature* **340**, 404-7.

Raag, R. & Poulos, T. L. (1991). Crystal Structures of Cytochrome $P450_{cam}$ Complexed with Camphane, Thiocamphor, and Adamantane: Factors Controlling P450 Substrate Hydroxylation. *Biochemistry* **30**, 2674-2684.

Radmer, R. J. & Kollman, P. A. (1998). The application of three approximative free energy calculation methods to structure based ligand design: Trypsin and its complex inhibitors. *J Comput Aided Mol Des* **12**, 215-227.

Rarey, M., Kramer, B. & Lengauer, T. (1995). Time-efficient docking of flexible ligands into active sites of proteins. *Ismb* **3**, 300-308.

Rarey, M., Kramer, B. & Lengauer, T. (1997). Multiple automatic base selection: protein-ligand docking based on incremental construction without manual intervention. *J Comput Aided Mol Des* **11**, 369-84.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996a). A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**, 470-89.

Rarey, M., Wefing, S. & Lengauer, T. (1996b). Placement of medium-sized molecular fragments into active sites of proteins. *J Comput Aided Mol Des* **10**, 41-54.

Reddy, M. R., Bacquet, R. J., Zichi, D., Matthews, D. A., Welsh, K. M., Jones, T. R. & Freer, S. (1992). Calculation of Solvation and Binding Free Energy Differences for Folate-Based Inhibitors of the Enzyme Thymidilate Synthase. *J Am Chem Soc* **114**, 10117-10122.

Reddy, M. R., Viswanadhan, V. N. & Erion, M. D. (1998). Rapid Estimation of Relative Binding Affinities of Enzyme Inhibitors. In *3D QSAR in drug design: ligand protein interactions and molecular similarity* (Kubinyi, H., Folkers, G. & Martin, Y. C., Ed.). Kluwer/Escom, Leiden.

Ren, J., Esnouf, R., Garman, E., Somers, D., Ross, C., Kirby, I., Keeling, J., Dardy, G., Jones, Y., Stuart, D. & Stammers, D. (1995). High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat Struct Biol* **2**, 293-302.

Reynolds, J. A., Gilbert, D. B. & Tanford, C. (1974). *Proc Natl Acad Sci USA* **71**, 2925.

Roe, S. M. & Teeter, M. M. (1993). Patterns for prediction of hydration around polar residues in proteins. *J Mol Biol* **229**, 419-27.

Rognan, D., Lauemoller, S. L., Holm, A., Buus, S. & Tschinke, V. (1999). Predicting Binding Affinities of Protein Ligands from Three-Dimensional Models: Application to Peptide Binding to Class I Major Histocompatibility Proteins. *J Med Chem* **42**, 4650-4658.

Ross, P. D. & Rekharsky, M. V. (1996). Thermodynamics of hydrogen bond and hydrophobic interactions in cyclodextrin complexes. *Biophys. J.* **71**, 2144.

Rost, B. (1998). Marrying structure and genomics. *Structure* **6**, 259-263.

Roth, G. J., Stanford, N. & Majerus, P. W. (1975). Acetylation of prostaglandin synthase by aspirin. *Proc Natl Acad Sci USA* **72**, 3073.

Rouvray, D. H. (1995). Similarity in chemistry: Past, present and future. In *Molecular Similarity I* (Sen, K., ed.), Vol. 173, S. 1-30. Springer-Verlag, Heidelberg.

Samanta, U., Pal, D. & Charkrabarti, P. (1999). Packing of aromatic rings against tryptophan residues in proteins. *Acta Cryst* **D55**, 1421-1427.

Samudrala, R. & Moult, J. (1998). An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction. *J Mol Biol* **275**, 895-916.

Sanz, F., Giraldo, J. & Manaut, F., Eds. (1995). QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications. Barcelona: J. R. Prous.

Schaefer, M., van Vlijmen, H. W. T. & Karplus, M. (1998). Electrostatic contributions to molecular free energies in solution. *Adv Prot Chem* **51**, 1-57.

Searle, M. S. & Williams, D. H. (1992). The Cost of Conformational Order: Entropy Changes in Molecular Association. *J Am Chem Soc* **114**, 10690-10697.

Searle, M. S., Williams, D. H. & Gerhard, U. (1992). Partitioning of Free Energy Contributions in the Estimation of Binding Constants: Residual Motions and Consequences for Amide-Amide Hydrogen Bond Strengths. *J Am Chem Soc* **114**, 10697-10704.

