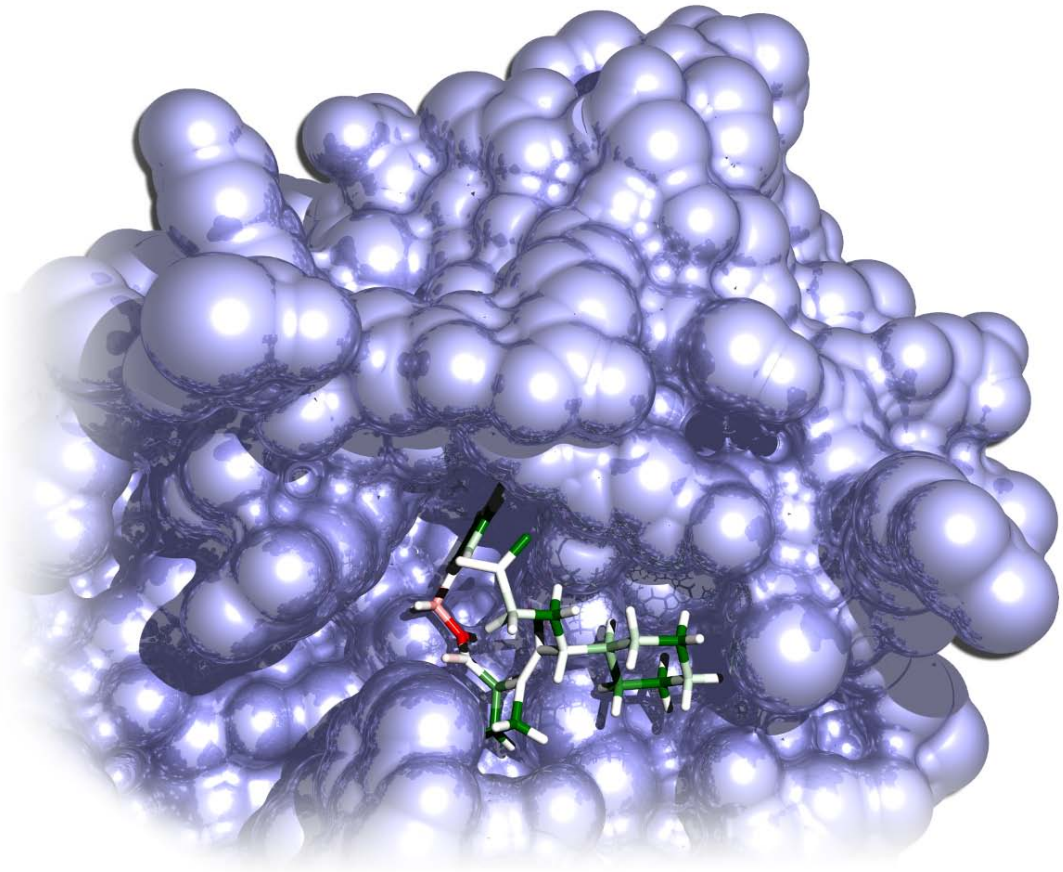
Ingo Reulecke

Adaptives Scoring im strukturbasierten Wirkstoffentwurf.

HYDE - Eine neue Scoring-Funktion für die konsistente Bewertung von Protein-Ligand Komplexen



Universität Hamburg

Zentrum für Bioinformatik ZBH
Algorithmisches Molekulares Design
Bundesstr. 43
20146 Hamburg

Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades

Vorgelegt zur Begutachtung am Department Informatik, Universität Hamburg

Thema	Adaptives Scoring im strukturbasierten Wirkstoffentwurf. HYDE - Eine neue Scoring-Funktion für die konsistente Bewertung von Protein-Ligand Komplexen
Gutachter	Prof. Dr. Matthias Rarey Prof. Dr. Leonie Dreschler-Fischer Dr. Gudrun Lange
Tag der Disputation	5. November 2008

Ingo Reulecke, Hamburg, 16. Februar 2009

Kurzfassung

Die schnelle und präzise Berechnung von Bindungs-Affinitäten zwischen Biomolekülen ist von zentraler Bedeutung für den Computer-gestützten Entwurf neuer Leitstrukturen - sowohl in der pharmazeutischen Forschung als auch im chemischen Pflanzenschutz. Im Rahmen dieser Dissertation wird eine neuartige, empirische Scoring-Funktion HYDE für die Bewertung von Protein-Ligand Komplexen vorgestellt und ihre Umsetzung in einer Software für das virtuelle Screening.

HYDE beruht auf einer Theorie, derzufolge der Beitrag von intermolekularen Wechselwirkungen zur Bindungs-Affinität zwischen zwei Molekülen unmittelbar auf die Eigenschaften des umgebenden Lösungsmittels, dem Wasser, zurückzuführen und auf deren Basis berechenbar ist. Mit einfachen Termen nur für die Desolvatation und für Wasserstoff-Brückenbindungen berücksichtigt HYDE alle entscheidenden Einflüsse auf die Bindungs-Affinität in einem Protein-Ligand Komplex.

Im Gegensatz zu herkömmlichen Scoring-Funktionen berücksichtigt HYDE auch destabilisierende Beiträge infolge der Desolvatisierung, welche einer Komplex-Bildung entgegenwirken. Nur durch eine konsistente Berücksichtigung dieser Beiträge ist eine korrekte Diskriminierung zwischen Bindern und Nicht-Bindern möglich. In umfangreichen theoretischen und experimentellen Untersuchungen konnte gezeigt werden, dass die Verwendung von HYDE zu besseren Vorhersagen im strukturbasierten Wirkstoffentwurf führt.

Für die Umsetzung des Modells wurden neuartige Deskriptoren für molekulare Eigenschaften entwickelt und auf Basis der FlexX Software-Bibliothek implementiert. Erstmals wird die zugängliche Oberfläche eines Moleküls bezüglich der bevorzugten Orientierung von Lösungsmittel-Molekülen gewichtet, was gewährleistet, dass lokale Eigenschaften der Atome wesentlich besser repräsentiert werden als bisher. Neben einer konsistenteren Gewichtung der beiden Scoring-Terme sorgt dies auch für eine gute Adaptierbarkeit der Scoring-Funktion an unterschiedliche Protein-Strukturen.

Herkömmliche Ansätze verwenden zur Parametrisierung strukturelle Informationen von Protein-Ligand Komplexen, in denen destabilisierende Beiträge stark unterrepräsentiert sind. Um dieses Problem zu umgehen, wurde HYDE nur anhand von logP Verteilungskoeffizienten kalibriert, was eine korrekte Quantifizierung aller wesentlichen Energie-Terme ermöglicht.

Abstract

The fast and accurate calculation of binding affinity between biomolecules is of major importance in the context of structure-based drug design. This holds for applications in the pharmaceutical research and in the chemical pest management as well. This thesis introduces a novel empirical scoring function HYDE for the evaluation of protein ligand complexes and its implementation into a virtual screening software.

Due to a novel theory the impact of intermolecular interactions on the binding affinity is directly connected to the properties of the solvent and based on this easily computable. Realizing these assumptions, HYDE estimates binding affinity based on two terms for desolvation and hydrogen bonds only, taking the essential energetic contributions into account.

In contrast to conventional scoring functions HYDE accounts consistently for destabilizing contributions as a result of desolvation, which antagonize the formation of a molecular complex. The correct classification of binders and non-binders seems only feasible by accounting for these contributions. In initial studies, we could demonstrate the improved results in structure-based drug design which are achieved by using the HYDE scoring function both in theoretical and experimental investigations.

Several novel desriptors of molecular properties were developed for the realization of this theoretical model and implemented into the FlexX programming framework. It is the first time, that the accessible surface of a molecule is described with regards to the preferrend orientation of the surrounding solvent molecules. This not only leads to an enhanced consistency between the two scoring terms but also makes the scoring function adaptable to different protein structures, as the local features of individual atoms are much better represented.

In addition, the parameterization of HYDE using logP partition coefficients enables the correct quantification of destabilizing contributions which are not feasible by conventional approaches relying on structural data of protein-ligand complexes.

Inhaltsverzeichnis

1. Einleitung **13**
1.1. Motivation . . . 13
1.2. Projektbeschreibung . . . 15
1.3. Übersicht . . . 15

2. Wirkstoffdesign und Scoring **17**
2.1. Enzyme als pharmazeutisches Target . . . 17
2.1.1. Enzym-Inhibition durch Übergangszustandsanaloga . . . 18
2.1.2. ADME-Parameter und Toxizität . . . 19
2.1.3. Wirkstoffdesign im Pflanzenschutz . . . 21
2.2. Strukturbasierte Methoden im Wirkstoffentwurf . . . 21
2.3. Modellierung von Protein-Ligand Wechselwirkungen . . . 25
2.3.1. Chemische Gleichgewichte . . . 25
2.3.2. Freie Gibbsche Energie . . . 27
2.3.3. Intermolekulare Wechselwirkungen . . . 28
2.4. Scoring-Funktionen zur Approximation von Bindungs-Affinitäten . . . 30
2.4.1. Scoring-Methoden . . . 30
2.4.2. Consensus Scoring . . . 37
2.4.3. Berücksichtigung von Desolvatations-Effekten . . . 38
2.5. Intermolekulare Wechselwirkungen in wässriger Lösung . . . 40

3. HYDE **43**
3.1. Allgemeine Form . . . 43
3.2. Desolvatation . . . 44
3.3. Wasserstoff-Brücken . . . 47
3.4. Bewertung eines Protein-Ligand Komplexes . . . 50

4. Atom-Deskriptoren **59**
4.1. Desolvatation . . . 59
4.1.1. Atombasierte logP Inkrement-Verfahren . . . 59
4.1.2. Partielle logP-Werte als Maß der Desolvatations-Energie . . . 60
4.1.3. Parametrisierung durch multiple lineare Regression . . . 61

4.1.4. Berücksichtigung von geladenen Gruppen . . . 62
4.2. Lösungsmittel-Zugänglichkeit . . . 63
4.2.1. Gewichtete Lösungsmittel-zugängliche Oberfläche WSAS . . . 63
4.2.2. Berechnung der WSAS . . . 64
4.2.3. Alternative Methoden zur Berechnung der Zugänglichkeit . . . 66
4.3. Wechselwirkungen . . . 70
4.3.1. Optimierung des Wechselwirkungs-Modells . . . 70
4.3.2. Sättigung als Funktion der Zugänglichkeit . . . 73

5. Implementierung 77
5.1. Beschreibung der Algorithmen . . . 77
5.1.1. Berechnung der Wechselwirkungs-Energie . . . 77
5.1.2. Berechnung der Desolvatations-Energie . . . 78
5.2. Laufzeitanalyse . . . 83
5.3. Alternativen . . . 85

6. Validierung 89
6.1. Anreicherung aktiver Verbindungen . . . 90
6.1.1. Enrichment-Faktoren und Enrichment-Plots . . . 90
6.1.2. Validierung am DUD-Datensatz . . . 92
6.1.3. Estrogen-Rezeptor . . . 93
6.1.4. Thrombin . . . 99
6.1.5. p38 MAP Kinase . . . 103
6.2. Vorhersage von Bindungs-Affinitäten . . . 108
6.3. Virtuelles Screening mit experimenteller Validierung . . . 115
6.4. Diskussion der Ergebnisse . . . 115
6.5. Anwendbarkeit von HYDE . . . 118

7. Zusammenfassung und Ausblick 121
7.1. Zusammenfassung . . . 121
7.2. Ausblick . . . 122

Literaturverzeichnis 123

A. Dokumentation 141
A.1. HYDE Einstellungen . . . 141
A.1.1. Konfiguration . . . 141
A.1.2. SETTINGS . . . 141
A.1.3. LOGP . . . 142
A.1.4. GEOMETRY . . . 143
A.1.5. AMINO und CONTACT . . . 146

A.1.6. GRAPHIC 146
A.2. Parametrisierung (PLOGP Menü) 147

B. plogP Parameter **153**

C. DUD Datensatz **155**

1. Einleitung

1.1. Motivation

Die Entwicklung eines neuen Medikaments ist ein kosten- und zeitintensiver Prozess. Von den frühen Phasen und ersten Untersuchungen im Labor bis hin zur Marktreife eines Präparates entstehen selbst bei vorsichtigen Schätzungen Kosten in der Größenordnung von mehreren hundert Millionen US-$ [2]. Diese Kosten werden unter anderem durch aufwändige Untersuchungen von Verbindungen verursacht, die sich in den klinischen Test als nicht wirksam herausstellen oder Nebenwirkungen im Organismus verursachen. Von mehreren tausend Kandidaten in der Anfangsphase der Entwicklung schafft es nur eine Handvoll bis hin zur Marktreife. Durch aussagekräftige theoretische Modelle und Vorhersagen zu Beginn der Entwicklung eines neuen Wirkstoffs können deutliche Kosteneinsparungen erzielt werden. Eine Methode, die dabei zum Einsatz kommt, ist das strukturbasierte virtuelle Screening, bei dem aus einer Moleküldatenbank mit mehreren hunderttausend Wirkstoff-ähnlichen Molekülen solche herausgesucht werden, die mit großer Wahrscheinichkeit durch eine Komplexbildung mit einem Protein dieses in seiner Funktion beeinträchtigen bzw. inhibieren können. Durch die Anwendung eines solchen Verfahrens ist es möglich, eine große chemische Diversität in den frühen Phasen der Forschung zu gewährleisten und damit Sackgassen im Entwicklungsprozess zu vermeiden [119].

Ein Docking-Programm generiert für ein Protein-Ligand Paar verschiedene Vorschläge, wie ein Komplex zwischen den Molekülen mit großer Wahrscheinlichkeit zustande kommen könnte. Obwohl ein erfahrener Medizinal-Chemiker oder Kristallograph in vielen Fällen vorhersagen kann, ob eine Bindung in der gegebenen Struktur möglich ist, erfordert alleine die Vielzahl an Verbindungen und möglichen Konformationen, die ein Docking-Programm generiert, eine Automatisierung dieser Aufgabe. Essentieller Bestandteil eines Docking-Programms ist daher die sogenannte Scoring-Funktion, die eine gegebene Konformation bewertet und ein physikalisches Maß für die Stärke der Bindung liefert. Neben einer ausreichenden Genauigkeit, die erlaubt zwischen Bindern und Nicht-Bindern zu differenzieren, muss die Scoring-Funktion auch hohe Anforderungen bezüglich der Rechengeschwindigkeit erfüllen, da für jedes der betrachteten Moleküle tausende von möglichen

Konformationen erzeugt und bewertet werden.

Aufgrund der besonderen Bedeutung einer Scoring-Funktion im Kontext des strukturbasierten Wirkstoffdesigns ist die Verbesserung der existierenden Methoden ein sehr aktuelles Thema. Alleine aus der Anzahl der Publikationen auf diesem Gebiet ist die Dringlichkeit ersichtlich, eine genauere Berechnung der Bindungs-Affinität zu ermöglichen. Dementsprechend hoch ist die Anzahl an Reviews und Vergleichs-Studien, in denen Stärken und Schwächen der unterschiedlichen Scoring-Funktionen analysiert werden. Interessanterweise existiert derzeit keine Scoring-Methode, die uneingeschränkt empfohlen wird. Abhängig vom Protein und auch von dem verwendeten Docking-Programm, ergeben unterschiedliche Scoring-Funktion die besten Resultate in solchen Studien [91].

Da die zugrunde liegenden physikalischen Phänomene immer dieselben sind, ist es wahrscheinlich, dass die aktuellen Scoring-Funktionen wesentliche Einflüsse auf die Bindungs-Affinität nicht berücksichtigen oder falsch bewerten. Zum Beispiel werden die grundlegenden Unterschiede in der strukturellen Beschaffenheit der Binderegion von verschiedenen Proteinen üblicherweise nicht bei die Bewertung eines Komplexes in Erwägung gezogen. Auch das Fehlen von sinnvollen Straftermen in den meisten Scoring-Funktionen, etwa für die Dehydratisierung polarer Atome, führt zu schlechten Resultaten. Im Organismus ist ein Protein-Ligand Komplex stets von Wasser-Molekülen des Lösungsmittels umgeben, welche einen essentiellen Einfluss auf die Bindungs-Affinität haben. Durch das Fehlen von Straftermen kommt es beim virtuellen Screening oftmals zu einer großen Anzahl an falsch positiven Hits. Dies sind Moleküle, für die eine gute Bindung vorausgesagt wurde, die experimentell jedoch nicht nachvollziehbar ist.

Im Rahmen dieser Dissertation wird eine neue Scoring-Funktion HYDE für den strukturbasierten Wirkstoffentwurf vorgestellt. Das wesentliche Merkmal dieser Scoring-Funktion ist die konsistente Beschreibung von hydrophobem Effekt, Desolvatation und Wasserstoff-Brücken. HYDE beruht auf einer Theorie, welche die Stärke von nichtbindenden Wechselwirkungen in wässrigem Milieu in Abhängigkeit der Netzwerk-Eigenschaften des Lösungsmittels beschreibt [87]. Im Gegensatz zu den gängigen Scoring-Funktionen berücksichtigt HYDE auch solche Beiträge, die zu einer Schwächung des Protein-Ligand Komplexes führen. Destabilisierende Wechselwirkungen von Protein und Ligand werden quantitativ beschrieben.

Die generelle Anwendbarkeit von HYDE als Scoring-Funktion im strukturbasierten Wirkstoffentwurf wird anhand von einigen Beispiel-Rechnungen und Vergleich mit aktuellen Daten aus der Literatur diskutiert. Zudem liegen erste experimentelle Ergebnisse vor, welche eindrucksvoll belegen, dass HYDE in der Lage ist, aus einer großen Bibliothek von Molekülen solche mit messbarer Affinität zu einem gegebenen Protein auszuwählen.

1.2. Projektbeschreibung

Die Forschungen zu dieser Dissertation fanden im Rahmen des ADASCOR Projekts in enger wissenschaftlicher Kooperation mit der *Scientific Computing* Gruppe der Bayer CropScience AG, Frankfurt am Main, statt. Das Projekt umfasste die Entwicklung einer neuen Scoring-Funktion für den strukturbasierten Wirkstoffentwurf mit dem Ziel einer verbesserten Vorhersageleistung im Protein-Ligand Docking. Neben der Entwicklung der eigentlichen Funktion und der dafür benötigten molekularen Deskriptoren war es ein wesentlicher Schwerpunkt des Projekts, neue Verfahren für eine geometrische und energetische Parametrisierung zu schaffen. Der wissenschaftliche Beitrag des Kooperations-Partners umfasste im Besonderen die theoretischen Grundlagen für die Bewertung von Wasserstoff-Brücken auf Basis der Eigenschaften des Wasserstoff-Brücken-Netzwerkes im Lösungsmittel. Zudem wurden umfangreiche Untersuchungen der neuen Scoring-Funktion erst durch die Expertise des Kooperations-Partners möglich. Die Entwicklung von neuen Termen und Algorithmen zur Berechnung der Bindungsenergien fand ausschließlich am Zentrum für Bioinformatik statt.

Für die Implementation der neuen Scoring Funktion wurde mit Genehmigung der BioSolveIT GmbH die Flex* Software-Bibliothek verwendet. Diese stellt umfangreiche Datenstrukturen und Algorithmen zur Verwaltung und Manipulation von chemischen Strukturen zur Verfügung. Zudem erlaubt die direkte Einbindung der neuen Scoring-Funktion in die FlexX Docking-Software [120] eine direkte Validierung bezüglich der Einsetzbarkeit im virtuellen Screening. HYDE wurde in ANSI C implementiert, eine Dokumentation ist im Anhang der Arbeit zu finden.

1.3. Übersicht

In Kapitel 2 erfolgt eine Übersicht über die grundlegenden Fragestellungen im strukturbasierten Wirkstoffdesign. Es wird erläutert, wie Enzyme die Umsetzung von chemischen Substanzen im Organismus beschleunigen und damit maßgeblich für den Verlauf von Krankheiten verantwortlich sind. Das Prinzip der Übergangszustands wird erläutert und wie man durch Analoga des Überganzgszutandes Enzyme in ihrer Wirkung inhibieren kann. In einer kurzen Übersicht wird gezeigt, wo die Ansatzpunkte des strukturbasierten Wirkstoffdesigns liegen und welche Funktion dabei die Scoring-Funktion einnimmt. Einer Erläuterung der wichtigsten physikalischen Charakteristika von intermolekularen Wechselwirkungen folgt eine Übersicht über die verschiedenen Arten von Scoring-Funktionen, die im strukturbasierten Wirkstoffdesign eingesetzt werden. Im Besonderen wird der ener-

getische Beitrag durch die Desolvatisierung eines Protein-Ligand Komplexes diskutiert, welche eine maßgebliche Rolle in der anschließend vorgestellten Theorie zur Beschreibung von intermolekularen Wechselwirkungen spielt. Diese Theorie, die maßgeblich durch den Kooperations-Partner entwickelt wurde, dient als Grundlage für die Entwicklung von HYDE.

Die Scoring-Funktion HYDE wird in Kapitel 3 vorgestellt. Nach einer Erläuterung der allgemeinen Form der Scoring-Funktion wird näher auf die beiden essentiellen Scoring-Terme für die Desolvatation und Wasserstoff-Brücken und deren Einfluss auf die freie Bindungs-Energie $\Delta G_{binding}$ eingegangen. Diese Terme werden anhand eines Beispiels exemplarisch erläutert.

Die Berechnung wichtiger Deskriptoren, die für den HYDE-Score verwendet werden, wird in Kapitel 4 ausführlicher erläutert. Diese sind die partiellen Verteilungs-Koeffizienten und die gewichtete Zugänglichkeit WSAS als Maß der Desolvatations-Energie sowie die Wasserstoff-Brücken-Sättigung.

In Kapitel 5 werden die zentralen Algorithmen vorgestellt, die für die Implementierung von HYDE verwendet wurden. Es folgt eine ausführliche Analyse des Laufzeit-Verhaltens, und es werden Möglichkeiten einer beschleunigten Berechenbarkeit diskutiert.

Die Anwendbarkeit von HYDE wird in Kapitel 6 diskutiert. Anhand von öffentlich erhältlichen Test-Datensätzen werden Studien vorgestellt, welche zeigen, dass HYDE in der Lage ist, zwischen Inhibitoren und inaktiven Molekülen zu unterscheiden. Es erfolgt ein Vergleich mit experimentellen Daten. Zudem wird das Ergebnis eines ersten Anwendungs-Falles vorgestellt. Vor- und Nachteile von HYDE gegenüber aktuellen Scoring-Methoden werden diskutiert und die Grenzen der Anwendbarkeit aufgezeigt.

Nach einer Zusammenfassung der Arbeit in Kapitel 7 schließt die Dissertation mit einem Ausblick auf Möglichkeiten, HYDE weiter zu verbessern. Eine technische Dokumentation der HYDE Scoring-Funktion sowie eine Übersicht über den in der Validierung verwendeten DUD Datensatz ist im Anhang zu finden.

2. Wirkstoffdesign und Scoring

2.1. Enzyme als pharmazeutisches Target

Noch vor wenigen Jahrzehnten beruhte die Entwicklung neuer Medikamente meist auf zufälligen Beobachtungen. Die Entdeckung des Penizillins durch Alexander Flemming zum Beispiel ist dem Umstand zu verdanken, dass eine von ihm angesetzte Bakterienkultur von Schimmel befallen war und sich in unmittelbarer Nachbarschaft des Schimmelpilzes die Bakterien aufgelöst hatten [44]. Erst viele Jahre später wurde der genaue Wirkmechanismus des Penizillins bekannt und erst durch intensive Forschungen während des zweiten Weltkrieges konnte das Penizillin als Medikament nutzbar gemacht werden. Heutzutage ist durch die Erforschung des menschlichen Genoms und die Aufklärung vieler Protein-Strukturen ein viel größeres Grundlagenwissen vorhanden, das eine gezielte Forschung nach einem wirksamen Therapeutikum erlaubt.

Mit dem Wissen über Enzyme, welche bei bestimmten Krankheiten eine wichtige Rolle spielen, ist häufig der Ansatzpunkt für die Suche nach einer neuen chemischen Leitstruktur gegeben, die im Idealfall irgendwann als neues Medikament auf den Markt kommt. Enzyme sind Proteine, die im Organismus die Rolle eines Biokatalysators übernehmen. Sie beschleunigen eine chemische Umsetzung, indem sie für geeignete Reaktionsbedingungen sorgen. Zum einen binden sie die Edukte einer Reaktion und bringen sie so in räumliche Nähe zueinander, zum anderen senken sie die Aktivierungs-Energie E_A, die für eine Reaktion benötigt wird. Gemäß der Arrhenius-Gleichung

$$k = A \cdot e^{\frac{-E_A}{RT}} \tag{2.1}$$

für eine einfache chemische Umsetzung stehen die Geschwindigkeits-Konstante k und die Aktivierung-Energie E_A in Korrelation zueinander. R ist die Gaskonstante und T die absolute Temperatur. A ist die sogenannte Stoßfrequenz, die ein Maß für die Wahrscheinlichkeit ist, mit der sich die an der Reaktion beteiligten Moleküle treffen. Je niedriger die benötigte Aktivierungs-Energie E_A ist bzw. je größer die Stoßfrequenz A, desto schneller

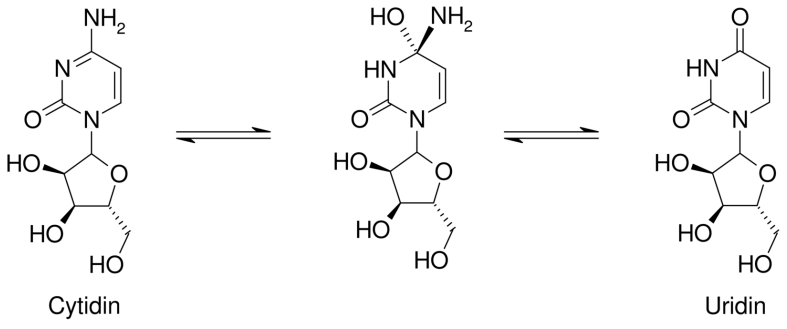


Abbildung 2.1.: Enzymatische Umwandlung von Cytidin zu Uridin. Der Übergangszustand hat eine tetraedrische Geometrie, die durch strukturelle Komplementarität der Bindetasche stabilisiert wird.

kann eine Reaktion ablaufen.

Wenn man es nun schafft ein Enzym in seiner Funktion auszuschalten, das für ein bestimmtes Krankheitsbild eine besondere Rolle spielt, so kann dies einen positiven Effekt auf deren Verlauf haben. Im Falle des Penizillins ist es das Enzym D-Alanin-Transpeptidase, welches eine wichtige Funktion für die Zellwand-Synthese von Bakterien hat. Durch den Einfluss des Penizillins kann es seine Aufgabe nicht mehr erfüllen und die für die Reproduktion der Bakterien notwendige Zellwand-Synthese findet nicht mehr statt.

2.1.1. Enzym-Inhibition durch Übergangszustandsanaloga

Während das Penizillin durch eine zufällige Beobachtung entdeckt wurde, ist es bei genauer Kenntnis der von einem Enzym katalysierten Reaktion theoretisch möglich, einen Inhibitor für das Enzym aufgrund des Reaktionsmechanismusses vorauszusagen. Dazu macht man sich das Prinzip des Übergangszustandes zunutze. Während einer chemischen Umsetzung verändert das Substrat meist seine Geometrie und durchläuft einen metastabilen Zustand, bevor die Reaktion zum endgültigen Produkt erfolgt. Die Aktivierungs-Energie E_A entspricht der Energie-Differenz zwischen Edukt und Übergangszustand und wirkt sich gemäß der Arrhenius-Gleichung 2.1 auf die Reaktionsgeschwindigkeit aus. Ein Enzym stabilisiert den Übergangszustand durch physikalische und geometrische Komplementarität und setzt so die benötigte Aktivierungsenergie herab.

Als Beispiel sei die Umsetzung von Cytidin in Uridin angeführt, die durch die Cytidin-Desaminase katalysiert wird (Abbildung 2.1). Die endständige Amin-Gruppe von Cytidin ist an einen sp^2-hybridisierten Kohlenstoff gebunden und liegt in der Ebene des Sechsringes. Durch den nukleophilen Angriff eines Wasser-Moleküls an das Kohlenstoff-Atom verändert dieses seine Geometrie zu einem sp^3-hybridisierten, tetraedrisch koordinierten Zustand. Amino- und Hydroxyl-Gruppe liegen oberhalb bzw. unterhalb des Ringes. Nach Abspaltung von Ammoniak geht das Kohlenstoff-Atom wieder in den planaren sp^2-Zustand über. Die Cytidin-Desaminase unterstützt diese Reaktion, indem sie den Übergangs-Zustand besser (energetisch günstiger) zu binden vermag als das Cytidin und Uridin. Zudem wird durch ein Zinc-Ion im aktiven Zentrum des Enzyms ein Wasser-Molekül stark polarisiert und so der nukleophile Angriff auf das Kohlenstoff-Atom begünstigt.

Der picomolare Inhibitor 4-Hydroxy-3,4-dihydro-zebularin ist in seiner dreimdimensionalen Struktur dem Übergangszustand dieser Reaktion sehr ähnlich ist und bindet mit perfekter Komplementarität ins aktive Zentrum der Cytidin-Desaminase (Abbildung 2.2). Durch die Bildung des Protein-Ligand Komplexes wird verhindert, dass Substrat-Moleküle ins aktive Zentrum des Enzyms gelangen und die Umsetzung durch das Enzym wird unterbunden. Da dieser Inhibitor der Geometrie des Übergangszustandes der Reaktion nachempfunden wurde, spricht man von einem Übergangszustandsanalogon.

Neben diesem kompetetiven Modus der Inhibition, bei der das natürliche Substrat durch den Inhibitor aus der Bindetasche verdrängt wird, gibt es auch die allosterische Inhibition. Dabei wird durch das Binden eines Inhibitors außerhalb des reaktiven Zentrums eine Konformations-Änderung innerhalb des Proteins bewirkt. Das Substrat bzw. der Übergangszustand der Reaktion bindet nicht mehr im aktiven Zentrum und die enzymatische Umsetzung kann nicht mehr stattfinden.

2.1.2. ADME-Parameter und Toxizität

Die Affinität einer Substanz zu einem Enzym ist eine wichtige Voraussetzung für ihre Wirksamkeit als Therapeutikum. Um die Wirksamkeit jedoch entfalten zu können, muss ein Wirkstoff nach der Applikation zunächst zum Ort der Wirkung gelangen. Dabei spielen Parameter wie die Absorption, Distribution, Metabolisierung und Exkretion eine entscheidende Rolle, welche häufig als ADME-Parameter bezeichnet werden. Die Absorption kann über den Magen, die Schleimhäute, die Haut oder per direkter Injektion in die Blutbahn erfolgen, und je nach Wirkort muss die Substanz wässrige Phasen und Lipidmembranen durchdringen können. Je nach Umgebung und durch Einwirkung verschiedener Enzyme ist die Substanz Edukt von verschiedenen chemischen Reaktionen,

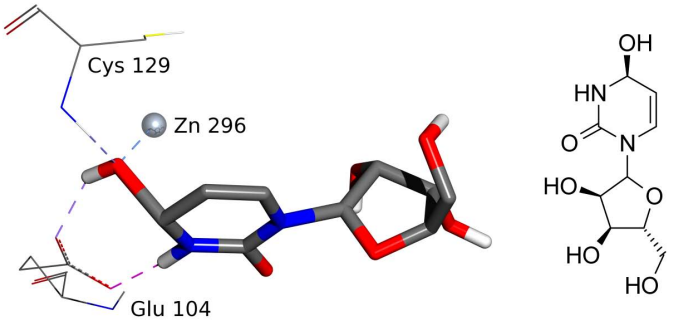


Abbildung 2.2.: 4-Hydroxy-3,4-dihydro-zebularin inhibiert das Reaktionszentrum der Cytidin-Desaminase (Xiang *et al.* [160]).

bis sie schliesslich über Niere, Galle oder Darm vom Körper ausgeschieden wird. Es ist schwierig vorauszusagen, wie ein Wirkstoff beschaffen sein muss, damit er vom Körper in hinreichender Menge aufgenommen wird und zum Wirkort gelangen kann und nicht frühzeitig abgebaut und ausgeschieden wird. Allerdings gibt es grobe Richtlinien wie zum Beispiel die *Rule of Five* von Lipinski [93], welche einen Wertebereich für verschiedene Parameter wie Lipophilie, Basizität, Molekulargewicht oder die Anzahl an Hetero-Atomen angeben, die eine Substanz erfüllen sollte um z.B. im Fall der *Rule of Five* über eine orale Aufnahme zugänglich zu sein.

Von entscheidender Bedeutung für die Anwendbarkeit einer Substanz als Therapeutikum ist ebenso die Toxizität dieser Substanz. Selbst bei hervorragenden Eigenschaften bzgl. Wirksamkeit (Pharmakodynamik) und ADME-Parametern (Pharmakokinetik) kann eine Substanz unerwünschte Nebenwirkungen hervorrufen, die von leichten Irritationen bis hin zum Tod einer Person reichen können. Dies kann direkt als Folge der Inhibition des pharmazeutischen Targets geschehen, welches nicht nur verantwortlich für ein bestimmtes Krankheitsbild sein muss, sondern auch an anderen essentiellen Abläufen im Organismus beteilgt sein kann. Es können aber auch Wechselwirkungen des Wirkstoffs oder seiner Abbauprodukte mit anderen Biomolekülen auftreten, die unerwünschte Nebenwirkungen zur Folge haben. Nicht zuletzt durch Skandale um Medikamente wie Contergan® oder Lipobay® rückt diese Problematik immer wieder ins Licht der Öffentlichkeit und es werden immer stärkere Auflagen für die Zulassung eines Medikaments erhoben, um die körperliche Unversehrtheit der Patienten zu gewährleisten. Bei der Suche nach einem neuen Medikament müssen alle genannten Parameter optimiert werden. Es ist ein ständiges Abwägen zwischen Wirksamkeit und Applizierbarkeit notwendig (Abbildung 2.3).

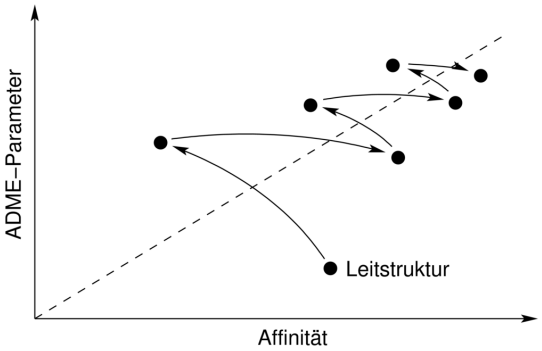


Abbildung 2.3: Typischer Entwicklungsverlauf bei der Optimierung einer Leitstruktur. Verbesserungen in der Bindungs-Affinität bewirken oftmals eine Verschlechterung der ADME-Parameter und umgekehrt.

2.1.3. Wirkstoffdesign im Pflanzenschutz

Enzyme sind nicht nur maßgeblich am Verlauf von Krankheiten im menschlichen Organismus beteiligt, sie sind ebenso essentieller Bestandteil von pflanzlichen Zellen und damit auch Forschungsgegenstand im chemischen Pflanzenschutz. Abhängig vom genauen Einsatzgebiet einer Substanz (entweder als Herbizid, Pestizid oder Insektizid) sind die ADME-Parameter sehr verschieden von denen, die in der pharmazeutischen Forschung beachtet werden müssen. Zudem muss sichergestellt werden, dass eine Substanz sich nicht negativ auf das ökologische Gleichgewicht auswirkt und entsprechend gut abgebaut werden kann. Die Applikation eines Pflanzenschutzmittels erfolgt in wesentlich größeren Mengen als dies bei einem Einsatz als Medikament der Fall ist.

Die Fragestellung, ob eine Substanz ein guter Inhibitor für ein Enzyms ist bzw. ob sie eine möglichst große Affinität zu diesem aufweist, unterscheidet sich jedoch kaum von derjenigen im pharmazeutischen Wirkstoffdesign. Es gelten die selben physikalischen Gesetze und die anwendbaren Methoden für den zielgerichteten Wirkstoffentwurf unterscheiden sich in diesem Punkt nicht voneinander.

2.2. Strukturbasierte Methoden im Wirkstoffentwurf

Die primäre Voraussetzung für die Anwendendbarkeit von strukturbasierten Methoden im Wirkstoffdesign ist natürlich das Vorhandensein von dreidimensionalen Koordinaten des Ziel-Proteins. Solche Koordinaten erhält man durch Röntgenstrukturanalysen oder Kernresonanspektroskopie (NMR). Durch die Protein Datenbank PDB [11] ist eine immer größere Anzahl an aufgeklärten Strukturen von Makromolekülen der Allgemeinheit frei zugänglich (Abbildung 2.4). Großen Pharmaunternehmen stehen zudem üblicherweise

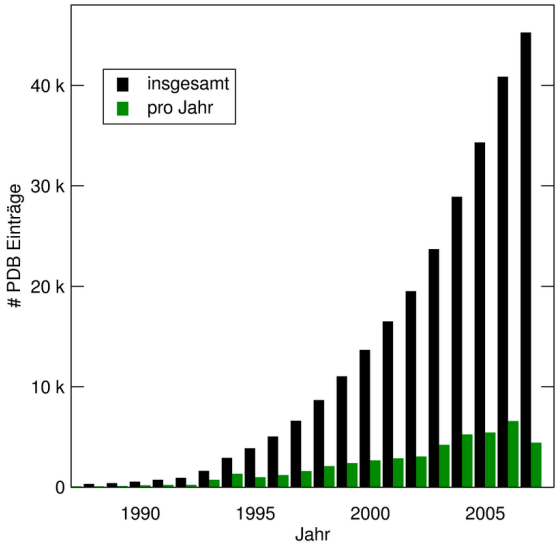


Abbildung 2.4: Anzahl der aufgeklärten Strukturen in der PDB. Dargestellt ist neben der Gesamtzahl der vorliegenden Strukturen (schwarze Balken) auch die Anzahl der pro Jahr neu veröffentlichten Strukturen (grüne Balken). Stand: 15.8.2007.

eigene Kapazitäten für die Struktur-Aufklärung zur Verfügung, so dass ein Ansatzpunkt für strukturbasierte Methoden meist gegeben ist.

Der erste Schritt bei einer vorhandenen Struktur ist nun die Analyse des Proteins und die Suche nach dem aktiven Zentrum bzw. der Bindetasche für einen Inhibitor. Dazu kann man evolutionär basierte Methoden verwenden, welche Bereiche favorisieren, in denen viele konservierte Aminosäuren zu finden sind [6]. Diese Aminosäuren sind mit hoher Wahrscheinlichkeit für die Funktion des Proteins verantwortlich. Aminosäuren in Bereichen, welche eher strukturellen Zwecken dienen, unterliegen häufiger Mutationen als solche in funktionell wichtigen Bereichen des Proteins. Eine weitere Variante ist die geometrische Analyse der Protein-Oberfläche (z.B. durch LIGSITE [64]). Vor allem in hydrophoben Spalten und tiefen Taschen ist eine starke und selektive Bindung eines Inhibitors möglich. Oftmals ist die Lage des aktiven Zentrums aber auch aus aufgeklärten Strukturen von Protein-Ligand Komplexen bekannt oder aus dem Vergleich mit verwandten Proteinen.

Darüber hinaus gehen Methoden, mit denen eine Vorhersage von bevorzugten Wechselwirkungen möglich ist. X-SITE [88] verwendet statistische Daten aus bekannten Protein-Ligand Komplexen. Solche Daten werden aus geometrischen Analysen von PDB Strukturen generiert und sind auch in Datenbanken wie ISOSTAR [20] abrufbar. Das Computer-Programm GRID [57] verwendet kleine Molekül-Fragmente als Sonden, um mithilfe von Kraftfeld-Rechnungen räumliche Potenziale für bestimmte Atom-Typen vorherzusagen.

Diese Daten lassen sich für die gezielte Modellierung von potenziellen Inhibitoren nutzen, welche die pharmakophoren Voraussetzungen des Rezeptors erfüllen. Ein Pharmakophor ist ein dreidimensionales Modell eines oder mehrerer Moleküle, welches die essentiellen Eigenschaften beinhaltet, die für die biologische Aktivität eines Medikaments verantwortlich sind. Dies kann manuell erfolgen - ein geübter Modeller kann anhand der Struktur einer Bindetasche abschätzen, wie ein Inhibitor aufgebaut sein könnte, um alle essentiellen Wechselwirkungen zum Protein auszubilden. Vor allem wenn ein Inhibitor mit niedriger Affinität oder ungewünschten physiologischen Eigenschaften bereits bekannt und in seiner dreidimensionalen Struktur aufgeklärt ist, kann ein Modeller aufgrund seiner Erfahrung gezielt einzelne funktionelle Gruppen eines Liganden austauschen. Dies kann auch aus Gründen der chemischen Zugänglichkeit und Synthetisierbarkeit einer Substanz notwendig sein, welche nicht unwesentlich für die spätere Einsetzbarkeit als Medikament ist. Für dieses sogenannte *scaffold hopping* existieren Computer-Programme, welche entweder für einen gegebenen Teil eines Liganden automatisch Molekül-Fragmente aus einer Bibliothek heraussuchen, welche die geforderten Eigenschaften bezüglich des Pharmakophors erfüllen, oder Teile eines Liganden mutieren (z.B. CAVEAT [89], BREED [118], RECORE [94]). Üblicher ist jedoch die automatische Generierung von komplett neuen Vorschlägen durch *de novo*-Design und Docking [119].

Beim *de novo*-Design [131] werden in der Bindetasche des Proteins Moleküle generiert, welche die gegebenen sterischen und physikalischen Anforderungen des Rezeptors erfüllen. Dafür gibt es unterschiedliche Vorgehensweisen. Im Programm LUDI [14] werden zunächst kleine Molekülfragmente in sogenannten Wechselwirkungszentren der Bindestasche platziert und anschliessend miteinander verbunden. SKELGEN [146, 147] generiert ein Molekülgerüst in der Bindetasche und mutiert dieses mehrfach, um eine Diversität in den erzeugten Molekülen zu gewährleisten. FlexNovo [30] verwendet einen inkrementellen Aufbau-Algorithmus. Nach Platzierung mehrerer möglichst gut bindender Start-Fragmente werden diese über sogenannte Linker-Atome nach und nach durch Anbau von weiteren Fragmenten unter Berücksichtigung der Eigenschaften der Bindetasche bis zum fertigen Molekül erweitert. Die Verknüpfungs-Regeln des verwendeten Fragmentraumes gewährleisten eine gute chemische Synthetisierbarkeit der generierten Moleküle.