Sen, S. & Nilsson, L. (1999). Some Practical Aspects of Free Energy Calculations from Molecular Dynamics Simulation. *J Comput Chem* **20**, 877-885.

Serrano, L., Neira, J.-L., Sancho, J. & Fersht, A. R. (1992). Effect of alanine versus glycine in alpha-helices on protein stability. *Nature* **356**, 453.

Sharman, G. J., Searle, M. S., Benhamu, B., Groves, P. & Williams, D. H. (1995). Kooperative Verstärkung elektrostatischer Bindungen durch das Verbergen von Kohlenwasserstoffen. *Angew Chem* **107**, 1644-1646.

Sharp, K. A., Nicholls, A., Friedman, R. & Honig, B. (1991). Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry* **30**, 9686-97.

Shoichet, B., Stroud, R., Santi, D., Kuntz, I. & Perry, K. (1993). Structure-Based Discovery of Inhibitors of Thymidylate Synthase. *Science* **259**, 1445-1450.

Shoichet, B. K., Bodian, D. L. & Kuntz, I. D. (1992). Molecular Docking Using Shape Descriptors. *J Comput Chem* **13**, 380-397.

Shoichet, B. K., Leach, A. R. & Kuntz, I. D. (1999). Ligand Solvation in Molecular Docking. *Proteins* **34**, 4-16.

Shortle, D., Stites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* **29**, 8033.

Silvermann, R. B. (1994). *Medizinische Chemie für Organiker, Biochemiker und Pharmazeutische Chemiker*, VCH, Weinheim.

Silverstein, K. A. T., Haymet, A. D. J. & Dill, K. A. (1998). A Simple Model of Water and the Hydrophobic Effect. *J Am Chem Soc* **120**, 3166-3175.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859-83.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des* **7**, 473-501.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr Opin Struct Biol* **5**, 229-35.

Sippl, M. J., Ortner, M., Jaritz, M., Lackner, P. & Flöckner, H. (1996). Helmholtz free energies of atom pair interactions in proteins. *Folding & Design* **1**, 289-298.

Skolnick, J. & Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic area. *Trends Biotechnol* **18**, 34-39.

Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci* **6**, 676-88.

So, S.-S. & Karplus, M. (1999). A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J Comput Aided Mol Des* **13**, 243-258.

Spaltmann, F., Blunck, M. & Ziegelbauer, K. (1999). Computer-aided target selection - priotizing targets for antifungal drug discovery. *Drug Discov Today* **4**, 17-26.

Spark, M. J., Winkler, D. A. & Andrews, P. R. (1982). Conformational Analysis of Folates and Folate Analogues. *Int J Quantum Chem* **9**, 321.

Srinivasan, J., Cheatham, T. E., Cieplak, P. & Kollman, P. A. (1998). Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J Am Chem Soc* **120**, 9401-9409.

Stahl, M. & Böhm, H.-J. (1998). Development of filter functions for protein-ligand docking - Fast, fully automated docking of flexible ligands to protein binding sites. *J Molec Graph Model* **16**, 121-132.

Stahl, M., Rarey, M. & Klebe, G. (2000). Screening of Drug Databases. *in press*.

Sternberg, M. J. E., Bates, P. A., Kelley, L. A. & MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol* **9**, 368-373.

Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **112**, 6127-6129.

Straatsma, T. P. (1996). Free Energy by Molecular Simulation. In *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 9, S. 81-127. VCH Publishers, New York.

Sturtevant, J. M. (1977). Heat capacity and entropy changes in processes involving proteins. *Proc Natl Acad Sci USA* **74**, 2236-2240.

SYBYL. Molecular Modeling Software 6.6 edit., Tripos Inc., St. Louis, MO.

Takamatsu, Y. & Itai, A. (1998). A New Method for Predicting Binding Free Energy Between Receptor and Ligand. *Proteins* **33**, 62-73.

Tame, J. R. H. (1999). Scoring functions: A view from the bench. *J Comput Aided Mol Des* **13**, 99-108.

Tame, J. R. H., Sleigh, S. H., Wilkinson, A. J. & Ladbury, J. E. (1996). The role of water in sequence-independent ligand binding by an oligopeptide transporter protein. *Nat Struct Biol* **3**, 998-1001.