Im Gegensatz dazu geht man beim Docking [79] von kompletten Molekülen aus, die möglichst sinnvoll in der Bindetasche eines Proteins platziert werden. Dabei unterscheidet man je nach Vorgehensweise zwischen starren und flexiblen Docking-Methoden. Bei starren Docking-Algorithmen wird ein Ligand-Molekül in einer bzw. mehreren vorberechneten Konformationen in der Bindetasche platziert. Dabei beschränkt sich der Suchraum auf sechs Dimensionen, die die Rotation und Translation des Liganden bezüglich eines durch das Protein definierten Koordinaten-Systems beschreiben. Einer der ersten Ansätze zur Lösung dieses Problems durch das Programm DOCK [86] beruht auf einer Distanz-basierten Cliquen-Suche, bei der Eigenschaften des Liganden auf komplementä-

re Eigenschaften des Rezeptors projeziert werden und die Lösung durch die maximale Anzahl an Übereinstimmungen gegeben ist. Den starren Docking-Algorithmen liegt die Annahme zugrunde, dass der konformationelle Raum des Liganden und des Rezeptor sehr klein ist, dadurch arbeiten entsprechende Programme sehr schnell.

Flexible Docking-Algorithmen sind wesentlich aufwändiger in Bezug auf den Suchraum und die Rechenzeit. Die Anzahl an möglichen Konformationen eines Moleküls steigt exponentiell mit der Anzahl drehbarer Bindungen an. Eine Möglichkeit zur Lösung dieses Problems ist der systematische Aufbau eines in starre Fragmente unterteilten Liganden in der Bindetasche wie zum Beispiel in FlexX [120]. Ein Basisfragment wird im zuvor definierten aktiven Zentrum des Proteins möglichst gut platziert und an die besten Lösungen wird das nächste Fragment mit unterschiedlichen Bindungswinkeln angebaut. Die besten Lösungen werden dann für weitere Anbauschritte herangezogen bis das Molekül komplett ist. Andere Methoden beruhen auf einer stochastischen Suche im konformationellen Raum des Liganden. AutoDock [100] und ICM [4] verwenden Monte Carlo Verfahren und GOLD [72, 73] benutzt einen genetischen Algorithmus zum Auffinden der korrekten Docking-Lösung. Moleküldynamische Simulationen dienen häufig zur lokalen Nachoptimierung, z.B. in ICM und neueren Versionen des DOCK-Programms [39, 40]. Glide [46] versucht den Suchraum möglichst weiträumig abzudecken und verwendet dazu ein hierarchisches Modell, in dem zunächst verschiedene, energieminimierte Konformationen des Liganden starr in der Bindetasche des Rezeptors platziert werden. Nach einer moleküldynamischen Minimierung im Kraftfeld des Rezeptors erfolgt für die aussichtsreichsten Kandidaten ein zusätzliches konformationelles Sampling durch Monte Carlo Rechnungen.

Diese Methoden gehen davon aus, dass ein Rezeptor in einer starren Konformation vorliegt und ein Inhibitor in diesen hineinpasst wie ein Schlüssel in ein Schloss. Allerdings liegt auch ein Rezeptor nicht grundsätzlich in einer starren Konformation vor. Durch Bewegungen in den Seitenketten aber auch durch größere konformationelle Änderungen in der Tertiärstruktur kann sich ein Protein dem jeweiligen Inhibitor anpassen (*induced fit*). Bei der Auswahl der beim Docking verwendeten Protein-Struktur muss man diesen Umstand berücksichtigen und bei bekannter Flexibilität der Bindetasche eventuell mehrere unterschiedliche Strukturen verwenden. Dies kann auch direkt von einem Docking-Programm unterstützt werden. Zum Beispiel erlaubt FlexX-Ensemble [26] die Verwendung mehrerer Protein-Strukturen und unterteilt diese zur Minimierung der Rechenzeit in identische Bereiche und solche, die sich in ihrer Struktur unterscheiden.

Die auf diesen Wegen generierten Vorschläge müssen anschließend durch *in vitro* Experimente auf ihre tatsächliche Wirksamkeit im Organismus überprüft werden. Die exakte Voraussage der Bindungs-Affinität ist eins der größten Probleme im strukturbasierten Wirkstoffdesign. Viele Einflüsse auf die Bindungs-Affinität können aufgrund der

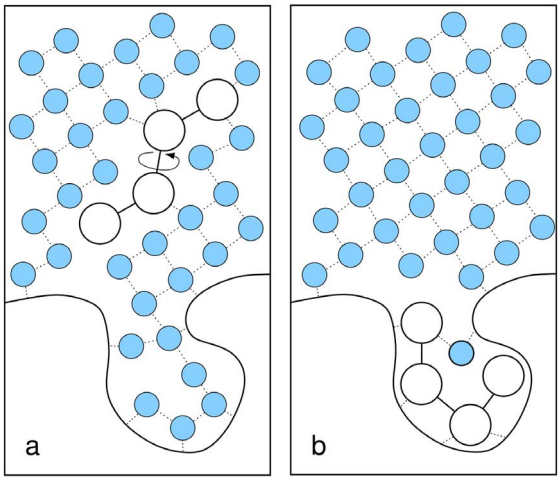


Abbildung 2.5: Schematische Darstellung der Komplexbildung. Im unkomplexierten Zustand (a) sind Protein und Ligand vollständig von Wasser-Molekülen umgeben und bilden Wechselwirkungen zu diesen aus. In (b) ist der Ligand an das Protein gebunden und Wasser-Moleküle wurden aus der Bindetasche verdrängt.

Komplexität des Suchraums nicht im Detail berücksichtigt werden und gängige Scoring-Funktionen zur Bewertung von intermolekularen Wechselwirkungen sind häufig ein Kompromiss zwischen Genauigkeit und Rechengeschwindigkeit.

2.3. Modellierung von Protein-Ligand Wechselwirkungen

In den vorangegangenen Abschnitten wurde beschrieben, wie die Funktion eines Proteins durch das Binden eines Liganden beeinflusst werden kann. Um gezielt auf die Suche nach geeigneten Kandidaten für neue Wirkstoffe gehen zu können, muss man verstehen, wieso es überhaupt zu der Bildung eines Protein-Ligand Komplexes kommt bzw. was die treibenden Kräfte dafür sind. Ein universelles Maß wird benötigt, welches die Stärke einer Bindung bzw. die Affinität eines Liganden zu einem Protein beschreiben kann und Vergleiche zulässt, ob ein bestimmtes Molekül eher als Inhibitor geeignet ist als ein anderes.

2.3.1. Chemische Gleichgewichte

In Abbildung 2.5 ist die Bildung eines Protein-Ligand Komplexes dargestellt. (a) sei der unassoziierte Zustand, in dem sowohl der Ligand als auch das Protein vollständig solvatisiert sind, (b) stellt den Komplex zwischen Ligand und Protein dar. In einem dynamischen Gleichgewicht wird zu jedem Zeitpunkt ein Teil der Moleküle in Zustand (a) vorliegen und ein anderer Teil in Zustand (b). Die Frage nach einem Maß für die

Stärke der Assoziation von Protein und Ligand ist gleichbedeutend mit der Frage, welcher Zustand häufiger vorzufinden ist bzw. dominiert.

Ist das System keinen äußeren Einflüssen wie z.B Neubildung oder Abbau des Liganden unterworfen, ist das Verhältnis zwischen den Molekülen in Zustand (a) und in Zustand (b) immer gleich und man kann das Massenwirkungsgesetz anwenden demzufolge gilt:

$$K_D = \frac{[P] \cdot [L]}{[PL]} \tag{2.2}$$

K_D ist die Gleichgewichtkonstante für die reversible Umsetzung (Dissoziation) gemäß folgender Reaktionsgleichung:

$$PL(aq) \rightleftarrows P(aq) + L(aq) \tag{2.3}$$

P steht für das Protein, L für den Liganden. PL symbolisiert den Protein-Ligand Komplex, jeweils in wässriger Lösung. Die eckigen Klammern in Gleichung 2.2 stehen für die Stoff-Konzentrationen der einzelnen Spezies (genauer: die chemische Aktivität). K_D ist umso kleiner, je höher die Konzentration des Komplexes im Vergleich zum unassoziierten Zustand ist und stellt damit einen geeigneten Wert zur Beschreibung der Bindungs-Affinität zwischen den Molekülen dar. Die Gleichgewichtskonstante für die Bindung K_B beschreibt den umgekehrten Prozess:

$$K_B = 1/K_D \tag{2.4}$$

Je nach Größenordnung von K_D unterscheidet man zwischen milli- (10^{-3}), mikro- (10^{-6}), nano- (10^{-9}), femto- (10^{-12}) und picomolaren (10^{-15}) Inhibitoren. Da es sich bei guten Bindern um sehr kleine Werte handelt, wird üblicherweise der pK_D Wert angegeben, welches der negative dekadische Logarithmus des K_D Wertes ist.

$$pK_D = -\log_{10} K_D \tag{2.5}$$

Eine weitere Möglichkeit, die Güte eines Moleküls als Inhibitor für ein Protein auszudrücken, ist der IC_{50} Wert. Dieser Wert gibt die Konzentration eines Liganden an, welche benötigt wird, um die Aktivität eines Enzyms um 50% zu vermindern. Der IC_{50} steht

mit dem K_D Wert des Systems in unmittelbarem Zusammenhang. Er spiegelt allerdings nicht direkt die Qualität des Protein-Ligand Komplexes wider, sondern die Art und Weise, wie sich die Bindung des Liganden auf die Funktionsweise des Proteins auswirkt, was von weiteren Faktoren wie dem genauen Funktionsmechanismus, dem Ort der Bindung und der Kinetik des Enzyms abhängt.

Obwohl der IC_{50} gerade bei Proteinen mit unterschiedlicher Funktion nur bedingt vergleichbar ist, wird er sehr häufig als Kriterium für die Bindungs-Affinität einer Test-Reihe verwendet. Im Gegensatz zum K_D Wert, der die Konzentration der Moleküle im dynamischen Gleichgewicht beschreibt, lässt sich ein Maß für die Funktion eines Enzyms (und damit der IC_{50}) häufig leichter experimentell bestimmen. Zudem ist aus therapeutischer Sicht die Wirksamkeit eines Inhibitors entscheidender als die reine Bindungs-Affinität. Ein Ligand, welcher mit hoher Affinität weit entfernt vom katalytischen Zentrum bindet, ist als potenzielle Leitstruktur zur Entwicklung eines neuen Therapeutikums wahrscheinlich ungeeignet.

2.3.2. Freie Gibbsche Energie

Die freie Gibbsche Energie G eines Systems setzt sich zusammen aus der Enthalpie H und der Entropie S des Systems. Während die Enthalpie H eine Zustandsfunktion für ein System ist, die maßgeblich von elektronischen Wechselwirkungen beeinflusst wird, ist die Entropie S eine Zustandsfunktion für die Ordnung eines Systems. Nach dem 2. Hauptsatz der Thermodynamik nimmt die Entropie eines Systems ohne äußere Einflüsse niemals ab. Ein System strebt nach einem Zustand maximaler Unordnung. Die Änderung der Gibbschen freien Energie ΔG eines Systems bei konstanter Temperatur ist gegeben durch

$$\Delta G = \Delta H - T\Delta S \tag{2.6}$$

Reaktionen laufen spontan ab, wenn die damit verbundene Änderung in der freien Energie ΔG negativ ist, es wird ein Zustand der geringsten freien Energie angestrebt. Für eine Substanz A in wässriger Lösung berechnet sich die Gibbsche freie Energie aus einem Grundwert $G_0(A)$ und der Konzentration $[A]$.

$$G(A) = G_0(A) + RT \cdot \ln[A] \tag{2.7}$$

R ist die Gaskontante und T die absolute Temperatur. Für den Energieunterschied ΔG zwischen Protein P, Ligand L und Komplex PL ergibt sich entsprechend

$$\Delta G = G(PL) - G(P) - G(L) = (G_0(PL) - G_0(P) - G_0(L)) + RT \cdot \ln \frac{[PL]}{[P] \cdot [L]} \quad (2.8)$$

Im Gleichgewichtszustand ($\Delta G = 0$) gilt das Massenwirkungsgesetz.

$$\frac{[PL]}{[P] \cdot [L]} = K_B \quad (2.9)$$

Damit ist die Gibbsche freie Energie für die Assoziation von P und L gegeben durch

$$\Delta G_{binding} = G_0(PL) - G_0(P) - G_0(L) = -RT \cdot \ln K_B \quad (2.10)$$

Von Bedeutung für die Berechnung von $\Delta G_{binding}$ sind vor allem die Beiträge der nicht-bindenden Wechselwirkungen und Entropie-Änderungen aufgrund von Einschränkungen der Freiheitsgrade der Translation, Rotation und Konformation. Einen wesentlichen Einfluss auf die Bindungs-Energie hat aber auch das Lösungsmittel. Protein und Ligand sind ständig von Wasser-Molekülen umgeben. Eine Freisetzung von Wasser-Molekülen im Interface zwischen Protein und Ligand kann sich sowohl günstig als auch ungünstig auf die Bindungs-Energie auswirken.

2.3.3. Intermolekulare Wechselwirkungen

Bei der Bildung eines Protein-Ligand Komplexes spielen vor allem die nicht-bindenden oder schwachen Wechselwirkungen eine Rolle. Man unterscheidet:

Elektrostatische Wechselwirkungen. Hierzu zählen alle Wechselwirkungen zwischen permanenten elektrischen Momenten. Diese sind weiter unterscheidbar in Coulomb-Kräfte zwischen geladenen Ionen und richtungsabhängige Keesom-Kräfte zwischen Dipol- oder Quadrupol-Momenten. Mit zunehmender Ordnung des elektrischen Moments erhöht sich die Potenz n in der Abhängigkeit des Potentials V vom Abstand r zwischen den Momenten ($V \propto 1/r^n$).

Induktions-Wechselwirkungen (Debye-Kräfte). Durch ein permanentes elektrisches Moment wird in einem anderen Atom bzw. Molekül ein Dipol induziert. Die Stärke dieser Wechselwirkung hängt von der Polarisierbarkeit α des Atoms ab.

Dispersions-Wechselwirkungen (London-Kräfte). Die Ladungsdichte um ein Atom ist zu jedem Zeitpunkt einer Fluktuation unterworfen. Dadurch kommt es zum Auftreten von sich gegenseitig induzierenden Dipolen.

Während die elektrostatischen Wechselwirkungen je nach Art der Ladung und Ausrichtung der elektrischen Momente sowohl abstoßender als auch anziehender Natur sein können, sind Induktions- und Dispersions-Wechselwirkungen immer attraktiv. Sie werden auch als *van der Waals* Kräfte bezeichnet. Kommen sich zwei Atome so nahe, dass sich ihre Elektronen-Hüllen durchdringen, so kommt es aufgrund des Pauli-Prinzips, das in einer einfachen Interpretation verbietet, dass sich zwei Elektronen mit gleichem Spin zu nahe kommen, zu einer starken Abstoßung zwischen den Atomen.

Wasserstoff-Brückenbindungen. Einen Spezialfall der intermolekularen Wechselwirkungen bilden die sogenannten Wasserstoff-Brückenbindungen. Eingeführt wurde dieser Begriff durch Pauling [109]. Dieser hatte beobachtet, dass bei der Interaktion zwischen einem an ein elektronegatives Atom X gebundenem Wasserstoff-Atom und dem freien Elektronenpaar eines anderen elektronegativen Atoms Y in der Konstellation X-H$\cdots$Y das Wasserstoff-Atom keinem der beiden Atome eindeutig zuzuordnen ist und eine überbrückende Position einnimmt. Typischerweise bezeichnet man als Wasserstoff-Brücken solche Wechselwirkungen, in denen die Atome X und Y entweder Stickstoff oder Sauerstoff sind. Obwohl die Wasserstoff-Brücke meist als Dipol-Dipol Wechselwirkung klassifiziert wird, kann sie auch einen kovalenten Anteil haben, der durch die gemeinsame Benutzung von Orbitalen charakterisiert ist. Gerade bei den sogenannten Ladungs-unterstützten Wasserstoff-Brücken treten sehr große Wechselwirkungs-Energien auf, die mit denen von kovalenten Bindungen vergleichbar sind. Abweichend von der klassischen Definition können an Wasserstoff-Brückenbindungen auch C-H Gruppen, Fluor, Schwefel oder π-Elektronen beteiligt sein. Die Stärke einer Wasserstoff-Brücke variiert in einem großen Bereich von etwa 4–15 kJ mol^{-1} (schwache Wasserstoff-Brücke), über 15–60 kJ mol^{-1} bei normalen bis hin zu 60–170 kJ mol^{-1} bei starken Wasserstoff-Brückenbindungen [108].

Wasserstoff-Brückenbindungen sind von besonderer Bedeutung für die Ausbildung eines Protein-Ligand Komplexes. Aufgrund der starken Abhängigkeit der Bindungs-Energie von der gegenseitigen Orientierung der beteiligten Atome muss ein Ligand-Molekül ein sehr spezifisches dreidimensionales Profil seiner funktionellen Gruppen aufweisen, um an einen Rezeptor binden zu können. Zudem sind die Eigenschaften des umgebenden Wassers ebenso auf Wasserstoff-Brücken zurückzuführen. Im Gegensatz dazu wirken sich van der Waals Kräfte nur wenig auf die Bindungs-Affinität eines Komplexes aus. Die Moleküle

sind zu jedem Zeitpunkt, also im Komplex aber auch im ungebundenen Zustand, immer von gleichartigen Atomen umgeben, so dass sich in der Differenz nur ein kleiner Beitrag bei dieser Art von Wechselwirkung ergibt.

2.4. Scoring-Funktionen zur Approximation von Bindungs-Affinitäten

Eine Bewertungs- bzw. Scoring-Funktion verwendet man im Kontext des strukturbasierten Wirkstoffentwurfs, um den Unterschied in der freien Gibbschen Energie $\Delta G_{binding}$ zwischen unassoziiertem Zustand und Protein-Ligand Komplex vorherzusagen und damit die Wahrscheinlichkeit zu bestimmen, mit der ein Molekül in einer spezifischen Konformation an ein Protein binden wird. Anhand einer Scoring-Funktion werden verschiedene Konformationen eines Liganden in der Bindetasche des Proteins miteinander verglichen. Die Konformation, die dem natürlichen Bindungsmodus am ähnlichsten ist, sollte durch eine Scoring-Funktion am besten bewertet werden. Weiterhin sollte der berechnete Wert auch ein abolutes Maß für die Bindungs-Affinität darstellen, da bei der strukurbasierten Suche nach einem Wirkstoff sehr viele potenzielle Liganden miteinander verglichen werden. Anhand der Scoring-Funktion muss zwischen tatsächlichen Bindern und solchen Molekülen, die keine Affinität aufweisen, unterschieden werden. Zudem werden bei einem virtuellen Screening oftmals mehrere Millionen verschiedene Moleküle gedockt und bewertet. Eine Scoring-Funktion muss daher den Spagat zwischen Genauigkeit und Anwendbarkeit (Geschwindigkeit) leisten können.

2.4.1. Scoring-Methoden

Grundsätzlich unterscheidet man im strukturbasierten Wirkstoffentwurf drei unterschiedliche Ansätze zur Bewertung von Protein-Ligand Komplexen. Die Kraftfeld-basierten Scoring-Funktionen verwenden klassische Potenzial-Terme zur Beschreibung der wichtigsten Wechselwirkungen zwischen den Molekülen. Empirische Scoring-Funktionen beruhen auf der mathematischen Beschreibung von beobachtbaren Phänomenen und der Parametrisierung der entsprechenden Terme anhand von experimentell bestimmten Bindungs-Parametern. Sogenannte wissensbasierte Ansätze leiten ein Maß für die Bindungs-Affinität aus der statistischen Häufigkeit her, mit der verschiedene Atom-Paare in strukturell aufgeklärten Komplexen beobachtet werden. Es sei angemerkt, dass sich diese Namensgebung teilweise von der in den Informations-Wissenschaften üblichen Nomenklatur

wissensbasierter Systeme unterscheidet. Zudem sind die Grenzen zwischen den drei anfgeführten Methoden fließend.

Gemäß der zentralen Bedeutung der Scoring-Funktion im strukturbasierten Wirkstoffentwurf existieren zahlreiche Übersichts-Artikel, in denen aktuelle Methoden miteinander verglichen werden [79, 132, 84, 91, 119]. Grundlegende Ansätze der verschiedenen Funktions-Typen seien im folgenden dargestellt.

Kraftfeld-basierte Scoring-Funktionen

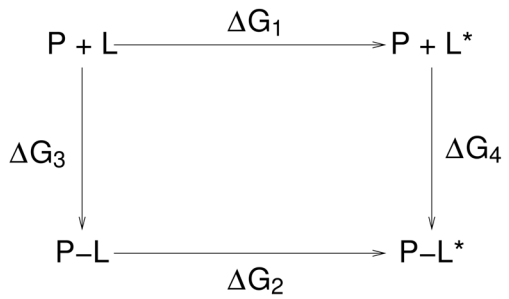


Abbildung 2.6: Ein thermodynamischer Kreislauf für die Berechnung der relativen freien Bindungs-Energie $\Delta\Delta G_{binding}^{L \rightarrow L^*}$ von zwei Liganden L und L^* bezüglich desselben Proteins P.

Mit einem physikalischen Kraftfeld werden die zwischen den Atomen eines oder mehrerer Moleküle wirkenden Kräfte beschrieben. Eine häufig zur Berechnung der Energien in Biomolekülen verwendete Form eines Kraftfeldes ist:

$$\begin{aligned} E_{total} &= \sum_{bond} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 \\ &+ \sum_{dihedrals} \frac{V_n}{2} \left[1 + \cos(n\phi - \gamma)\right] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right] \end{aligned} \tag{2.11}$$

Die Energie E_{total} eines Systems wird berechnet als Summe über Anteile kovalenter Schwingungs-Zustände und nicht-kovalente Bindungs-Energien. Abweichungen von Standard-Bindungslängen r_{eq} und -winkeln θ_{eq} werden mit einem harmonischen Potenzial-Term bestraft. Änderungen im Torsionswinkel ϕ unterliegen einem periodischen Energie-Term. Intermolekulare Kräfte zwischen zwei Atomen i und j werden durch Lennard-Jones- und Coulomb-Ausdrücke beschrieben, die vom Abstand R_{ij} zwischen den Atomen abhängen. Kraftfeld-Gleichungen beinhalten viele Parameter, die jeweils für das

untersuchte molekulare System optimiert werden. Für die Modellierung von Wechselwirkungen in Biomolekülen verwendete Kraftfelder sind z.B. AMBER [29], OPLS-AA [74], CHARMM [95] und GROMOS [105].

Wie aus Gleichung 2.11 ersichtlich ist, beinhalten Kraftfelder keinen Ausdruck für entropische Beiträge zur freien Bindungs-Energie. Da die Entropie S gemäß Gleichtung 2.6 wesentlicher Bestandteil der Gibbschen Energie G ist, werden Kraftfelder häufig in Kombination mit Methoden wie Moleküldynamik (MD) oder Monte Carlo Simulationen (MC) verwendet. Entropische Beiträge werden durch die Generierung von statistischen Ensembles und Anwendung der Boltzmann Gleichung bestimmt [12]. Einflüsse des Lösungsmittels werden entweder durch explizite Berücksichtigung von Wasser-Molekülen in der Simulation oder implizit berechnet, z.B. durch Verwendung einer Abstands-abhängigen dielektrischen Konstante. Diese simuliert die elektrostatische Abschirmung von geladenen Gruppen durch das Lösungsmittel gemäß dem Poisson-Boltzmann Modell [65].

Durch eine entsprechende Simulation erhält man die freie Gibbsche Energie G des betrachteten Protein-Ligand Komplexes, aber per se noch nicht die freie Bindungs-Energie $\Delta G_{binding}$. Um entscheiden zu können, welcher von zwei Liganden L und L^* besser an ein Protein P binden wird, benötigt man einen gemeinsamen Referenz-Zustand. Über den thermodynamischen Kreisprozess in Abbildung 2.6 lässt sich die relative Bindungs-Energie $\Delta\Delta G_{binding}^{L \rightarrow L^*}$ bestimmen. Mit ΔG_3 und ΔG_4 sind die relativen Bindungs-Energien von L bzw. L^* bezüglich des Proteins P bezeichnet. $\Delta\Delta G_{binding}^{L \rightarrow L^*}$ ist die Differenz zwischen ihnen. Da in einem thermodynamischen Kreislauf die Energie unabhängig vom Weg ist, über den sie ermittelt wird, kann man bei Kenntnis von ΔG_3 (z.B. aus einem Experiment) die Bindungs-Energie ΔG_4 über ΔG_1 und ΔG_2 bestimmen. Dies sind die Energie-Beiträge, welche aufgebracht werden müssten, um den Liganden L in den Liganden L^* zu 'transformieren', entweder im unassoziierten Zustand oder im Komplex mit dem Protein. Diese können über eine schrittweise thermodynamische Integration über die Reaktionskoordinate berechnet werden. In der Praxis wird die Transformation zwischen den Systemen durch eine Serie von Simulationen von nicht-physikalischen Übergangszuständen ersetzt. Diese Methode wird *free energy perturbation* (FEP) genannt [80].

$$\Delta\Delta G_{binding}^{L \rightarrow L^*} = \Delta G_4 - \Delta G_3 = \Delta G_2 - \Delta G_1 \tag{2.12}$$

Weiterhin ist es auch möglich, die totale freie Bindungs-Energie ΔG_4 von L^* zu berechnen. Dazu wird anstelle des Liganden L ein *dummy*-Molekül ohne Atome verwendet. Damit ist $\Delta G_3 = 0$ und die relative Bindungs-Energie entspricht ΔG_4 und lässt sich entsprechend Gl. 2.12 direkt aus den thermodynamischen Berechnungen bestimmen. Da die FEP Methode jedoch sehr viele Simulations-Schritte entlang des Reaktions-Pfades benö-

tigt, um die Ensemble-Energien für möglichst viele Überganszustände zu bestimmen, ist sie sehr zeitaufwendig und nur beschränkt für das virtuelle Screening anwendbar.

Åqvist *et al.* haben die *linear interaction energy* (LIE) Methode [3] vorgestellt, bei der die moleküldynamischen Simulationen nur für den Start- und den Endpunkt der Reaktionskoordinate durchgeführt werden. Anstatt eine Integration über viele Zwischenzustände vorzunehmen, wird die Bindungs-Energie durch eine lineare Näherung bestimmt. Die Energie hängt von der Änderung im polaren und im unpolaren Wechselwirkungs-Beitrag vor und nach Bildung des Komplexes ab und wird berechnet wie folgt:

$$\Delta G_{binding} = \alpha \Delta \left\langle V_{l-s}^{vdW} \right\rangle + \beta \Delta \left\langle V_{l-s}^{el} \right\rangle + \gamma \tag{2.13}$$

Dabei entsprechen die $\langle V^{l-s} \rangle$-Terme den Energie-Unterschieden zwischen Start- und Endpunkt. Sie sind aufgeteilt in einen van der Waals Term vdW und elektrostatische Beiträge el des Liganden mit den umliegenden Atomen. Die griechischen Buchstaben α, β und γ stehen für Konstanten, die von der Art des betrachteten Systems abhängen.

Eine weitere Möglichkeit zur Beschleunigung von Kraftfeld-basierten Docking-Methoden beruht darauf, dass vorberechnete Energien für unterschiedliche Arten von Wechselwirkungen mit dem Protein an den Knotenpunkten eines Gitters gespeichert werden, welches die Bindetasche des Proteins ausfüllt. Zur Berechnung der paarweisen Wechselwirkungs-Energien zwischen Protein und Ligand ist es danach nur noch nötig, die entsprechenden Werte aus den zu den Ligand-Atomen benachbarten Knotenpunkten auszulesen. Diese Methode wurde zuerst in der immer noch häufig eingesetzten GRID Software verwendet [57]. Pearlman hat mit der OWFEG [110] Methode (*one-window free energy grid*) eine Erweiterung vorgeschlagen, bei der die Gitterpunkt-Energien direkt aus FEP Rechnungen erhalten werden und somit mit freien Bindungs-Energien korrelieren.

Empirische Scoring-Funktionen

In empirischen Scoring-Funktionen werden beobachtbare geometrische Eigenschaften eines Protein-Ligand Komplexes mit gemessenen freien Gibbschen Bindungs-Energien korreliert. Die berechnete Affinität setzt sich aus einer Summe von Einzelbeiträgen zusammen, welche unterschiedliche physiko-chemische Eigenschaften repräsentieren. Dabei werden sowohl entropische als auch enthalpische Einflüsse berücksichtigt. Eine erste Anwendung im strukturbasierten Wirkstoffentwurf geht zurück auf Böhm [16], der eine solche Formel in das *de novo*-Design Programm LUDI [14, 15] implementiert hat. Eine ver-

änderte Version der Böhm-Funktion kommt im Docking Programm FlexX [120] zum Einsatz:

$$\begin{aligned}\Delta G_{binding} = \Delta G_0 \quad & +\Delta G_{rot} \times N_{rot} \\ & +\Delta G_{hb} \sum_{neutral\ H-bonds} f(\Delta R, \Delta\alpha) \\ & +\Delta G_{io} \sum_{ionic\ int} f(\Delta R, \Delta\alpha) \\ & +\Delta G_{aro} \sum_{aro\ int} f(\Delta R, \Delta\alpha) \\ & +\Delta G_{lipo}\ f^*(\Delta R) \end{aligned} \tag{2.14}$$

Die freie Gibbsche Bindungs-Energie $G_{binding}$ wird in FlexX als Summe von Beiträgen von Wasserstoff-Brücken ΔG_{hb}, ionischen Wechselwirkungen ΔG_{io}, Wechselwirkungen zwischen π-Systemen ΔG_{aro}, hydrophoben Kontakten zwischen Protein- und Ligand-Atomen ΔG_{lipo} sowie der Anzahl frei rotierbarer Bindungen im Liganden ΔG_{rot} berechnet. Im Falle von Wasserstoff-Brücken oder ionischen und aromatischen Wechselwirkungen werden Abweichungen von der experimentell bestimmten, bevorzugten Geometrie mit einem Straf-Faktor multipliziert, der vom Abstand R und Winkel α zwischen den betrachteten funktionellen Gruppen abhängt. Die Anzahl hydrophober Kontakte wird als Maß für den hydrophoben Effekt verwendet, und die Anzahl der frei drehbaren Bindungen im Liganden repräsentiert den Entropieverlust durch den Wegfall von möglichen Konformationen des Liganden im Zuge der Komplex-Bildung. Die jeweiligen Energie-Parameter G_i der einzelnen Terme wurden durch eine Regressionanalyse bestimmt. Die Regressionskonstante G_0 wird als entropischer Beitrag zur Bindungs-Affinität interpretiert, der auf dem Wegfall von Freiheitsgraden der Translation und der Rotation beruht.

Viele der seitdem publizierten empirischen Scoring-Funktionen sind der Böhm Funktion sehr ähnlich (siehe Tabelle 2.1). Sie unterscheiden sich hauptsächlich in der genauen mathematischen Beschreibung der einzelnen Terme. Zum Beispiel wurden für die Modellierung des hydrophoben Effekts anstatt der Kontakt-Fläche paarweise Kontakt-Potenziale (z.B. in Gl. 2.14) oder partielle $logP_{o/w}$ Parameter eingeführt, oder der entropische Rotations-Term wurde unter dem Aspekt modifiziert, dass nicht alle Seitenketten eines Liganden zwingend an das Protein binden und die Anzahl möglicher Konformationen nicht proportional zur Anzahl der drehbaren Bindungen ist. Durch die bessere Verfügbarkeit von hochaufgelösten Protein-Strukturen und experimentell bestimmten Bindungs-Affinitäten wurden zudem die Parametrisierung-Datensätze umfangreicher

Jahr	Autor	Name	Ref.
1994	Böhm	SCORE1	[16]
1995	Gelhaar	PLP	[49]
1996	Head	VALIDATE	[62]
1996	Jain	F (Hammerhead)	[71]
1996	Rarey	F-Score	[120]
1997	Eldridge	ChemScore	[38]
1998	Böhm	SCORE2	[17]
1998	Morris	AutoDock	[100]
1998	Takamatsu		[144]
1998	Wang	Score	[154]
1999	Rognan	Fresno	[126]
2000	Stahl	F-Score (mod.)	[141]
2001	Kellogg	HINT	[76]
2002	Wang	X-CSCORE	[153]
2004	Friesner	GlideScore 2.5	[46]
2004	Giordanetto		[53]
2005	Krammer	LigScore	[82]
2006	Friesner	GlideScoreXP	[47]
2006	Pham	Surflex	[117]
2007	Reulecke	HYDE	[121]

Tabelle 2.1.: Übersicht empirische Scoring-Funktionen.

und es wurden neuere Regressionsmethoden wie *partial least squares* (PLS) [48], genetische Algorithmen [56] oder neuronale Netzwerke verwendet. Ein generelles Problem bei empirischen Scoring-Funktionen ist die Berücksichtigung destabilisierender Beiträge. Da ungünstige Konstellationen, welche zu einer Bestrafung führen würden, in Kristall-Strukturen von Protein-Ligand Komplexen unterrepräsentiert sind, erhalten entsprechende Scoring-Terme bei einer Regressions-Analyse ein zu geringes Gewicht. Pham versucht dieses Problem durch die Verwendung von künstlich generierten falschen Lösungen und einer entsprechenden Klassifikation als Nicht-Binder zu umgehen [117]. Solche Lösungen weichen allerdings beliebig weit von der Realität ab und sind hinsichtlich ihrer Affinität nicht quantifizierbar. Eine Parametrisierung führt zu keinem besseren Verständnis der auftretenden Kräfte und Energien.

Wissensbasierte Scoring-Funktionen

Neben den empirischen Scoring-Funktionen sind die sogenannten wissensbasierten Scoring-Funktionen sehr populär. Nachdem sie ursprünglich für die Struktur-Vorhersage von Proteinen eingesetzt wurden [138], erfreuen sie sich aufgrund ihrer einfachen Form und generellen Anwendbarkeit auch im strukturbasierten Wirkstoffentwurf großer Beliebtheit. Unter der Annahme, dass eine kristallographisch aufgeklärte Struktur eines Protein-Ligand Komplexes das energetische Optimum repräsentiert, lassen sich durch die Anwendung der Boltzmann-Statistik aus Abstands-Verteilungen paarweise Protein-Ligand Kontakt-Potenziale für unterschiedliche Atom-Typen bestimmen. Die PMF Funktion (*potential of mean force*) von Muegge [102] als typischer Repräsentant dieser Art von Scoring-Funktion hat die Form

$$A_{ij}(r) = -k_B T \ln \left[f^j_{Vol\_corr}(r) \frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}} \right] \tag{2.15}$$

wobei das Potenzial A_{ij} zwischen Rezepetor Atom i und Ligand Atom j als Funktion des Abstandes r zwischen diesen berechnet wird. ρ^{ij}_{seg} bezeichnet die Anzahl der Atom-Paare ij im Parametrisierungs-Datensatz, die innerhalb eines bestimmten Raum-Segmentes gefunden werden. Diese 'Dichte' steht im Bezug zur Anzahl der entsprechenden Kontakt-Paare im gesamten Raum ρ^{ij}_{bulk}. Der Faktor $f^j_{Vol\_corr}$ wurde eingeführt, um das Volumen zu berücksichtigen, das durch den Liganden selbst besetzt ist und wo keine Wechselwirkungen stattfinden können. k_B ist die Boltzmann Konstante und T die absolute Temperatur. Der PMF Score eines Protein-Ligand Komplexes berechnet sich als Summe über alle Protein-Ligand Atom-Paare innerhalb eines bestimmten Abstandes.

Andere wissensbasierte Scoring-Funktionen (siehe Tabelle 2.2) sind in ihrer Form der PMF Funktion sehr ähnlich und unterscheiden sich hauptsächlich in der Größe und Zusammensetzung des Trainingsdatensatzes. Gohlke hat in der DrugScore [55] Funktion einen zusätzlichen Term für der 'hydrophoben Kollaps' eingeführt. Muryshev [103] erhält die Potenzial-Funktionen aus Paar-Korrelations Studien, passt aber zusätzliche Parameter durch eine Regressionsanalyse bezüglich gemessener Bindungs-Affinitäten an.

Ein viel diskutiertes Problem bei wissensbasierten Scoring-Funktionen ist die Auswahl eines geeigneten Bezugs-Systems (entsprechend dem ρ^{ij}_{bulk} Term in der PMF Funktion). Die Boltzmann-Statistik basiert auf einer Häufigkeits-Verteilung, welche die jeweiligen Energie-Zustände repräsentiert. Um reversible Faltungs-Vorgänge zu ermöglichen, sind in Protein-Strukturen jedoch auch häufig energetisch ungünstige Konstellationen zu finden, wie zum Beispiel ungesättigte Wasserstoff-Brücken-Funktionen [130]. Huang umgeht

Jahr	Autor	Name	Ref.
1996	DeWitte	SMog96	[31]
1999	Mitchell	BLEEP	[98]
1999	Muegge	PMF	[102]
2000	Gohlke	DrugScore	[55]
2001	Ishchenko	SmoG2001	[70]
2003	Feher	BHB	[41]
2003	Muryshev		[103]
2004	Ozrin	PLASS	[106]
2005	Zhang	DFIRE	[162]
2006	Huang	ITScore	[69]
2006	Muegge	PMF04	[101]
2006	Yang	M-Score	[161]

Tabelle 2.2.: Übersicht wissensbasierte Scoring-Funktionen.

dieses Problem bei der Kalibrierung von ITScore [69] durch Verwendung eines iterativen Verfahrens, bei dem die Potenzial-Terme solange verbessert werden, bis die Funktion zwischen experimentell bestimmten Ligand-Konformationen und 'unnatürlichen' Komplexen (*decoys*) unterscheiden kann.

2.4.2. Consensus Scoring

Während für eine Validierung von Scoring-Funktionen im strukturbasierten Wirkstoffdesign noch Mitte der Neunziger Jahre nur sehr wenige Daten zur Verfügung standen, sind mittlerweile viele Datenbanken zugänglich, in denen Struktur-Daten von Protein-Ligand Komplexen mit experimentellen Werten verknüpft sind. Die bekanntesten von ihnen sind die Ligand Protein DataBase [125], CLiBE [23], Relibase [63], KiBank [163] und PDBbind [150, 151]. Es ist von großem Interesse, dass die verwendeten Scoring Funktionen eine gute Übereinstimmung mit experimentellen Werten zeigen. Aus diesem Grund werden in regelmäßigen Abständen Studien publiziert, welche verschiedene Scoring-Funktionen miteinander vergleichen, zuletzt von Bissantz [13], Pearlman [110], Peréz [112], Wang [156, 155], Ferrara [42], Marsden [96] und Perola [113]. Interessanterweise existiert keine Scoring-Funktion, die in allen Studien gleich gut abschneidet. Je nach Art der bewerteten Protein-Ligand Komplexe liefern unterschiedliche Funktionen die besten Resultate.

Wenn bei einem neuen Projekt kein explizites Wissen über die Anwendbarkeit von bestimmten Scoring-Funktionen vorliegt, kann es von Vorteil sein, verschiedene Scoring-

Funktionen zu verwenden und die Ergebnisse zu vereinen. Dieses sogenannte *Consensus Scoring* liefert für unterschiedliche Komplexe häufig verlässlichere Resultate als eine einzige Scoring-Funktion, da die Gefahr von Ausreißern geringer ist. Je nach der Art und Weise, wie man die Ergebnisse der unterschiedlichen Funktionen miteinander kombiniert, wird zwischen Rang-, Score- oder Regressions-basierten Methoden unterschieden [22, 25, 10, 104]. Es ist auch möglich einzelne Terme aus verschiedenen Scoring-Funktionen neu miteinander zu kombinieren. ScreenScore [142] basiert auf einer Analyse von vier unterschiedlichen Scoring-Funktionen und beinhaltet Terme der empirischen Funktionen PLP [49] und F-Score [120].

Man muss sich jedoch darüber im Klaren sein, dass durch Verwendung eines Consensus-Schemas nicht nur die Stärken sondern auch die Schwächen der zugrunde liegenden Funktionen in das Ergebnis eingehen. Eine Consensus-Funktion wird quantitativ niemals das Ergebnis erzielen, das mit einer optimalen oder für ein spezielles Protein optimierten Funktion möglich wäre. Deshalb empfiehlt sich eine solche Vorgehensweise vor allem als initialer Schritt in einem neuen Projekt. Wenn erste experimentelle Ergebnissen und bessere Kenntnisse über das Zielproteins vorliegen, sollten spezifischere Methoden wie zum Beispiel POEM [8] eingesetzt werden. In POEM wird der *design of experiment* (DOE) Ansatz [99] mit verschiedenen Regressionstechniken kombiniert, um eine empirische Scoring-Funktion für ein bestimmtes Protein zu optimieren.

2.4.3. Berücksichtigung von Desolvatations-Effekten

Ein großes Problem bei der Bewertung von Protein-Ligand Komplexen ist der Einfluss des Lösungsmittels auf die Bindungs-Affinität. Protein und Ligand liegen nicht isoliert vor. Sie sind zu jedem Zeitpunkt von Wasser-Molekülen umgeben. Bei der Bildung des Komplexes wird das Wasser aus der Bindetasche verdrängt, Wasserstoff-Brückenbindungen werden gebrochen und die Moleküle arrangieren sich neu um den Komplex herum. Während das Aufbrechen von Wasserstoff-Brücken ein endothermer Prozess ist ($\Delta H > 0$) und damit primär destabilisierend beiträgt, kann die Neuordnung der Wasser-Moleküle einen günstigen Einfluss auf die Bindungs-Affinität haben. Offensichtlich wird dieser hydrophobe Effekt durch das Aggregieren unpolarer Oberflächen. Dieses Phänomen wird häufig durch van der Waals Kräfte zwischen den Oberflächen erklärt, was jedoch nicht korrekt ist. Vielmehr ist der hydrophobe Effekt auf entropische und enthalpische Änderungen im Wasserstoff-Brücken-Netzwerk zurückzuführen. In einem anderen Medium als Wasser wären abweichende Phänomene zu beobachten. Im folgenden Kontext wird die Desolvatation mit der Dehydratation gleichgesetzt, da Wasser das natürliche Lösungsmittel im zellulären Organismus darstellt.

In der klassischen Böhm-Funktion [16] wird der hydrophobe Effekt durch die Änderung der zugänglichen hydrophoben Oberfläche berücksichtigt. Die Desolvatisierung polarer funktioneller Gruppen fließt indirekt in den Ausdruck für Wasserstoff-Brücken mit ein. Der Grundbeitrag einer Wasserstoff-Brücke zwischen Protein und Ligand ist gegenüber der Vakuum-Energie einer solchen Wechselwirkung stark vermindert. Unberücksichtigt bleibt allerdings die Dehydratisierung von polaren Funktionen, welche keine neuen Wechselwirkungen ausbilden. Dass diese einen Einfluss auf die Bindungs-Affinität und die Stabilität eines Komplexes haben, konnte z.B. durch Untersuchungen bei Pfizer gezeigt werden [53]. Bei der statistischen Analyse verschiedener Deskriptoren von Protein-Ligand Komplexen wurde die Änderung in der polaren zugänglichen Oberfläche als eine Hauptkomponente bezüglich des Einflusses auf die Bindungs-Affinität ermittelt.

Verschiedene neuere Ansätze für die Bewertung von Protein-Ligand Komplexen beinhalten Terme zur Beschreibung von Desolvatations-Effekten:

HiNT. Im HiNT Kraftfeld [76] werden Wechselwirkungs-Beiträge von Atomen auf Basis von partiellen logP-Werten und Abständen zwischen den Atomen berechnet. Je nach Art eines Kontaktes werden die logP-Werte mit unterschiedlichem Vorzeichen berücksichtigt. Kontakte zwischen hydrophoben Atomen und Säure-Base-Paare (Wasserstoff-Brücken) werden belohnt, polare Kontakte zwischen zwei Säure- oder Base-Atomen werden bestraft.