Tame, J. R. H. & Wilkinson, A. J. (1994). The structural basis of sequence-independent peptide binding by OppA. *Science* **264**, 1578-1581.

Tanaka, S. & Scheraga, H. A. (1976). Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules* **9**, 945-950.

Tanford, C. (1980). *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, Wiley, New York.

Taylor, R. & Kennard, O. (1982). Crystallographic Evidence for the Existence of C-H...O, C-H...N, C-H...Cl Hydrogen Bonds. *J Am Chem Soc* **104**, 5063-5070.

Teague, S. J., Davis, A. M., Leeson, P. D. & Oprea, T. (1999). Design kombinatorischer Leitstruktur-Bibliotheken. *Angew Chem* **111**, 3962-3967.

Tembe, B. L. & McCammon, J. A. (1984). Ligand-Receptor Interactions. *Comput Chem* **4**, 281.

Terrett, N. K., Gardner, M., Gordon, D. W., Kobylecki, R. J. & Steele, J. (1995). Combinatorial Synthesis - The Design of Compound Libraries and their Application to Drug Discovery. *Tetrahedron* **51**, 8135-8173.

Thibaut, U., Folkers, G., Klebe, G., Kubinyi, H., Merz, A. & Rognan, D. (1993). Recommendations to CoMFA Studies and 3D QSAR Publications. In *3D QSAR in Drug Design. Theory, Methods and Applications* (Kubinyi, H., ed.), S. 711-716. ESCOM, Leiden.

Thomas, P. D. & Dill, K. A. (1996). Statistical Potentials Extracted From Protein Structures: How Accurate Are They? *J Mol Biol* **257**, 457-469.

Thorson, J. S., Chapman, E., Murphy, E. C., Schultz, P. G. & Judice, J. K. (1995). Linear Free Energy Analysis of Hydrogen Bonding in Proteins. *J Am Chem Soc* **117**, 1157-1158.

Tintelnot, M. & Andrews, P. (1989). Geometries of functional group interactions in enzyme-ligand complexes: guides for receptor modelling. *J Comput Aided Mol Design* **3**, 67-84.

Torda, A. E. (1997). Perspectives in protein-fold recognition. *Curr Opin Struct Biol* **7**, 200-5.

Tronrud, D. E., Holden, H. M. & Matthews, B. W. (1987). Structures of Two Thermolysin-Inhibitor Complexes That Differ by a Single Hydrogen Bond. *Science* **235**, 571-574.

UNITY. Chemical Information Software 4.1 edit., Tripos Inc., St. Louis, MO.

Vacca, J. P. & Condra, J. H. (1997). Clinically effective HIV-1 protease inhibitors. *Drug Discov Today* **2**, 261-272.

Vajda, S., Sippl, M. & Novotny, J. (1997). Empirical potentials and functions for protein folding and binding. *Curr Opin Struct Biol* **7**, 222-8.

Vajda, S., Weng, Z., Rosenfeld, R. & DeLisi, C. (1994). Effect of Conformational Flexibility and Solvation on Receptor-Ligand Binding Free Energies. *Biochemistry* **33**, 13977-13988.

van de Waterbeemd, H., Ed. (1995). Chemometric Methods in Molecular Design. Vol. 2. Methods and Principles in Medicinal Chemistry Series. Weinheim: VCH Publishers.

van Gunsteren, W. F. & Berendsen, H. J. C. (1990). Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew Chem Int Ed Engl* **29**, 992.

van Vlijmen, H. W. T., Schaefer, M. & Karplus, M. (1998). Improving the Accuracy of Protein pKa Calculations: Conformational Averaging Versus the Average Structure. *Proteins* **33**, 145-158.

Vandonselaar, M., Hickie, R. A., Quail, J. W. & Delbaere, L. T. J. (1994). Trifluoperazine-induced conformational change in Ca(2+)-calmodulin. *Nat Struct Biol* **1**, 795-801.

Veale, C. A., Damewood, J. R., Steelman, G. B., Bryant, C., Gomes, B. & Williams, J. (1995). Nonpeptidic inhibitors of human leukocyte elastase. 5. Design, synthesis, and X-ray crystallography of a series of orally active 5-aminopyrimidin-6-one-containing trifluoromethyl ketones. *J Med Chem* **38**, 86-97.