X-CSCORE. Auch in der empirischen Scoring-Funktion X-CSCORE [153] werden logP-Inkremente verwendet. Hier dienen sie allerdings nur zur Beschreibung des hydrophoben Effekts. Wird ein hydrophobes Ligand-Atom in einer hydrophoben Tasche des Proteins platziert, so trägt es stabilisierend zur Bindungs-Affinität bei. Die Entscheidung, ob ein Atom zum hydrophoben Effekt beiträgt ist abhängig von dem entsprechenden Vorzeichen des logP-Inkrements und denen der benachbarten Atome. Die übrigen Terme von X-CSCORE berücksichtigen van der Waals Kräfte, Wasserstoff-Brücken und Deformations-Effekte. Eine Parametrisierung aller Terme erfolgt anhand von experimentellen Bindungs-Affinitäten. In alternativen Formen wird der hydrophobe Effekt auch über paarweise Kontakte von hydrophoben Atomen oder eine Änderung in der zugänglichen hydrophoben Oberfläche beschrieben.

LigScore. Die LigScore Funktion [82] besteht aus drei Termen. Neben den üblichen Ausdrücken für van der Waals und polare Wechselwirkungen wird explizit die Desolvatisierung von polaren Ligand- und Protein-Atomen infolge der Komplexbildung bestraft.

GlideXP. GlideXP [47] berücksichtigt Einflüsse des Lösungsmittels durch verschiedene Terme. Die von ChemScore [38] abgeleitete Scoring-Funktion verwendet eine Gitter-basierte Methode zur Berechnung der Desolvatations-Energie von polaren Atomen. Zu-

sätzlich werden strukturelle Motive berücksichtigt, welche eine Bindung zwischen Protein und Ligand begünstigen. Dazu zählen der Einschluss von hydrophoben Ligand-Atomen zwischen hydrophoben Regionen des Proteins (*hydrophobic enclosure*), der vergrößerte Energie-Beitrag von Wasserstoff-Brücken in hydrophoben Bindetaschen und unterschiedliche Einflüsse von Ladungen auf die Stärke von Wasserstoff-Brückenbindungen.

2.5. Intermolekulare Wechselwirkungen in wässriger Lösung

Als theoretische Grundlage für die Entwicklung der HYDE Scoring-Funktion dient der Ansatz von Lange *et al.* [87] zur Beschreibung von intermolekularen Wechselwirkungen in wässriger Lösung. Gemäß dieser Theorie lässt sich die freie Gibbsche Bindungs-Energie zwischen kleinen organischen Molekülen mit einem einheitlichen Konzept beschreiben, welches sowohl Wasserstoff-Brückenbindungen als auch stabilisierende und destabilisierende Desolvatationseffekte berücksichtigt. Wasser lässt sich durch eine Temperaturabhängige Größe f_{sat} charakterisieren, die den Anteil der abgesättigten Wasserstoff-Brücken-Funktionen im Lösungsmittel-Netzwerk angibt. Im Eiskristall ist jedes Wasser-Molekül in tetraedrischer Geometrie an vier benachbarte Wasser-Moleküle gebunden. In flüssiger Phase jedoch weichen die Abstände zwischen den Atomen vom Idealwert ab, Wasserstoff-Brückenbindungen zwischen den einzelnen Molekülen werden ständig gebrochen und neu geknüpft. Mit zunehmender Temperatur nimmt der Anteil an Wasserstoff-Brücken mit ungünstiger Geometrie zu, und der Anteil an ungesättigten Wasserstoff-Brücken-Funktionen in der flüssigen Phase $f_{unsat} = 1 - f_{sat}$ lässt sich unter der Annahme abschätzen, dass (1) in der festen Phase jedes Wasser-Molekül vier Wasserstoff-Brücken mit idealer Geometrie macht, (2) in der Gasphase alle Bindungen gebrochen sind und die Moleküle isoliert vorliegen, (3) die Energie zum Aufbrechen aller Bindungen durch die Summe von Schmelz-, Verdampfungsenthalpie und der Enthalpie für das Erhitzen von Wasser von 273 K auf 373 K berechnet werden kann und (4) der Anteil gebrochener Bindungen proportional zu der in dem System enthaltenen Energie ist.

$$f_{unsat} = \frac{\Delta H_{fusion} + c_p \cdot (T - 273K)}{\Delta H_{fusion} + c_p \cdot (373K - 273K) + \Delta H_{evaporation}} \tag{2.16}$$

ΔH_{fusion} ist die Schmelzenthalpie, $\Delta H_{evaporation}$ die Verdampfungsenthalpie, c_p die molare Wärmekapazität und T die absolute Temperatur. f_{unsat} hat Werte zwischen 0.11 nahe dem Gefrierpunkt und 0.25 am Siedepunkt. Unter physiologischen Bedingungen beträgt der Anteil ungesättigter Wasserstoff-Brücken-Funktionen etwa 17%.

Die einzelnen Wasser-Moleküle im Lösungsmittel sind nicht voneinander unterscheidbar und können nur unter statistischen Gesichtspunkten betrachtet werden. Damit verringert sich die mittlere Wechselwirkungs-Energie zwischen zwei Molekülen $\varepsilon^{H_2O\cdots H_2O}$ im Vergleich zu der Wechselwirkungs-Energie in der Gasphase $\varepsilon_0^{H_2O\cdots H_2O}$ entsprechend dem Faktor f_{sat}.

$$\varepsilon^{H_2O\cdots H_2O} = f_{sat} \cdot \varepsilon_0^{H_2O\cdots H_2O} \tag{2.17}$$

Es konnte experimentell gezeigt werden, dass sich die Struktur des Wassers um kleine polare Moleküle nicht wesentlich ändert, wenn diese Wasserstoff-Brückenbindungen zum Lösungsmittel mit einer Energie $\varepsilon_0^{polar\cdots H_2O} \approx \varepsilon_0^{H_2O\cdots H_2O}$ ausbilden [139]. Solche Moleküle integrieren sich in das Lösungsmittel-Netzwerk und damit reduziert sich die Wechselwirkungs-Energie einer solchen polaren Gruppe mit Wasser.

$$\varepsilon^{polar\cdots H_2O} = f_{sat} \cdot \varepsilon_0^{polar\cdots H_2O} \tag{2.18}$$

Bei der Dehydratisierung einer polaren Gruppe wird ein Wasser-Molekül wieder freigegeben und dieses stellt eine Wasserstoff-Brücken-Funktion für die Bildung einer Wechselwirkung mit einem zweiten Wasser-Molekül zur Verfügung. Damit lässt sich die freie Gibbsche Energie für die Dehydratisierung einer polaren Funktion abschätzen:

$$\begin{aligned} \Delta G_{dehydration} &= -f_{sat} \cdot \varepsilon_0^{polar\cdots H_2O} + \frac{1}{2} f_{sat} \cdot \varepsilon_0^{H_2O\cdots H_2O} \\ &\approx -\frac{1}{2} f_{sat} \cdot \varepsilon_0^{polar\cdots H_2O} \end{aligned} \tag{2.19}$$

Auf ähnliche Weise lässt sich auch der Beitrag einer apolaren funktionellen Gruppe (z.B. einer Methylgruppe) zur Dehydratisierungs-Energie in Abhängigkeit von der Größe f_{sat} bzw. der Temperatur herleiten. Durch die Dehydratisierung einer apolaren Gruppe wird ein ungesättigtes Wasser-Molekül wieder in das Lösungsmittel-Netzwerk integriert, dies ist mit einem Energie-Gewinn verbunden. Der hydrophobe Effekt wird damit als Eigenschaft des Wasserstoff-Brücken-Netzwerks im Lösungsmittel erklärt.

Anhand dieser Überlegungen lässt sich die freie Bindungs-Energie $\Delta G_{binding}$ in einem Protein-Ligand Komplex in erster Näherung aus nur zwei Größen ableiten, (a) den Beiträgen für die Desolvatisierung und (b) der Wasserstoff-Brücken-Energie $\varepsilon^{polar1\cdots polar2}$

der polaren Gruppen im Protein-Ligand Interface. Unter der Voraussetzung, dass die zu bewertende Struktur keine ungünstigen sterischen Überlappungen aufweist, lässt sich die freie Gibbsche Energie der Bindung von zwei Molekülen A und B berechnen als

$$\Delta G_{binding} = \Delta G_{dehydration} + \sum_{j=1}^{n} \sum_{k=1}^{m} \varepsilon^{j \cdots k} \cdot \delta_{H-bond}(j,k) \tag{2.20}$$

$$\delta_{H-bond}(j,k) = \begin{cases} 1 & : \text{H-bond zwischen } j \text{ und } k \\ 0 & : \text{ansonsten} \end{cases}$$

Während die Dehydratisierung-Energie $\Delta G_{dehydration}$ sowohl die günstigen Beiträge durch den hydrophoben Effekt als auch die ungünstigen durch die Desolvatisierung der polaren Gruppen von A und B beinhaltet, tragen die Wasserstoff-Brücken zwischen den Atomen j von Molekül A und k von Molekül B immer stabilisierend zur Bindungs-Energie bei. In Abhängigkeit von der Geometrie zwischen j und k kann der Beitrag im Vergleich zur Vakuum-Wechselwirkungs-Energie vermindert sein. Im Extremfall trägt eine Wasserstoff-Brücke übehaupt nicht zur Bindungs-Energie bei.

Anstatt die Wasserstoff-Brücken-Energie einer diskreten Wechselwirkung wie in Gleichung 2.21 als Ganzes zu berücksichtigen, lässt sich dieser Beitrag auch anteilig auf die beiden Atome j und k aufteilen, zwischen denen die Wechselwirkung stattfindet. Damit lässt sich die Bindungs-Energie als Summe über Beiträge von allen Atome i im Komplex beschreiben.

$$\Delta G_{binding} = \sum_{i=1}^{n+m} \Delta G^{i}_{dehydration} + \Delta G^{i}_{H-bond} \tag{2.21}$$

$\Delta G^{i}_{dehydration}$ ist die freie Energie der Dehydratisierung von Atom i und ΔG^{i}_{H-bond} dessen Beitrag durch neu gebildete Wasserstoff-Brücken.

3. HYDE

3.1. Allgemeine Form

Die Beschreibung der Bindungs-Energie von Molekülen in wässriger Lösung in Abschnitt 2.5 beruht auf vereinfachenden Annahmen. Die Einteilung in das Schema polar/unpolar ist für viele Atome nicht anwendbar. Aufgrund unterschiedlicher elektrischer Momente und durch die Polarisierbarkeit der Atome ist der Übergang zwischen einem polaren Charakter (Einbindung in das Wasserstoff-Brücken-Netzwerk) und einem unpolaren (Störung des Netzwerks) fließend. Die HYDE Scoring-Funktion beruht auf einem empirischen Ansatz, welcher die Dehydratations-Energie einzelner Atome auf Basis von inkrementellen Löslichkeits-Parametern beschreibt. In Einklang mit der beschriebenen Theorie werden durch diesen Ansatz sowohl die destabilisierenden Beiträge von desolvatisierten Wasserstoff-Brücken-Funktionen als auch der hydrophobe Effekt in konsistenter Weise modelliert. Der Energie-Gewinn durch Wasserstoff-Brückenbindungen im Protein-Ligand Interface wird ebenfalls durch die Löslichkeits-Parameter der beteiligten Atome beschrieben. Dies geschieht unter Berücksichtigung des Faktors f_{sat}, der den Anteil gesättigter Wasserstoff-Brücken-Funktionen im Wasser beschreibt.

Die freie Bindungs-Energie $\Delta G_{binding}$ eines Protein-Ligand Komplexes wird mit HYDE als Summe über die Beiträge aller Rezeptor- und Ligand-Atome i zur Wasserstoff-Brücken- und zur Desolvatations-Energie berechnet. Wasserstoff-Atome werden implizit als Teil der entsprechenden Schweratome behandelt und leisten keinen eigenen Energie-Beitrag. Ihre genaue Ausrichtung ist aber wichtig für die Berechnung richtungsabhängiger Größen. Wie in Abbildung 3.1 schematisch dargestellt, werden für die Desolvatisierung und für die Berechnung der Beiträge von Wasserstoff-Brücken der Komplex und der unassoziierte Zustand unabhängig voneinander betrachtet und Unterschiede im Grad der Solvatisierung und in der Anzahl der diskreten Wechselwirkungen berechnet. Die allgemeine Formel für die freie Bindungs-Energie lautet

$$\Delta G_{binding} = \sum_i \Delta G^i_{dehydration} + \Delta G^i_{interaction} \tag{3.1}$$

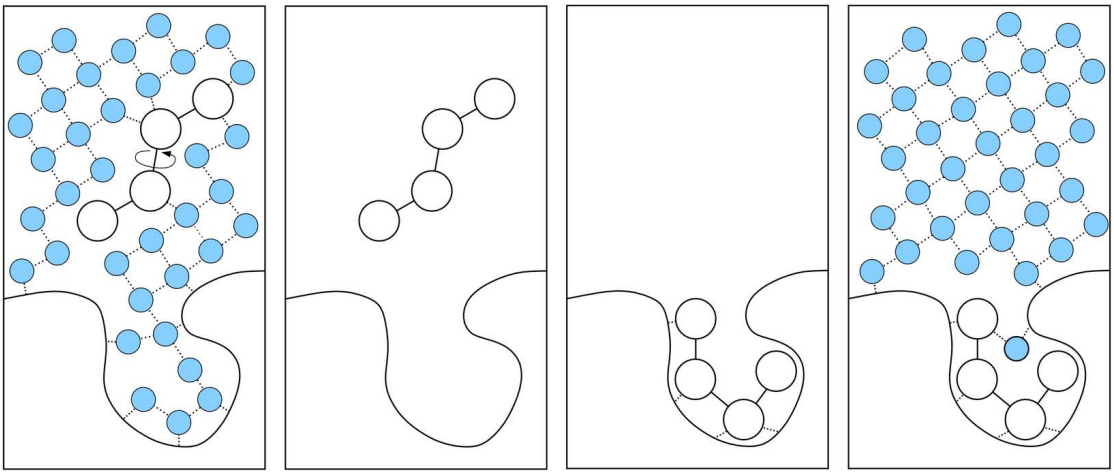


Abbildung 3.1.: Der Assoziations-Prozess von Protein und Ligand wird durch HYDE aufgeteilt. Desolvatation und Interaktionen werden unabhängig voneinander bewertet.

Dabei entspricht $\Delta G^i_{dehydration}$ dem Energie-Beitrag für die Änderung in der Lösungsmittel-Zugänglichkeit eines Atoms infolge des Übergangs aus dem unassoziierten Zustand in das Protein-Ligand Interface. $\Delta G^i_{interaction}$ ist die Energie-Änderung durch neue bzw. gebrochene Wasserstoff-Brückenbindungen. Nicht berücksichtigt in dieser Formel sind Entropie-Änderungen durch den Wegfall von Freiheitsgraden der Translation und Rotation der beteiligten Moleküle. Diese Beiträge werden in einer Scoring Funktion üblicherweise durch einen konstanten Term berücksichtigt, welcher für deren Anwendbarkeit allerdings keine Relevanz besitzt. Weiterhin setzt HYDE voraus, dass die zu bewertenden Strukturen eine spannungsfreie Konformation und keine sterischen Überlappungen aufweisen. HYDE wurde unter dem Schwerpunkt entwickelt, durch Docking-Programme generierte Protein-Ligand Komplexe erneut zu bewerten (*Rescoring*). Diese sollten eine entsprechende Qualität bezüglich der genannten Parameter aufweisen. Unrealistische Konformationen, auch solche welche am Rand des definierten aktiven Zentrums liegen, sollten nach Möglichkeit durch entsprechende Filter-Terme aussortiert werden.

3.2. Desolvatation

In HYDE werden zur Beschreibung der Desolvatation partielle logP Atom-Inkremente ($plogP$) verwendet. Der $logP_{o/w}$-Wert einer chemischen Verbindung ist der negative dekadische Logarithmus des Verteilungskoeffizienten $K_{o/w}$ der Verbindung im Zwei-Phasen-System n-Octanol/Wasser. Je größer der logP-Wert einer Verbindung ist, desto hydro-

phober ist diese und desto größer ist der Anteil an Molekülen, die in der Octanol-Phase gelöst sind verglichen mit denen im Wasser. Der plogP Wert eines Atoms sei dessen inkrementeller Beitrag zum logP des gesamten Moleküls.

Unter der Annahme, dass die freie Energie der Desolvatisierung in Octanol verglichen mit der in Wasser sehr gering ist, kann der logP-Wert als Maß für die freie Energie der Dehydratisierung verwendet werden. Analog dem in Abschnitt 2.3.2 beschriebenem Zusammenhang zwischen der Gleichgewichtskonstante K und der freien Energie ΔG im chemischen Gleichgewicht gilt

$$\Delta G_{dehydration} = -RT \ln K_{o/w} = -RT \frac{\log K_{o/w}}{\log e} = -2.3RT \cdot logP_{o/w} \tag{3.2}$$

R ist die Gaskontante und T die absolute Temperatur. $logP_{o/w}$ ist der Verteilungskoeffizient. Entsprechend wird mit HYDE der Beitrag eines Atoms i zur freien Energie der Dehydratisierung $\Delta G^i_{dehydration}$ auf Basis seines partiellen plogP-Wertes $plogP_i$ berechnet.

$$\Delta G^i_{dehydration} = -2.3RT \cdot \frac{acc^i_{free} - acc^i_{complex}}{acc^i_{mean}} \cdot plogP_i \tag{3.3}$$

Neben dem plogP-Wert des Atoms hängt $\Delta G^i_{dehydration}$ auch von einem Ausdruck für die Änderung in der Lösungsmittel-Zugänglichkeit acc^i des Atoms bei der Bildung des Protein-Ligand Komplexes ab. $acc^i_{complex}$ bezeichnet die Zugänglichkeit des Atoms unter Berücksichtigung aller Nachbar-Atome im Protein-Ligand Komplex. acc^i_{free} ist die Zugänglichkeit im unassoziierten Zustand. acc^i_{mean} ist ein mittlerer Wert für die statistische Zugänglichkeit eines Atom-Typs. Eine genauere Beschreibung dieser Größen ist in Abschnitt 4.1 zu finden.

Für ein Atom, dessen Zugänglichkeit im unassoziierten Zustand seiner statistischen Zugänglichkeit acc^i_{mean} entspricht, und welches bei der Bildung des Komplexes vollkommen vom Lösungsmittel abgeschirmt wird, ergibt sich nach dieser Formel ein Beitrag von $-2.3 \cdot RT \cdot plogP_i$ zur Dehydratisierungenergie. Hydrophobe Atome ($plogP_i > 0$) tragen demnach entsprechend dem hydrophoben Effekt mit einem negativen Beitrag stabilisierend zur Gesamtenergie bei, während polare Atome für den Verlust von Wechselwirkungen zum umgebenden Wasser bestraft werden.

Die Berechnung der Zugänglichkeit acc basiert auf dem SAS-Modell von Richards [122].

Dieses Modell wurde weiterentwickelt, um der bevorzugten Ausrichtung von Lösungsmittel-Molekülen in Nachbarschaft von polaren Atomen gerecht zu werden, und wird in Abschnitt 4.2.1 genauer erläutert.

Während die Zugänglichkeit im Protein-Ligand Komplex $acc^i_{complex}$ anhand der vorliegenden Konformation des Liganden in der Bindetasche des Proteins eindeutig berechenbar ist, stellt sich die Frage, welche Konformation für die Berechnung von acc^i_{free} zugrunde gelegt werden sollte. Der Ligand und in geringerem Maße auch das Protein werden je nach Beschaffenheit im unassoziierten Zustand eine andere Konformation einnehmen als im Komplex. Einhergehend damit ändert sich auch die Zugänglichkeit einzelner Atome zum Lösungsmittel. Es muss also ein Referenz-Zustand definiert werden, über den die Änderung in der Zugänglichkeit infolge aller mit der Desolvatisierung verbundenen Prozesse berechnet werden kann.

Genaugenommen müsste man eine Konformations-Analyse für jedes Molekül durchführen und darüber die mittlere Zugänglichkeit der Atome im Lösungsmittel bestimmen. Dies kann aber nicht die Aufgabe einer Scoring-Funktion sein, die für die Bewertung einer gegebenen Konformation angewendet wird, da solch eine Rechnung sehr viel Zeit in Anspruch nehmen würde. Eine einfachere Möglichkeit wäre die Verwendung der Energieminimierten Eingabe-Struktur des Docking-Programms zur Berechnung von acc^i_{free}. Allerdings müsste man bei dieser Vorgehensweise auch explizit einen Term für die Änderung der Konformations-Energie zwischen dem ungebundenen Zustand und dem Komplex verwenden. Stattdessen wird für die Berechnung der Zugänglichkeit im freien Zustand die gleiche Konformation des Liganden verwendet, wie sie im Komplex zu finden ist. Auf Seite des Rezeptors geht man üblicherweise ohnehin von einer starren Konformation aus.

Eine weitere Näherung ist die Verwendung der Atom-Typen und damit derselben Parameter für $plogP_i$ und acc^i_{mean} sowohl für Atome der Liganden als auch beim Protein. Diese Werte stammen aus einer Regressions-Analyse von kleinen, Inhibitor-ähnlichen Molekülen mit experimentell bestimmten logP-Werten. Für Makromoleküle wie Proteine, bei denen viele Atome im Inneren verborgen sind und deren Eigenschaften vorwiegend durch die Tertiärstruktur bestimmt werden, ist die Berechnung eines globalen logP-Wertes anhand von Atom-Inkrementen nicht möglich. Die verwendeten Atom-Typen sind jedoch in begrenztem Maße als lokale Charakteristika auch in einem Protein anwendbar. Sie klassifizieren die Zugänglichkeits-abhängigen Wechselwirkungen mit dem Lösungmittel in unmittelbarer Umgebung. Im Einklang mit experimentellen Beobachtungen [81] tragen in HYDE sehr exponierte Regionen an der Protein-Oberfläche stärker zur Desolvatation bei als solche, die in einer engen, unzugänglichen Bindetasche liegen.

3.3. Wasserstoff-Brücken

In der FlexX-Implementierung [120] der Böhm Funktion [16] erhält jede Wasserstoff-Brücke denselben Energie-Beitrag zugeschrieben, wenn die Abweichungen von der idealen Wechselwirkungs-Geometrie gering sind. Dieses Schema in Verbindung mit dem im vorherigen Abschnitt hergeleiteten Dehydratations-Term, welcher unterschiedlichen Atom-Typen individuelle Strafterme zuordnet, würde zu einem systematischen Fehler bei der Berechnung der Bindungs-Energie führen. Wasserstoff-Brücken-Akzpetoren und -Donoren mit einem großen plogP-Wert und einer entsprechend großen Bestrafung als Folge der Desolvatisierung würden durch eine Ausbildung von Wasserstoff-Brücken niemals den gleichen Energie-Gewinn liefern können wie solche mit kleinen plogP-Werten und dadurch bei der Bewertung immer benachteiligt werden.

Das in HYDE verwendete Wechselwirkungs-Schema beruht auf dem in Kapitel 2.5 hergeleiteten Zusammenhang zwischen der Wechselwirkungs-Energie einer polaren Gruppe mit einem Lösungsmittel-Molekül und der idealen Wechselwirkungs-Energie mit Wasser in der Gasphase. Abweichend von der FlexX Scoring-Funktion werden explizit nur Wasserstoff-Brücken modelliert und keine Wechselwirkungen zwischen aromatischen Systemen. Die Wechselwirkungs-Energie wird zu gleichen Teilen auf beide an der Wasserstoff-Brücke beteiligten Atome aufgeteilt, der Beitrag eines Atoms i ist

$$\Delta G^i_{interaction} = \frac{1}{2} \cdot \varepsilon_0^{polar \cdots H_2O} \tag{3.4}$$

Einsetzen von Gleichung 2.19 ergibt

$$\Delta G^i_{interaction} = -\frac{1}{f_{sat}} \cdot \Delta G^i_{dehydration} \tag{3.5}$$

und aus Gleichung 3.3 folgt unter der Annahme der vollständigen Desolvatisierung des Atoms

$$\Delta G^i_{interaction} = \frac{2.3 \cdot RT}{f_{sat}} \cdot plogP_i \tag{3.6}$$

Damit ist der Energie-Beitrag fuer eine diskrete Wechselwirkung eines Atoms um den Faktor $1/f_{sat} \approx 1.2$ höher als der entsprechende Energie-Verlust durch die Dehydratisierung. Der Austausch einer Wasserstoff-Brückenbindung zu einem Lösungmittel-Molekül

durch eine Wechselwirkung im Protein-Ligand Interface (oder durch eine intramolekulare Wechselwirkung) trägt im Einklang mit der Theorie insgesamt stabilisierend zur Bindungs-Energie bei.

Die in HYDE angewendete Formel für die Berechnung des Wechselwirkungs-Beitrages beinhaltet einen zusätzlichen Ausdruck sat_i für die *Sättigung* eines Atoms:

$$\Delta G^i_{interaction} = \frac{2.3 \cdot RT}{f_{sat}} \cdot \left(sat^i_{complex} - sat^i_{free}\right) \cdot plogP_i \tag{3.7}$$

$sat^i_{complex}$ ist die Sättigung des Atoms im Protein-Ligand Komplex und sat^i_{free} die Sättigung im unkomplexierten Zustand. Die Sättigung eines Atoms ergibt sich aus der Anzahl und Güte der gemachten Wasserstoff-Brücken mit den Atomen j in der Umgebung von i und wird berechnet als

$$sat^i = \sqrt{\sum_j f_{dev}(i,j)} \tag{3.8}$$

Dabei ist $f_{dev}(i,j)$ der Bestrafungsfaktor für Abweichungen in der Geometrie entsprechend dem FlexX-Wechselwirkungsmodell (s. Kapitel 4.3.1). $f_{dev}(i,j)$ ist 1 für eine gute Wechselwirkung und nimmt linear auf 0 ab, wenn die Abweichungen einen Toleranzwert überschreiten.

Der Wurzel-Ausdruck über die Summe der Bestrafungsfaktoren in Gleichung 3.8 wurde eingeführt, um Atome, die multiple Wechselwirkungen ausbilden können, nicht übermäßig zu belohnen. Der plogP-Beitrag eines Atoms repräsentiert dessen vollständige Desolvatisierung bei dem Übergang in die hydrophobe Octanol-Phase, und er ist unabhängig von der konkreten Anzahl an Wassertoff-Brücken zum Lösungsmittel. Demzufolge wäre es falsch, diesen Wert als Maß für die Stärke jeder einzelnen Wechselwirkung anzusetzen. Zum Beispiel würde eine Hydroxyl-Gruppe, welche drei Wasserstoff-Brücken in der Bindetasche macht, ein plogP-Äquivalent als Beitrag der Desolvatation erhalten. Diese Strafe würde durch 3 entsprechende Beiträge infolge der gemachten Wechselwirkungen bei weitem überkompensiert und die Hydroxyl-Gruppe wäre durch den potenziell sehr großen Energie-Gewinn gegenüber anderen funktionellen Gruppen unverhältnismäßig bevorzugt.

In Einklang mit der Beschreibung der Desolvatation sollte der Wechselwirkungs-Beitrag

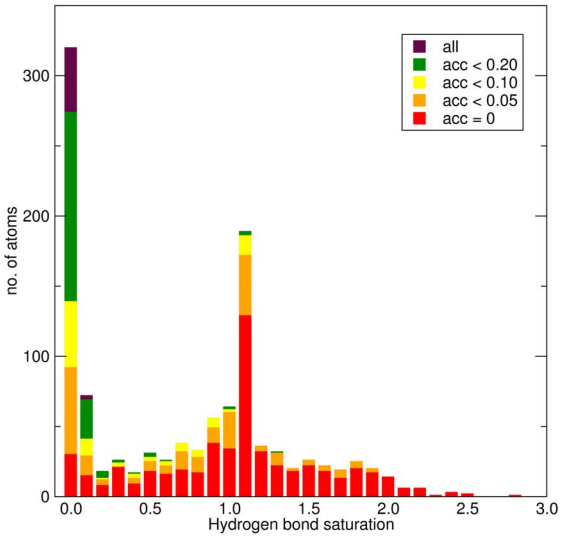


Abbildung 3.2.: Berechnete Anzahl an Wechselwirkungen als Summe der Bestrafungsfaktoren f^{ij}_{dev} in Abhängigkeit der berechneten Zugänglichkeit. Untersucht wurden 1123 Hydroxyl-Gruppen der Liganden in 800 Komplexen der PDBbind (Version 2002) [150].

eines Atoms sich auf einen Referenz-Zustand sat^{i}_{mean} beziehen, welcher angibt, wieviele Wasserstoff-Brücken dieses Atom im Normalfall machen kann. Idealerweise wäre dies die Summe von Wasserstoff-Brücken-Donoren und -Akzeptoren, also Wasserstoff-Atomen und freien Elektronen-Paaren. Allerdings hat eine Untersuchung von Protein-Ligand Komplexen aus dem PDBbind Datensatz [150] gezeigt, dass diese ideale Anzahl an Wechselwirkungen nur selten erreicht wird. In Abbildung 3.2 ist als Beispiel ein Histogramm zu sehen, welches die Anzahl an Wasserstoff-Brücken der 1123 Ligand-Hydroxyl-Gruppen im Datensatz darstellt. Diese Zahlen wurden als Summe über die Bestrafungsfaktoren f^{ij}_{dev} in FlexX und in Abhängigkeit von der berechneten Zugänglichkeit $acc^{i}_{complex}$ ermittelt. Der theoretische Wert von drei Wasserstoff-Brücken wird unter anderem aus sterischen Gründen nur selten erreicht. Die Verwendung eines Referenz-Zustandes analog der Zugänglichkeit acc^{i}_{mean} in Gleichung 3.3 würde also zu einer systematischen Unterbewertung von manchen Atomen führen, die ihre ideale Anzahl an Wechselwirkungen nicht erreichen können.

Durch Verwendung des Wurzel-Terms kann jedes Atom den Energie-Verlust durch die Dehydratisierung mit einer ausgebildeten Wasserstoff-Brücke mit guter Geometrie ausgleichen. Weitere Wasserstoff-Brücken bringen einen zusätzlichen Energie-Gewinn, der allerdings deutlich abgeschwächt ist und damit nicht zu einer Überbewertung führt.

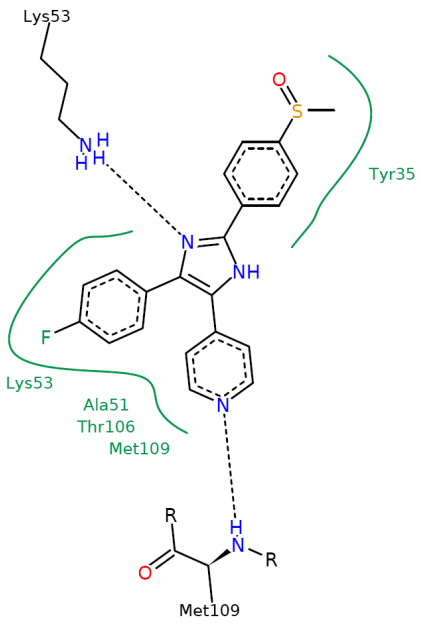


Abbildung 3.3: Poseview [143] Darstellung der p38 MAP Kinase im Komplex mit dem Inhibitor SB203580 (PDB 1a9u [157]).

3.4. Bewertung eines Protein-Ligand Komplexes

Als Beispiel für die Funktionsweise von HYDE sei die Bewertung eines Komplexes der p38 MAP Kinase mit dem Inhibitor 4-(4-Fluorophenyl)-2-(4-methylsulfinylphenyl)-5-(4-pyridyl)-imidazol angeführt (PDB 1a9u [157]). Dieser Inhibitor bindet in charakteristischer Weise in die ATP-Bindetasche der Kinase und bildet zwei essentielle Wasserstoff-Brücken zu diesem aus. Der Pyridin-Ring überlappt mit der Adenin-Position des ATPs und bindet in der sogenannten Hinge-Region an den Amid-Stickstoff von Methionin 109 (Abbildung 3.3). Die zweite Wechselwirkung findet zwischen dem zentralen Imidazol-Ring und Lysin 53 statt. Diese Amino-Säure wird üblicherweise als *Gatekeeper* bezeichnet, da sie durch konformationelle Änderungen die Bindetasche öffnen und verschließen kann. Der Fluorophenyl-Ring füllt die hydrophobe Bindetasche aus und leistet damit einen essentiellen Beitrag zur Bindungs-Affinität.

Die Bewertung mit HYDE ergibt eine Bindungs-Energie von -42.4 kJ mol^{-1}, das entspricht einer Bindungs-Affinität von $K_D = 36.2$ nM. In der Aufschlüsselung der Beiträge zum Gesamt-Score (Tabelle 3.1) wird deutlich, dass dem Energie-Gewinn durch die Wasserstoff-Brücken eine äquivalente Bestrafung für die Desolvatisierung von hydrophilen Atomen gegenübersteht. Der Energie-Unterschied zum ungebundenen Zustand ergibt sich demnach hauptsächlich aus dem hydrophoben Effekt.

```
Receptor            : 1a9u
Ligand              : 1a9u_crysth [ Fix ]
Score contributions
  dG_saturation    : -25.719 kJ/mol
  dG_desolvation   : -16.693 kJ/mol
  -----------------------------------
  ligand hydrophobic   : -24.328
  ligand hydrophilic   :  10.868
  receptor hydrophobic : -13.803
  receptor hydrophilic :  10.570
  -----------------------------------
  dG_total         : -42.412 kJ/mol
```

Tabelle 3.1.: HYDE Score-Beiträge für die Bewertung von 1a9u.

Die Ausgabe für die Bewertung einzelner Atome ist in Beiträge durch Wechselwirkungen und Beiträge durch den Desolvatations-Term aufgeteilt. Die Wechselwirkungs-Beiträge von Ligand- und Rezeptor-Atomen werden unabhängig voneinander berechnet (Tabelle 3.2). Die Atome `NB1` (Pyridin) und `NC3` (Imidazol) werden für die ausgebildeten Wasserstoff-Brücken zu `N MET 109` und `NZ LYS 53` entsprechend ihrem plogP-Wert belohnt. Der Beitrag für `NZ LYS 53` ist etwas geringer, da dieses Rezeptor-Atom bereits eine intramolekulare Wechselwirkung ausbildet und die zweite Wasserstoff-Brücke durch Verwendung des Wurzel-Terms (Gl. 3.8) einen kleineren Beitrag zur Wechselwirkungs-Energie liefert.

Die Desolvatations-Energie ist durch die Beiträge aller Atome gegeben, bei denen sich infolge der Komplex-Bildung die Zugänglichkeit verändert. In der Übersicht sind nur diejenigen Rezeptor-Atome aufgeführt, die mehr als 0.1 kJ mol^{-1} zur Gesamt-Energie beitragen (Tabellen 3.3 und 3.4). Die Beiträge der Ligand-Atome sind von den absoluten Werten her größer als die der Rezeptor-Atome. Dafür sind diese zahlreicher, da der Ligand in der Bindetasche von Rezeptor-Atomen umgeben ist. Je nach Vorzeichen des plogP-Wertes tragen die Atome entweder stabilisierend oder destabilisierend zur Bindungs-Affinität bei. Vor allem die an den Wasserstoff-Brücken beteiligten Atome werden stark für die Desolvatisierung bestraft.

Zum einfacheren Verständnis des berechneten Scores ist es möglich, die Energie-Beiträge der Atome in einer Farb-Codierung darzustellen (Abbildung 3.4). Dazu werden die einzelnen Beiträge für jedes Ligand-Atom aufsummiert. Die Beiträge von Rezeptor-Atomen werden dem geometrisch nächstgelegenen Ligand-Atom zugeordnet. Grün eingefärbte Atome bewirken eine negative Änderung der Bindungs-Energie (stabilisierend), rot eingefärbte Atome sind destabilisierend, entweder durch den eigenen Beitrag oder durch die

```
 >> Hydrogen bond saturation energy
-----------------------------------------------------------------
Scoring parameter: 1.20 (Cm) * 5.70 (Cd) = 6.84 (Cs)
Energy term: E_sat = Cs * plogP * sat_f
             sat_f = sqrt(sat_c) - sqrt(sat_s)
-----------------------------------------------------------------

Ligand atom match contribution
-----------------------------------------------------------------
atom name            | sat_c sat_s sat_m |  sat_f  plogP |  E_sat
  1 O2               | 0.000 0.000   -   |  0.000 -0.800 | -0.000
 10 NB1              | 0.903 0.000   -   |  0.950 -1.050 | -6.824
 16 NC1              | 0.000 0.000   -   |  0.000 -0.550 | -0.000
 18 NC3              | 0.920 0.000   -   |  0.959 -1.050 | -6.889
-----------------------------------------------------------------
Ligand saturation match energy: -13.713

Receptor atom match contribution
-----------------------------------------------------------------
atom name            | sat_c sat_s sat_m |  sat_f  plogP |  E_sat
 NZ  LYS     53      | 1.351 0.431   -   |  0.506 -1.310 | -4.532
 N   MET    109      | 0.903 0.000   -   |  0.950 -1.150 | -7.474
-----------------------------------------------------------------
Receptor saturation match energy: -12.006
```

Tabelle 3.2.: Bewertung von 1a9u, Wasserstoff-Brücken-Energie.

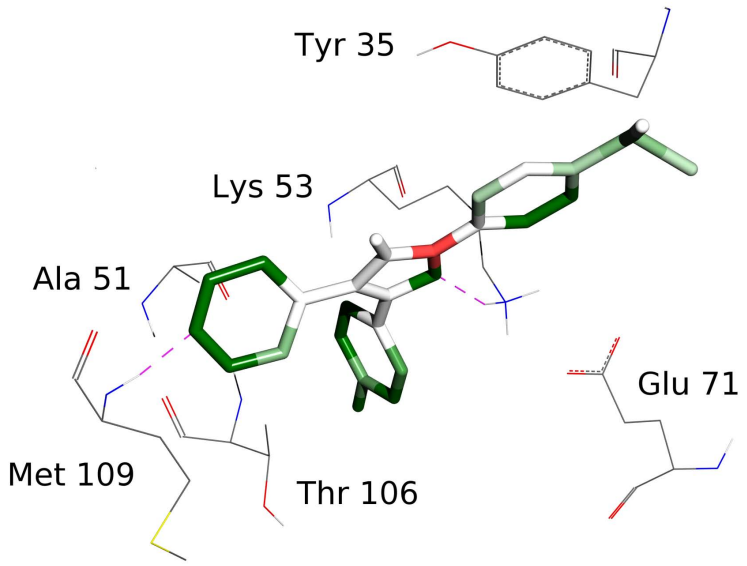


Abbildung 3.4.: Farbcodierte Darstellung von 1a9u. Ligand-Atome mit einem negativen Beitrag zu $\Delta G^i_{binding}$ werden in grün dargestellt, Atome mit einem positiven (destabilisierendem) Beitrag in rot. Der Beitrag der Rezeptor-Atome zu $\Delta G^i_{binding}$ ist in den Beitrag der Ligand-Atome einberechnet.

Desolvatisierung von Rezeptor-Atomen.

Der Beitrag von Wasserstoff-Brücken zur Bindungs-Energie liegt in der gleichen Größenordnung wie die durch die FlexX Scoring-Funktion berechnete Energie. Die Wechselwirkung zwischen `NB1` und `N MET 109` trägt insgesamt mit -4.28 kJ mol^{-1} zur Gesamt-Energie bei, die Wasserstoff-Brücke zwischen `NC3` und `NZ LYS 53` mit -5.72 kJ mol^{-1}. Im Gegensatz zur FlexX Scoring-Funktion machen diese Beiträge aber nur einen kleinen Teil der Gesamt-Energie aus, da sie durch Strafterme der gleichen Größenordnung nahezu kompensiert werden.

Besonders deutlich wird der Unterschied zu herkömmlichen Scoring-Funktionen, wenn man einzelne Atome austauscht. Durch Mutation des Pyridin-Ringes in einen Phenyl-Ring (Austausch von `NB1` durch ein Kohlenstoff-Atom `CB1`) verringert sich die berechnete Bindungs-Energie um +5.1 auf -37.3 kJ mol^{-1}, dies entspricht einer Änderung um eine Größenordnung in der Dissoziationskonstante ($K_D = 285.9$ nM). Durch den Wegfall der Wasserstoff-Brücke kann die Desolvatation des Methionin-Stickstoffs nicht mehr kom-

```
>> plogP desolvation term
-----------------------------------------------------------------------
Parameter set    : Ghose/Crippen (atom-based)
Ligand reference : pose coordinates
Site reference   : fix coordinates
Access mode      : weighted solvent accessible surface (WSAS)
Scoring parameter: Cd = 5.70

dG = Cd * plogp(occ) * (acc(c) - acc(s)) / acc(m)
-----------------------------------------------------------------------
```

Ligand atoms

atom	acc(c)	acc(s)	acc(m)	f(acc)	plogp	dG_desolv
1 O2	0.378	0.378	0.170	0.000	-0.800	-0.000
2 C1	0.476	0.563	0.540	-0.161	0.740	-0.677
3 S1	0.021	0.120	0.000	0.000	0.000	0.000
4 CA1	0.138	0.222	0.260	-0.322	0.430	-0.788
5 CA2	0.172	0.228	0.260	-0.216	0.430	-0.530
6 CA3	0.005	0.022	0.020	-0.859	0.000	0.000
7 CA4	0.071	0.172	0.260	-0.388	0.430	-0.950
8 CA5	0.045	0.234	0.260	-0.727	0.430	-1.782
9 CA6	0.000	0.038	0.020	-1.914	0.000	0.000
10 NB1	0.000	0.099	0.080	-1.234	-1.050	7.385
11 CB2	0.091	0.370	0.260	-1.073	0.430	-2.629
12 CB3	0.104	0.271	0.260	-0.643	0.430	-1.576
13 CB4	0.011	0.011	0.020	0.000	0.000	0.000
14 CB5	0.042	0.149	0.260	-0.412	0.430	-1.009
15 CB6	0.000	0.358	0.260	-1.376	0.430	-3.373
16 NC1	0.037	0.037	0.090	-0.000	-0.550	0.000
17 CC2	0.000	0.057	0.020	-2.852	0.000	0.000
18 NC3	0.000	0.047	0.080	-0.582	-1.050	3.482
19 CC4	0.000	0.005	0.020	-0.234	0.000	0.000
20 CC5	0.000	0.017	0.020	-0.859	0.000	0.000
21 CD1	0.063	0.212	0.260	-0.574	0.430	-1.407
22 CD2	0.000	0.277	0.260	-1.064	0.430	-2.607
23 CD3	0.000	0.051	0.020	-2.539	0.000	0.000
24 CD4	0.000	0.291	0.260	-1.118	0.430	-2.740
25 CD5	0.000	0.215	0.260	-0.826	0.430	-2.025
26 CD6	0.000	0.000	0.020	0.000	0.000	0.000
27 FD3	0.005	0.475	0.360	-1.306	0.300	-2.234

Tabelle 3.3.: Bewertung von 1a9u, Desolvatation Ligand.