Vedani, A., Zbinden, P., Snyder, J. P. & Greenidge, P. A. (1995). Pseudoreceptor Modeling: The Construction of Three-Dimensional Receptor Surrogates. *J Am Chem Soc* **117**, 4987-4994.

Venkatarangan, P. & Hopfinger, A. J. (1999). Prediction of Ligand-Receptor Binding Thermodynamics by Free Energy Force Field Three-Dimensional Quantitative Structure-Activity Relationship Analysis: Application to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *J Med Chem* **42**, 2169-2179.

Verdonk, M. L., Cole, J. C. & Taylor, R. (1999). SuperStar: A knowledge-based approach for identifying interaction sites in proteins. *J Mol Biol* **289**, 1093-108.

Verkhivker, G., Appelt, K., Freer, S. T. & Villafranca, J. E. (1995). Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng* **8**, 677-91.

Verwer, P. & Leusen, F. J. J. (1998). Computer Simulation To Predict Possible Crystal Polymorphs. In *Reviews in Computational Chemistry* (Lipkowitz, K. B. & Boyd, D. B., Ed.), Vol. 12, S. 327-365. Wiley-VCH, New York.

Vieth, M., Hirst, J. D. & Brooks III, C. L. (1998a). Do active site conformations of small ligands correspond to low free-energy solution structures? *J Comput Aided Mol Des* **12**, 563-572.

Vieth, M., Hirst, J. D., Dominy, B. N., Daigler, H. & Brooks III, C. L. (1998b). Assessing Search Strategies for Flexible Docking. *J Comput Chem* **19**, 1623-1631.

Viswanadhan, V. N., Reddy, M. R., Wlodawer, A., Varney, M. D. & Weinstein, J. N. (1996). An Approach to Rapid Estimation of Relative Binding Affinities of Enzyme Inhibitors: Application to Peptidomimetic Inhibitors of the Human Immunodeficiency Virus Type 1 Protease. *J Med Chem* **39**, 705-712.

von Itzstein, M., Dyason, J. C., Oliver, S. W., White, H. F., Wu, W.-Y., Kok, G. B. & Pegg, M. S. (1996). A study of the active site of influenza virus sialidase: an approach to the rational design of novel anti-influenza drugs. *J Med Chem* **39**, 388-391.

Wade, R. C. (1998). Molecular Interaction Fields. In *3D QSAR in Drug Design* (Kubinyi, H., Folkers, G. & Y., M., Ed.), Vol. 2, S. 486-505. Kluwer Academic Publishers, Dordrecht.

Wall, I. D., Leach, A. R., Salt, D. W., Ford, M. G. & Essex, J. W. (1999). Binding Constants of Neuraminidase Inhibitors: An Investigation of the Linear Interaction Energy Method. *J Med Chem* **42**, 5142-5152.

Waller, C. L. & Marshall, G. R. (1993). Three-Dimensional Quantitative Structure-Activity Relationship of Angiotensin-Converting Enzyme and Thermolysin Inhibitors. II. A Comparison of CoMFA Models Incorporating Molecular Orbital Fields and Desolvation Free Energies Based on Active-Analog and Complementary-Receptor-Field Alignement Rules. *J Med Chem* **36**, 2390-2403.

Waller, C. L., Oprea, T. I., Giolitti, A. & Marshall, G. R. (1993). 3D-QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. I. A CoMFA Study Employing Experimentally Determined Alignment Rules. *J Med Chem* **36**, 4152.

Wallqvist, A. & Covell, D. G. (1996). Docking enzyme-inhibitor complexes using a preference-based free-energy surface. *Proteins* **25**, 403-19.

Wallqvist, A., Jernigan, R. L. & Covell, D. G. (1995). A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci* **4**, 1881-903.

Walters, W. P., Stahl, M. T. & Murcko, M. A. (1998). Virtual screening - an overview. *Drug Discov Today* **3**, 160-178.

Walther, D. & Cohen, F. E. (1999). Conformational attractors on the Ramachandran map. *Acta Cryst* **D55**, 506-517.

Wang, H. & Ben-Naim, A. (1996). A Possible Involvement of Solvent-Induced Interactions in Drug Design. *J Med Chem* **39**, 1531-1539.

Wang, J., Dixon, R. & Kollman, P. A. (1999a). Ranking Ligand Binding Affinities With Avidin: A Molecular Dynamics-Based Interaction Energy Study. *Proteins* **34**, 69-81.