Receptor atoms

atom	acc(c)	acc(s)	acc(m)	f(acc)	plogp	dG_desolv
CG1 VAL 30	0.070	0.113	0.540	-0.081	0.740	-0.342
CB TYR 35	0.041	0.184	0.240	-0.596	0.480	-1.630
CD1 TYR 35	0.020	0.132	0.260	-0.430	0.430	-1.053
CD2 TYR 35	0.000	0.029	0.260	-0.111	0.430	-0.272
CE1 TYR 35	0.054	0.132	0.260	-0.300	0.430	-0.736
CE2 TYR 35	0.000	0.021	0.260	-0.081	0.430	-0.199
OH TYR 35	0.055	0.075	0.210	-0.096	-1.130	0.620
CG1 VAL 38	0.000	0.070	0.540	-0.130	0.740	-0.549
CG2 VAL 38	0.000	0.137	0.540	-0.253	0.740	-1.068
CB ALA 51	0.000	0.116	0.540	-0.214	0.740	-0.903
CB LYS 53	0.000	0.059	0.240	-0.244	0.480	-0.668
CD LYS 53	0.000	0.048	0.240	-0.202	0.480	-0.552
CE LYS 53	0.000	0.039	0.240	-0.163	0.480	-0.445
NZ LYS 53	0.020	0.111	0.210	-0.431	-1.310	3.218
CD2 LEU 75	0.000	0.084	0.540	-0.155	0.740	-0.653
CD1 ILE 84	0.045	0.170	0.540	-0.231	0.740	-0.976
CB LEU 104	0.000	0.046	0.240	-0.192	0.480	-0.525
CB THR 106	0.000	0.030	0.080	-0.371	0.120	-0.254
CG2 THR 106	0.000	0.157	0.540	-0.291	0.740	-1.227
CD1 LEU 108	0.033	0.088	0.540	-0.103	0.740	-0.433
N MET 109	0.000	0.016	0.040	-0.393	-1.150	2.577
O MET 109	0.066	0.122	0.170	-0.333	-0.800	1.517
CB MET 109	0.079	0.128	0.240	-0.205	0.480	-0.561
CG MET 109	0.000	0.059	0.240	-0.244	0.480	-0.668
CG LEU 167	0.015	0.025	0.080	-0.127	0.120	-0.087
OD1 ASP 168	0.149	0.278	0.270	-0.477	-0.970	2.638

Tabelle 3.4.: Bewertung von 1a9u, Desolvatation Protein.

```
>> Ligand atom energy contributions
 atom   |   sat   :   rec   |  plogp  :   rec   |  total
---------+---------+---------+---------+---------+---------
  1   O2 |   -0.00 |    0.00 |   -0.00 |    0.00 |    0.00
  2   C1 |    0.00 |    0.00 |   -0.68 |    0.00 |   -0.68
  3   S1 |    0.00 |    0.00 |    0.00 |   -1.05 |   -1.05
  4  CA1 |    0.00 |    0.00 |   -0.79 |    0.00 |   -0.79
  5  CA2 |    0.00 |    0.00 |   -0.53 |    0.62 |    0.09
  6  CA3 |    0.00 |    0.00 |    0.00 |   -0.74 |   -0.74
  7  CA4 |    0.00 |    0.00 |   -0.95 |   -1.90 |   -2.85
  8  CA5 |    0.00 |    0.00 |   -1.78 |   -0.20 |   -1.98
  9  CA6 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00
 10  NB1 |   -6.82 |   -7.47 |    7.39 |    2.63 |   -4.28 <
 11  CB2 |    0.00 |    0.00 |   -2.63 |   -0.78 |   -3.40
 12  CB3 |    0.00 |    0.00 |   -1.58 |   -0.55 |   -2.13
 13  CB4 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00
 14  CB5 |    0.00 |    0.00 |   -1.01 |   -0.09 |   -1.10
 15  CB6 |    0.00 |    0.00 |   -3.37 |   -0.67 |   -4.04
 16  NC1 |   -0.00 |    0.00 |    0.00 |    0.00 |    0.00
 17  CC2 |    0.00 |    0.00 |    0.00 |    1.57 |    1.57
 18  NC3 |   -6.89 |   -4.53 |    3.48 |    2.22 |   -5.72 <
 19  CC4 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00
 20  CC5 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00
 21  CD1 |    0.00 |    0.00 |   -1.41 |    0.00 |   -1.41
 22  CD2 |    0.00 |    0.00 |   -2.61 |   -1.63 |   -4.24
 23  CD3 |    0.00 |    0.00 |    0.00 |   -1.23 |   -1.23
 24  CD4 |    0.00 |    0.00 |   -2.74 |    0.00 |   -2.74
 25  CD5 |    0.00 |    0.00 |   -2.03 |   -0.67 |   -2.69
 26  CD6 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00
 27  FD3 |    0.00 |    0.00 |   -2.23 |   -0.78 |   -3.01
---------+---------+---------+---------+---------+---------
   total |  -13.71 |  -12.01 |  -13.46 |   -3.23 |  -42.41
```

Tabelle 3.5.: Bewertung von 1a9u, Beiträge einzelner Ligand-Atome.

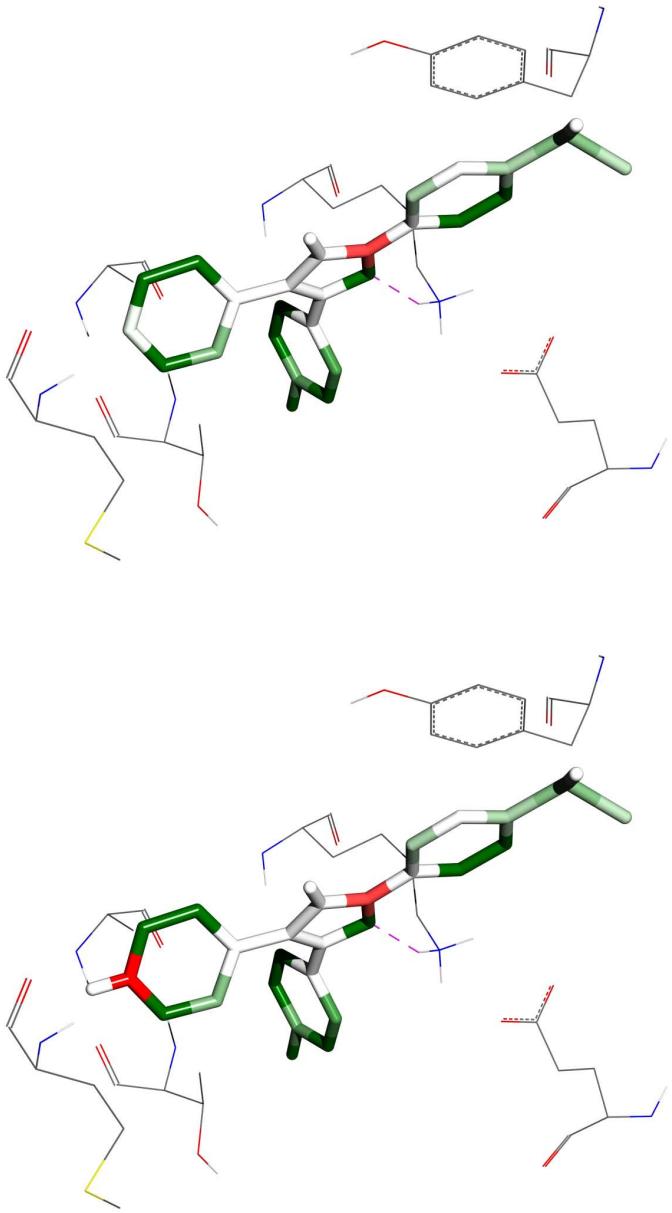


Abbildung 3.5.: Farbcodierte Darstellungen von 1a9u. Der Pyridin-Stickstoff des Liganden wurde durch [CH] (oben) bzw. [NH+] ausgetauscht. Während der Einbau eines Kohlenstoff-Atoms zu einem neutralen Energie-Beitrag führt, bewirkt die Protonierung des Stickstoffs einen destabilisierenden Beitrag.

pensiert werden. Allerdings wird die Desolvatation des eingefügten Kohlenstoff-Atoms belohnt, so dass sich ein Gesamtbeitrag dieser Konstellation von -0.12 kJ mol^{-1} ergibt. Entsprechend ist `CB1` in Abbildung 3.5 neutral (weiß) eingefärbt.

Eine noch deutlichere Auswirkung auf die Bindungs-Affinität ergibt sich durch das Zusammenbringen von zwei polaren Funktionen, die aufgrund von falscher Protonierung oder ungünstiger Geometrie keine Wechselwirkung ausbilden können. Eine Protonierung von `NB1` führt zu einer Bindungs-Energie von -32.6 kJ mol^{-1}, der Gesamt-Beitrag von `NB1` ist +5.51 kJ mol^{-1}. Der Energie-Unterschied zum Original-Liganden von +9.8 kJ mol^{-1} durch eine nicht zustande gekommene Wasserstoff-Brücke bewirkt eine Änderung der Dissoziationskonstante um zwei Größenordnungen auf $K_D = 1.9\ \mu$M. Dies erklärt, warum schon der Austausch eines einzigen Atomes den Unterschied ausmachen kann, ob ein Ligand an ein Protein bindet oder nicht. Ein Umstand, der nur durch die Berücksichtigung destabilisierender Beiträge quantitativ zu erklären ist.

4. Atom-Deskriptoren

4.1. Desolvatation

4.1.1. Atombasierte logP Inkrement-Verfahren

Die Approximation des Octanol/Wasser Verteilungskoeffizienten logP eines organischen Moleküls ist möglich unter Verwendung eines atombasierten, inkrementellen Verfahrens. Basierend auf Untersuchungen von Broto [19] zu atombasierten Struktur-Eigenschafts-Beziehungen, haben Ghose und Crippen bereits 1986 eine Methode veröffentlicht, bei der der berechnete logP sich über die Häufigkeit ergibt, mit der bestimmte Atom-Typen in einem Molekül auftreten [51]. Die unterschiedlichen Atom-Typen i tragen jeweils mit einem konstanten Wert a_i zum geschätzten logP des Moleküls bei.

$$logP = \sum_i n_i a_i. \tag{4.1}$$

n_i ist die Anzahl der Atome vom Typ i in dem betrachteten Molekül und a_i sind die individuellen Atom-Beiträge. Die Atom-Beiträge wurden durch multiple lineare Regression an einem Datensatz aus 494 Molekülen mit experimentell bestimmten Verteilungskoeffizienten ermittelt. Ghose und Crippen unterscheiden dabei zwischen 110 unterschiedlichen Atom-Typen. In Anlehnung an diese Methode sind zahlreiche Veröffentlichungen erschienen, in denen Verteilungskoeffizienten basierend auf Atom-Inkrementen mit einer guten Übereinstimmung mit experimentellen Daten vorhergesagt werden, unter anderem von Wang *et al.* [152] sowie *Hou et al.* [67]. Diese Methoden unterscheiden sich in der Definition der Atom-Typen, der Größe und der Zusammensetzung des Parametrisierungs-Datensatzes und in der bei der Parametrisierung eingesetzten Regressions-Methode. Hou *et al.* verwenden in ihrem Ansatz zudem die Lösungsmittel-zugängliche Oberfläche der Atome als einen zusätzlichen Deskriptor.

Bereits Ghose hat festgestellt, dass ein grundlegendes Problem bei solch einer Methode

die starke lineare Abhängigkeit zwischen den verschiedenen Atom-Typen ist [51]. Bedingt durch die sehr invarianten chemischen Umgebungen einzelner Atom-Typen lässt sich der genau Beitrag eines Atoms zum logP Wert des Moleküls häufig nicht exakt bestimmen. Verschiedene Atom-Typen treten in einem organischen Molekül überwiegend paarweise auf, zum Beispiel die Ghose/Crippen-Typen `17:=CR2` und `58:=O`, welche den Kohlenstoff und den Sauerstoff einer Carbonyl-Gruppe repräsentieren. Eine Regressions-Methode kann den Gesamt-Anteil der Carbonyl-Gruppe am logP des Moleküls bestimmen, die genaue Zuordnung zu einzelnen Atome ist jedoch nicht möglich und der Fehler in den berechneten Einzel-Werten ist relativ groß.

Dennoch ergeben atombasierte Inkrement-Methoden bei der Vorhersage von logP-Werten sehr gute Ergebnisse. Stark miteinander korrelierende Atom-Typen treten auch in den zu bewertenden Strukturen meist paarweise auf und die Summe der Einzel-Inkremente ergibt unabhängig von deren genauer Zuordnung den logP-Beitrag der entsprechenden funktionellen Gruppe.

4.1.2. Partielle logP-Werte als Maß der Desolvatations-Energie

In Abschnitt 3.2 wurde gezeigt, wie die Desolvatations-Energie eines Atoms in HYDE durch die Verwendung von partiellen (p)logP-Werten bestimmt wird. Da der plogP die individuelle Hydrophobizität eines Atoms beschreiben soll, müssen die verwendeten Parameter möglichst genau bestimmt werden. Eine Korrelation zwischen unterschiedlichen Atom-Typen würde dazu führen, dass die Desolvatations-Energie falsch berechnet wird. Neben dem plogP-Wert wird bei der Berechnung der Desolvatations-Energie (Gleichung 3.3) die Zugänglichkeit eines Atoms berücksichtigt, da die einzelnen Atome nicht immer komplett desolvatisiert werden bzw. im ungebundenen Zustand teilweise desolvatisiert vorliegen. Von daher ist es notwendig, schon bei der Bestimmung der plogP-Parameter einen zusätzlichen Term für die Lösungmittel-Zugänglickeit zu verwenden. Dies erscheint zudem sinnvoll, da sich verglichen mit dem einfachen Ansatz von Ghose und Crippen durch eine entsprechende Vorgehensweise bessere Ergebnisse bei der Berechnung von logP Werten erzielen lassen [67, 66].

Durch die Berücksichtigung der Lösungsmittel-zugänglichen Oberfläche kann in der Regressions-Analyse zwischen paarweise auftretenden Atom-Typen unterschieden werden. Diese haben je nach chemischer Umgebung eine unterschiedliche Zugänglichkeit und tragen unterschiedlich zum logP-Wert des Moleküls bei. Die bei der Parametrisierung von HYDE für ein Molekül mit n Atomen j vom Typ k verwendete Inkrementformel lautet:

$$logP = \sum_{j=1}^{n} \frac{acc^j}{acc^k_{mean}} \cdot plogP_k(j). \tag{4.2}$$

acc^j ist die Zugänglichkeit eines Atoms j. acc^k_{mean} ist die mittlere Zugänglichkeit des Atom-Typs k im Parametrisierungs-Datensatz und $plogP_k$ dessen logP Inkrement. Die Berechnung der Zugänglichkeit wird in Abschnitt 4.2 genauer beschrieben.

Neben der Berücksichtigung der zugänglichen Oberfläche wurde bei der Entwicklung von HYDE auch Wert darauf gelegt, die Atom-Typen so zu definieren, dass möglichst keine der oben geschilderten paarweisen Abhängigkeiten auftreten. Dies ist vor allem zu beobachten, wenn die Atom-Typen auf Basis der Zugehörigkeit zu funktionellen Gruppen definiert werden. Daher wurden nur die unmittelbaren Einflüsse auf die elektrostatischen Eigenschaften eines Atoms berücksichtigt: (a) die Ordnungszahl, (b) die Hybridisierung, (c) die Anzahl und Art der ausgehenden Bindungen und (d) die Anzahl gebundener Wasserstoff-Atome. Durch diese räumlich sehr begrenzte Differenzierung lässt sich eine Korrelation zwischen benachbarten Atom-Typen bei der Parametrisierung im wesentlichen vermeiden, und die Anzahl der unterschiedlichen Atom-Typen ist gering genug, dass es nicht zu einer Überbestimmung des Systems kommt.

4.1.3. Parametrisierung durch multiple lineare Regression

Für die Parametrisierung wurden aus der PHYSPROP Datenbank [1] ca. 700 Moleküle mit experimentell ermittelten logP-Werten ausgewählt. Die PHYSPROP Datenbank enthält Strukturen, Namen und physikalische Eigenschaften von über 25000 Chemikalien, die aus verschiedenen Quellen zusammengetragen wurden. Die ausgewählten Strukturen stammen aus der Zusammenstellung von Hansch und Leo [59] und nehmen einen logP-Wertebereich von -1.5 bis 9.4 ein. Neben dem hydrophoben Kohlenstoff-Gerüst enthalten sie jeweils nur eine funktionelle Gruppen bzw. mehrere funktionelle Gruppen vom selben Typ. Dies trägt zusätzlich zu einer Verringerung des Fehlers bei der Parametrisierung der logP-Inkremente bei.

Mit Corina [128] (Version 3.10) wurden für die im SMILES Format vorliegenden Moleküle 3D-Koordinaten berechnet und diese über das für diesen Zweck implementierte PLOGP Menü in FlexX eingelesen. Für die multiple lineare Regression wurde eine Matrix erstellt, in der die Zeilen die Moleküle repräsentieren und die Spalten die Atom-Typen. Die Felder der Matrix entsprechen den berechneten Zugänglichkeiten der Atome acc^j, die experimentellen logP-Werte sind die abhängigen Variablen der Regression. Aus

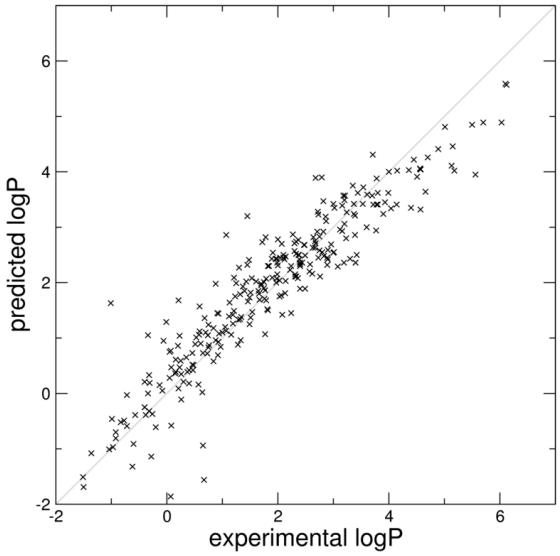


Abbildung 4.1: Experimentelle und vorhergesagte logP-Werte des Parametrisierungs-Datensatzes (Korrelations-Koeffizient: 0.93). Die Parametrisierung erfolgte in diesem Beispiel unter Verwendung des WSAS Algorithmus zur Bestimmung der Zugänglichkeit.

den Regressions-Koeffizienten ergeben sich die plogP-Parameter. Die Verwendung des PLOGP Menüs ist im Anhang A.2 geschildert.

Ein Korrelations-Koeffizient von 0.93 zwischen den experimentellen logP-Werten und den mit den Ergebnissen der Regressions-Analyse berechneten bestätigt die Anwendbarkeit des atombasierten Inkrement-Verfahrens (Abbildung 4.1) und der multiplen linearen Regression. Allerdings war das Ziel der Analyse nicht die Vorhersage von logP-Werten sondern eine Verwendung der Atom-Parameter zur Berechnung der Desolvatisierungs-Energie. Die gute Korrelation zu den experimentellen logP-Werten deutet jedoch darauf hin, dass die ermittelten Parameter in einem sinnvollen Werte-Bereich liegen. Tatsächlich haben die plogP-Werte von Kohlenstoff-Atomen positive Vorzeichen und klassifizieren diese als hydrophob, während für Hetero-Atome, die Donor oder Akzeptor in Wasserstoff-Brücken sind, negative plogP-Werte ermittelt worden sind.

4.1.4. Berücksichtigung von geladenen Gruppen

Geladene Spezies wurden in der Regression nicht beachtet, obwohl in Biomolekülen häufig geladene Gruppen wie Carboxylate oder Ammonium-Ionen zu finden sind. Solche Gruppen haben keinen spezifischen logP Wert, sie können in verschiedenen protomeren Formen vorliegen. Stattdessen werden ihnen je nach Anzahl der möglichen protomeren Zuständen unterschiedliche logD-Werte zugewiesen, die abhängig vom pH-Wert berechnet werden. Da sich aber der pH-Wert gerade in Bindetaschen von Proteinen von dem pH-Wert im

umgebenden Medium unterscheiden kann, erscheint eine explizite Berücksichtigung des pH-Werts beim Scoring schwierig realisierbar und zudem mit großen Fehlern belastet. Aus diesem Grund wurden geladenen Atomen die gleichen plogP-Parameter wie den jeweiligen korrespondierenden Basen bzw. Säuren zugewiesen. Als Spezialfall wurde den Sauerstoff-Atomen einer Carboxylat-Gruppe `O_sp2_a_HO` der Mittelwert der plogP-Werte von Carbonyl- und Hydroxyl-Sauerstoff zugewiesen (`O_sp2_d_HO` bzw. `O_sp3_ss_H1`). Metall-Atomen wurde ein fixer Wert zugeordnet.

Eine Übersicht über die verwendeten Atom-Typen und die in der Regressions-Analyse für verschiedene Methoden der Zugänglichkeits-Berechnung bestimmten plogP-Parameter ist in Anhang B zu finden. Coulomb-Wechselwirkungen, also die Anziehung oder Abstoßung zwischen Ionen, werden in HYDE nicht berücksichtigt. Es ist davon auszugehen, dass ionische Gruppen auch im unkomplexierten Zustand von Atomen gegensätzlicher Ladung umgeben sind, welche ein wesentlicher Bestandteil des zellulären Mediums sind.