Wang, J., Kollman, P. A. & Kuntz, I. D. (1999b). Flexible Ligand Docking: A Multistep Strategy Approach. *Proteins* **36**, 1-19.

Wang, R., Liu, L., Lai, L. & Tang, Y. (1998). SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. *J Mol Model* **4**, 379-394.

Warr, W. A. (1997). Combinatorial Chemistry and Molecular Diversity. An Overview. *J Chem Inf Comput Sci* **37**, 134-140.

Warshel, A. & Papazyan, A. (1998). Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr Opin Struct Biol* **8**, 211-217.

Warwicker, J. & Watson, H. C. (1982). Calculation of the Electric Potential in the Active Site Cleft due to α-Helix Dipoles. *J Mol Biol* **157**, 671-679.

Weber, I. T. & Harrison, R. W. (1998). Molecular Mechanics Calculations on Protein-Ligand Complexes. In *3D QSAR in Drug Design* (Kubinyi, H., Fokers, G. & Martin, H., Ed.), Vol. 2, S. 115-127. Kluwer Academic Publishers, Dordrecht.

Weber, P. C., Pantoliano, M. W., Simons, D. M. & Salemme, F. R. (1994). Structure-based Design of Synthetic Azobenzene Ligands for Streptavidin. *J Am Chem Soc* **116**, 2717-2727.

Weber, P. C., Wendoloski, J. J., Pantoliano, M. W. & Salemme, F. R. (1992). Crystallographic and Thermodynamic Comparison of Natural and Synthetic Ligands Bound to Streptavidin. *J Am Chem Soc* **114**, 3197-3200.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J Am Chem Soc* **106**, 765-784.

Weisberg, S. (1985). *Applied Linear Regression*, Wiley, New York.

Wells, J. A. (1990). Additivity of Mutational Effects in Proteins. *Biochemistry* **29**, 8509-8515.

Weng, Z., Vajda, S. & DeLisi, C. (1996). Prediction of protein complexes using empirical free energy functions. *Prot Sci* **5**, 614-626.

Wermuth, C. G., Ed. (1996). The practice of medicinal chemistry. London: Academic Press.

Westhead, D. R., Clark, D. E. & Murray, C. W. (1997). A comparison of heuristic search algorithms for molecular docking. *J Comput Aided Mol Des* **11**, 209-228.

Westhead, D. R. & Thornton, J. M. (1998). Protein structure prediction. *Curr Opin Biotechnol* **9**, 383-389.

Wieseman, T., Wiliston, S., Brandts, J. & Lin, L. (1989). Rapid Measurement of Binding Constants and Heats of Binding Using a New Titration Calorimeter. *Analyt Biochem* **179**, 131-137.

Wiley, R. A. & Rich, D. H. (1993). Peptidomimetics Derived from Natural Products. *Med Res Rev* **13**, 327-384.

Willett, P. (1995). Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J Mol Recog* **8**, 290-303.

Williams, D. H. & Bardsley, B. (1999). Estimating binding constants - The hydrophobic effect and cooperativity. *Persp Drug Discov Design* **17**, 43-59.

Williams, D. H., Cox, J. P. L., Doig, A. J., Gardener, M., Gerhard, U., Kaye, P. T., Lal, A. R., Nicholls, I. A., Salter, C. J. & Mitchell, R. C. (1991). Toward the Semiquantitive Estimation of Binding Constants. Guides for Peptide-Peptide Binding in Aqueous Solution. *J Am Chem Soc* **113**, 7020-7030.

Williams, D. H., Maguire, A. J., Tsuzuki, W. & Westwell, M. S. (1998). An Analysis of the Origins of a Cooperative Binding Energy of Dimerization. *Science* **280**, 711-714.

Williams, D. H., Searle, M. S., Mackay, J. P., Gerhard, U. & Maplestone, R. A. (1993). Toward an estimation of binding constants in aqueous solution: Studies of associations of vancomycin group antibiotics. *Proc Natl Acad Sci USA* **90**, 1172-1178.

Williams, D. H. & Westwell, M. S. (1998). Aspects of weak interactions. *Chem Soc Rev* **27**, 57-63.

Wlodawer, A. (1994). Rational drug design: The proteinase inhibitors. *Pharmacotherapy* **14**, 9S-20S.

Wlodawer, A., Nachman, J., Gilliland, G. L., Gallagher, W. & Woodward, C. J. (1987). Structure of form III crystals of bovine pancreatic trypsin inhibitor. *J Mol Biol* **198**, 469-480.

Wold, S., Johansson, E. & Cocchi, M. (1993). PLS - Partial Least Squares Projections to Latent Structures. In *3D-QSAR in Drug Design: Theory, Methods and Applications* (Kubinyi, H., ed.). ESCOM, Leiden.

Wold, S., Ruhe, A., Wold, H. & Dunn III, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares Approach to Generalized Inverses. *SIAM J Sci Stat Comput* **5**, 735-743.

Wong, C. F. & McCammon, J. A. (1986). Dynamics and Design of Enzymes and Inhibitors. *J Am Chem Soc* **108**, 3830-3831.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, Wiley, New York.

Yang, A., Sharp, K. A. & Honig, B. (1992). Analysis of the heat capacity dependence of protein folding. *J Mol Biol* **227**, 889-900.

Yu, L., Zhu, C. X., Tse-Dinh, Y. C. & Fesik, S. W. (1996). Backbone dynamics of the C-terminal domain of Escherichia coli topoisomerase I in the absence and presence of single-stranded DNA. *Biochemistry* **35**, 9661-9666.

Zauhar, R. J. & Morgan, R. S. (1985). A New Method for Computing the Macromolecular Electric Potential. *J Mol Biol* **186**, 815-820.

Zhang, T. & Koshland, D. E. (1996). Computational method for relative binding energies of enzyme-substrate complexes. *Prot Sci* **5**, 348-356.

Zidek, L., Novotny, M. V. & Stone, M. J. (1999). Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nat Struct Biol* **6**, 1118-1121.

Zou, X., Sun, Y. & Kuntz, I. D. (1999). Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J Am Chem Soc* **121**, 8033-8043.

Zwanzig, R. J. (1954). High-temperature equation of state by a perturbation method. *J Chem Phys* **22**, 1420-1426.

Holger Gohlke

Höhenweg 3
35041 Marburg

# Lebenslauf

| | |
|---|---|
| Geburtstag: | 7. April 1972 |
| Geburtsort: | Langen / Hessen |
| 08/1978 – 06/1984 | Grundschule und Förderstufe in Dietzenbach |
| 08/1984 – 06/1991 | Leibniz-Gymnasium in Offenbach |
| 10/1991 – 09/1992 | Ableistung des Grundwehrdienstes |
| 10/1992 | Immatrikulation im Studiengang Chemie an der Technischen Universität Darmstadt |
| 09/1994 | Diplom-Vorexamen |
| 02/1995 | Preisträger der Dr.-Anton-Keller-Stiftung des Fachbereichs Chemie der Technischen Universität Darmstadt |
| 12/1996 | Diplom-Hauptexamen |
| 01/1997 – 06/1997 | Diplomarbeit unter Anleitung von Prof. Dr. Dr. h.c. F. W. Lichtenthaler, Institut für Organische Chemie, Technische Universität Darmstadt: *Konformative Eigenschaften und Molekulare Lipophilie Profile von β(1→3)- und β(1→6)-verknüpften Cyclogalacto-furanosiden* |
| 06/1997 | Verleihung des akademischen Grades „Diplom-Ingenieur“ |
| 09/1997 – dato | Anfertigung der Promotionsarbeit unter Anleitung von Prof. Dr. G. Klebe, Institut für Pharmazeutische Chemie, Philipps-Universität Marburg |
| 06/1999 – 10/1999 | Forschungsaufenthalt in der Arbeitsgruppe von Prof. Dr. F. Diederich, Laboratorium für Organische Chemie, Eidgenössische Technische Hochschule Zürich mit einem Stipendium der ESCOM Science Foundation, Leiden, Niederlande |
| 11/1997 – dato | Wissenschaftlicher Mitarbeiter am Institut für Pharmazeutische Chemie, Philipps-Universität Marburg;<br>Betreuung des Studentenpraktikums im ersten Semester |

Holger Gohlke
Höhenweg 3
35041 Marburg

# Erklärung

Ich versichere, daß ich meine Dissertation

„**Entwicklung einer wissensbasierten Bewertungsfunktion zur Struktur- und Affinitätsvorhersage von Protein-Ligand-Komplexen**“

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 28. Mai 2000