4.2. Lösungsmittel-Zugänglichkeit

Die Lösungsmittel-zugängliche Oberfläche SAS (*solvent accessible surface*) eines Moleküls wurde von Richards definiert als die Spur des Zentrums einer Kugel, welche über die van der Waals Oberfläche eines Moleküls gerollt wird [122]. Diese Kugel repräsentiert ein idealisiertes Lösungsmittel-Molekül. Für das Lösungsmittel Wasser wird üblicherweise ein Kugel-Radius von $r_{SAS} = 1.4$ Å angenommen. Bei Anwendungen im Bereich der strukturbasierten Wirkstoffentwicklung hat es sich als vorteilhaft erwiesen, einen etwas geringeren Radius zu verwenden, um Ungenauigkeiten in den zugrunde liegenden Strukturen gerecht zu werden. FlexX und HYDE verwenden einen Radius von $r_{SAS} = 1.2$ Å.

4.2.1. Gewichtete Lösungsmittel-zugängliche Oberfläche WSAS

HYDE berechnet die Änderung in der Lösungsmittel-Zugänglichkeit als ein direktes Maß für die freie Energie der Desolvatisierung als Folge der Komplex-Bildung. Während Zugänglichkeits-Änderungen in hydrophoben Bereichen eines Moleküls hauptsächlich die benachbarten Wasser-Moleküle beeinflussen, und die damit verbundene Änderung in der freien Energie linear mit der Änderung in der SAS korreliert [35], werden im Gegensatz dazu bei der Dehydratisierung von polaren Atomen diskrete Wasserstoff-Brücken aufgebrochen. Eine Änderung in der SAS hat demzufolge dort größere Auswirkungen auf die freie Energie, wo solche Wechselwirkungen tatsächlich auftreten können, die Stärke von Wasserstoff-Brücken ist aufgrund des Dipol-Charakters stark orientierungsab-

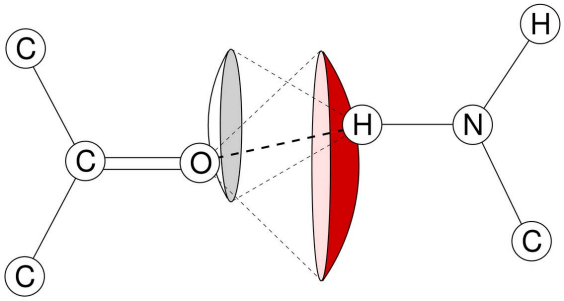


Abbildung 4.2: Schematische Darstellung einer Wasserstoff-Brücken-Wechselwirkung in FlexX. Die an der Wechselwirkung beteiligten Atome liegen auf der Wechselwirkungs-Fläche des jeweiligen Partners.

hängig. Daher wird in der HYDE Scoring-Funktion die SAS bezüglich des verwendeten Wechselwirkungs-Modells gewichtet. Die auf diese Weise berechnete Zugänglichkeit wird als *weighted solvent accessible surface* WSAS bezeichnet.

Bei der Berechnung der WSAS von hydrophilen Atomen ergeben nur solche Oberflächen-Abschnitte einen Beitrag, die in der Richtung einer FlexX-Wechselwirkungs-Fläche liegen [120]. Solche Flächen repräsentieren Regionen, in denen das an einer Wechselwirkung beteiligte Atom liegen darf, damit die Wechselwirkung den vollen Beitrag zur freien Bindungs-Energie leistet (Abbildung 4.2). Abweichungen im Winkel werden mit einem Faktor $f_{adev} < 1$ bestraft. Regionen, in denen keine Wechselwirkungen auftreten können, werden somit bei der Berechnung der WSAS nicht berücksichtigt. Hydrophobe Atome haben keine Vorzugsrichtung, hier entspricht die WSAS der Lösungsmittel-zugänglichen Oberfläche SAS.

4.2.2. Berechnung der WSAS

Die Berechnung der WSAS beruht auf der FlexX-Implementierung der Lösungsmittel-zugänglichen Oberfläche. Die SAS wird in FlexX über eine rekursive Tesselierungs-Funktion approximiert. Der van der Waals Radius r^i_{vdW} eines Atoms i wird um den Radius der Lösungsmittel-Kugel r_{SAS} erweitert und die Eckpunkte eines einbeschriebenen, regelmäßigen Ikosaeders werden auf Zugänglichkeit überprüft. Wenn die Entfernung zwischen einem Eckpunkt und der van der Waals Oberfläche eines benachbarten Atoms geringer als r_{SAS} ist, dann wird der Eckpunkt als unzugänglich gekennzeichnet, an diese Stelle kann kein Lösungsmittel-Molekül gelangen. Für jedes sphärische Dreieck j, das durch benachbarte Eckpunkte k des Ikosaeders gegeben ist, werden drei Fälle unterschieden:

1. Sind alle Eckpunkte k des sphärischen Dreiecks zugänglich, gilt auch der entsprechende Oberflächen-Abschnitt als zugänglich. Die Fläche des sphärischen Dreiecks wird zu der Gesamt-SAS des Atoms addiert.

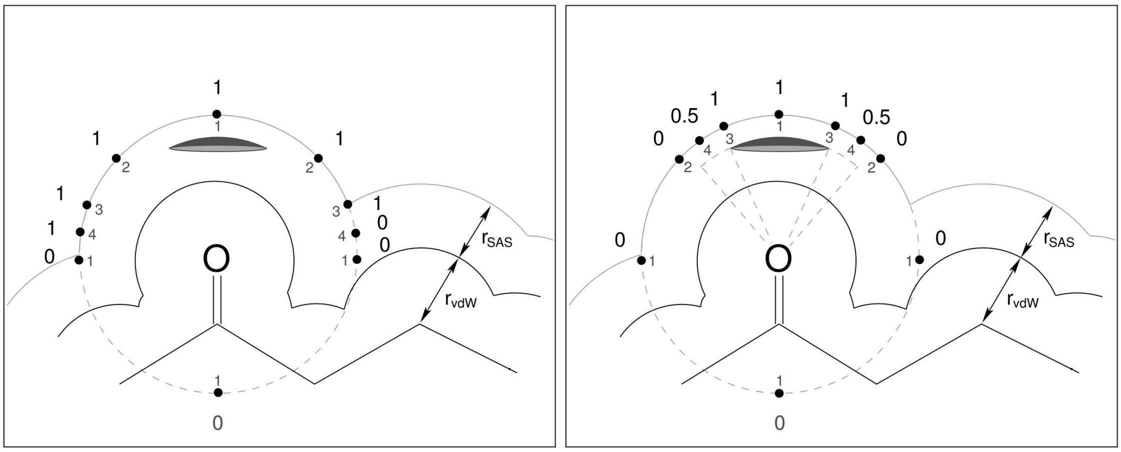


Abbildung 4.3.: Schematische Darstellung der WSAS Tesselierung. Die Zahlen innerhalb des gestrichelten Kreises geben die Rekursions-Tiefe an. Die Zahlen außerhalb des Kreises entsprechen den berechneten Zugänglichkeiten $a(i, j, k)$ der Eckpunkte auf der van der Waals Oberfläche eines Atoms. Im Gegensatz zum ursprünglichen SAS-Algorithmus (links) werden für die WSAS-Berechnung polarer Atome die Wechselwirkungs-Flächen berücksichtigt (rechts).

2. Sind alle Eckpunkte unzugänglich, ist der Oberflächen-Abschnitt ebenfalls nicht zugänglich und trägt nicht zur SAS bei.

3. Wenn mindestens ein Eckpunkt zugänglich und mindestens ein Eckpunkt unzugänglich ist, werden drei zusätzliche Punkte berechnet. Diese Punkte entsprechen den Mittelpunkten der Kanten des sphärischen Dreiecks und unterteilen dieses in vier kleinere regelmäsige Dreiecke. Die neuen Eckpunkte werden durch einen rekursiven Funktions-Aufruf ebenfalls auf Zugänglichkeit überprüft und der jeweilige Beitrag zur SAS ermittelt. Wenn die Abbruch-Bedingung der Rekursions-Schleife erreicht ist, werden die Beiträge der verbleibenden Dreiecke anhand der Anzahl der zugänglichen Eckpunkte abgeschätzt.

Analog zu der Berechnung der SAS erfolgt die Berechnung der gewichteten Oberfläche WSAS, allerdings mit dem Unterschied, dass die Entscheidung über die Zugänglichkeit eines Eckpunktes nicht länger eine ja/nein-Entscheidung ist. Zusätzlich wird die Winkel-Abweichung α des betrachteten Eckpunktes k von der Vorzugsrichtung der Wasserstoff-Brücken-Geometrien durch den Straf-Faktor $f_{adev}(i, k)$ in Betracht gezogen (Abbildung 4.3). Kann ein Atom mehrere Wasserstoff-Brücken ausbilden, so wird der größte Wert von $f_{adev}(i, k)$ verwendet. Hydrophobe Atome besitzen keine Vorzugrichtung für benachbarte Wasser-Moleküle, es gilt $f_{adev}(i, k) = 1$. Die gewichtete Zugänglichkeit $a(i, j, k)$ des Eckpunkte k eines Dreiecks j hängt vom Abstand $r_{k,l}$ zu allen umgebenden Atomen l ab

und wird berechnet als

$$a(i,j,k) = \begin{cases} f_{adev}(i,k) & : \forall\, l(r_{k,l} \geq r^l_{vdW} + r_{SAS}) \\ 0 & : \text{ansonsten} \end{cases} \tag{4.3}$$

Damit ergibt sich bei einer Rekursions-Tiefe n der Beitrag $wsas_{i,j}$ eines sphärischen Dreiecks j zur WSAS von Atom i als

$$wsas_{i,j} = \frac{4\pi \cdot \left(r_{SAS} + r^i_{vdW}\right)^2}{20 \cdot 4^n} \cdot \sum_{k=1}^{3} \frac{a(i,j,k)}{3}. \tag{4.4}$$

Die gewichtete Lösungsmittel-zugängliche Oberfläche $WSAS_i$ des Atoms entspricht der Summe über alle Oberflächen-Anteile.

$$WSAS_i = \sum_j wsas_{i,j} \tag{4.5}$$

Für eine einfachere Interpretierbarkeit wird die berechnete zugängliche Oberfläche eines Atoms $WSAS_i$ mit der Oberfläche einer Kugel mit entsprechendem Radius normiert. Das ergibt einen Wert acc_i zwischen 0 für ein unzugängliches Atom und 1 für ein isoliertes hydrophobes Atom ohne eine Präferenz für eine Ausrichtung der Wechselwirkung.

$$acc_i = \frac{WSAS_i}{4\pi \cdot \left(r_{SAS} + r^i_{vdW}\right)^2} \tag{4.6}$$

acc_i ist die Zugänglichkeit eines Atoms i, $WSAS_i$ die berechnete gewichtete zugängliche Oberfläche. r_{SAS} ist der Radius der Lösungsmittel-Kugel und r^i_{vdW} der *van der Waals*-Radius von i. Die Abbruch-Bedingungen der Rekursion und eine genauere Beschreibung des WSAS-Algorithmus sind in Abschnitt 5.1 aufgeführt.

4.2.3. Alternative Methoden zur Berechnung der Zugänglichkeit

Der WSAS-Algorithmus liefert überwiegend gute Ergebnisse. Allerdings gibt es Grenzfälle, bei denen die Verwendung der WSAS als Maß der Desolvatisierungs-Energie un-

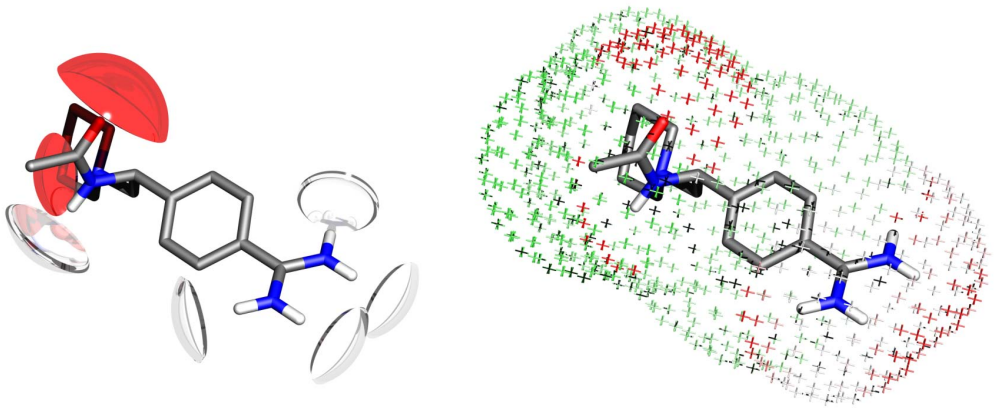


Abbildung 4.4.: Einfluss der Wechselwirkungs-Flächen auf die WSAS und die Desolvatisierungs-Energie. Die Kugel-Kappen im linken Bild repräsentieren Wechselwirkungs-Flächen von Wasserstoff-Brücken-Donoren (weiß) und -Akzeptoren (rot). Die Kreuze im rechten Bild sind die Eckpunkte der Tesselierungs-Prozedur. Ihre Farbe ist ein Maß für die Desolvatisierungs-Kosten und ergibt sich durch Multiplikation der berechneten Zugänglichkeit mit dem plogP-Wert des entsprechenden Atoms (grün: Energie-Gewinn, rot: Bestrafung, weiß: kein Beitrag).

zulänglich erscheint. Einzelne Wasser-Moleküle an schwer zugänglichen Stellen des Proteins haben oft eine große Bedeutung für den Bindungsmechanismus in einem Protein-Ligand Komplex. Sie sind strukturell konserviert und dienen als wichtiger Mediator von Wasserstoff-Brückenbindungen zwischen Protein und Ligand. Das Entfernen eines solchen Moleküls kann von entscheidender Bedeutung für die Affinität zwischen einem Liganden und einem Protein sein. Ein bekanntes Beispiel ist die HIV-1 Protease, bei der ein konserviertes Wasser-Molekül durch `N ILE 150A` und `N ILE 150B` fest gebunden ist. Während einige Inhibitoren dieses Molekül verdrängen, bleibt es bei anderen erhalten und bildet wichtige Wechselwirkung mit diesen aus [140]. Da das Wasser-Molekül praktisch nur an einer einzigen Stelle sitzen kann, wird die zugängliche Oberfläche der benachbarten Protein-Atome nur durch einen Punkt repräsentiert und ist damit gleich 0 (Abbildung 4.5). Das Verdrängen dieses Moleküls durch einen Inhibitor würde sich nicht in einer Änderung der WSAS äußern. Trotz des Aufbrechens wichtiger Wasserstoff-Brücken zwischen Protein und Lösungsmittel wäre der Beitrag zur Desolvatisierungs-Energie bei dieser Art der Zugänglichkeits-Berechnung unberücksichtigt.

Weiterhin ist es bei Verwendung des WSAS-Algorithmus nicht möglich, zwischen sehr ähnlichen Konformationen eines Liganden in der Protein-Bindetasche zu unterscheiden. Sobald ein Wasser-Molekül nicht mehr zwischen Ligand und Protein passt, ist der entsprechende Beitrag 0. Es kann nicht zwischen einem kleinen Hohlraum und einem Kontakt

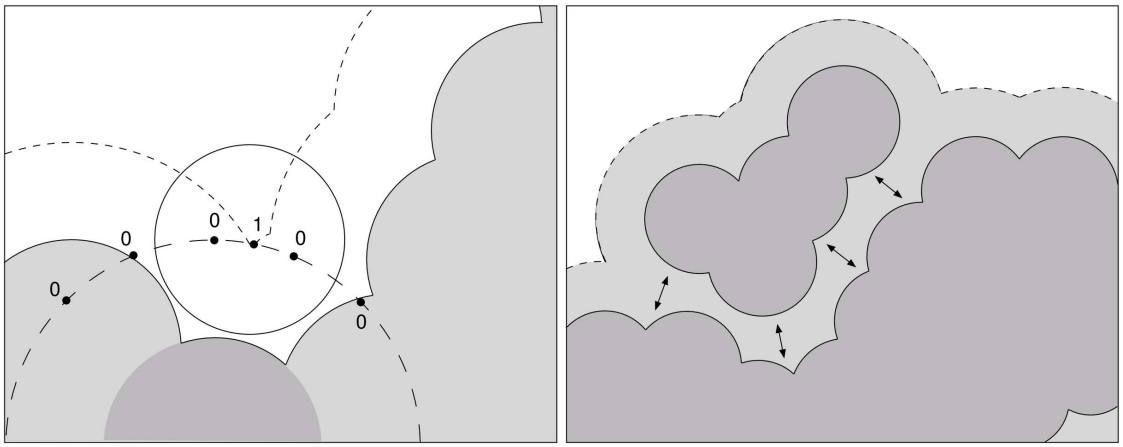


Abbildung 4.5.: Berechnung der WSAS in Kavitäten (links). Änderung in der Position des Liganden mit nur geringen Auswirkungen auf die WSAS (rechts).

der van der Waals Flächen von Protein und Ligand unterschieden werden. Eine exakte Komplementarität in der Oberfläche wird durch diesen Term nicht bevorzugt (vgl. Abbildung 4.5).

Gewichtetes Strahlen-Modell WRAY

Um die aufgeführten Fälle besser bei der Berechnung der Desolvatatisierungs-Energie berücksichtigen zu können, wurde eine zweite Methode zur Berechnung der Zugänglichkeit implementiert. Diese basiert ähnlich dem vom Stahl *et al.* [141] vorgeschlagenen Ansatz zur Skalierung der Wasserstoff-Brücken-Energie auf einem Strahlen-Modell. Ausgehend vom Mittelpunkt des betrachteten Atoms wird durch die vom SAS-Algorithmus berechneten Eckpunkte ein Strahl gezogen und überprüft, in welchem Abstand $r(i,j,k)$ er zuerst die van der Waals Oberfläche eines benachbarten Atoms trifft. Die Zugänglichkeit $a(i,j,k)$ des Eckpunktes wird berechnet als

$$a(i,j,k) = \min\left(1, \max\left(0, \left(\frac{r(i,j,k) - r^i_{vdW}}{2 \cdot r_{SAS}}\right)^3\right)\right) \cdot f_{adev}(i,k) \tag{4.7}$$

Der Exponent in der Funktion dient dazu, den Beitrag zur WRAY in engen Bereichen gegenüber gut zugänglichen Bereichen stark abzuschwächen. Die zugängliche Oberfläche wird analog Gleichung 4.4 berechnet. Für gut zugängliche Oberflächen-Bereiche ergibt der

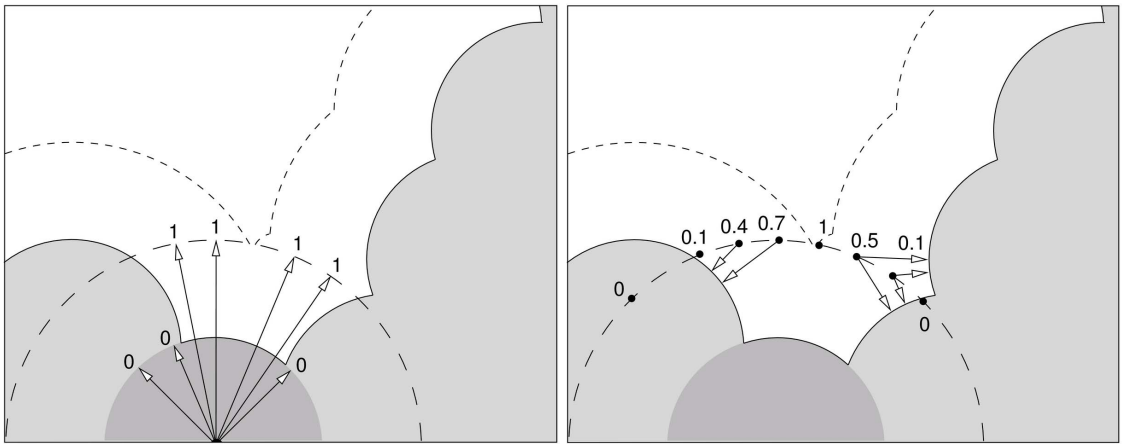


Abbildung 4.6.: Zugänglichkeits-Berechnung entsprechend dem WRAY- (links) und dem SWSA-Algorithmus (rechts).

WRAY-Algorithmus ähnliche Werte wie der WSAS-Algorithmus. In engen Bindetaschen ist die berechnete Zugänglichkeit größer. Die macht sich vor allem im Fall der im vorigen Abschnitt diskutierten konservierten Wasser-Moleküle bemerkbar. Eine Desolvatisierung solcher Bereiche führt nun zu einem größeren Beitrag in der Desolvatisierungs-Energie.

Der Vorteil des WRAY-Algorithmus gegenüber der WSAS ist die stärkere Berücksichtigung schwer zugänglicher Bereiche. Dies hat allerdings auch einen ungünstigen Effekt bei der Berechnung der Desolvatisierung-Energie. Da sich die durch Kugeln repräsentierten Atome von Protein und Ligand nicht überlappen können, gibt es bei einem Kontakt immer Eckpunkte für die eine Zugänglichkeit $a(i, j, k) > 0$ berechnet wird. Eine vollständige Desolvatisierung infolge der Komplex-Bildung wird somit selten erreicht und damit trägt ein Atom gemäß Gleichung 3.3 nicht mit dem gesamten Äquivalent für sein plogP-Inkrement bei. Im Gegenzug ist die Berechnung der Wechselwirkungs-Energie aber nicht abhängig von der Zugänglichkeit und Wasserstoff-Brücken werden im Endeffekt stärker belohnt, als dies nach der Theorie der Fall sein sollte.

Kombination von WSAS und WRAY: SWSA

Es ist schwierig, die beiden vorgestellten Algorithmen miteinander zu vergleichen und gegeneinander abzuwägen. Beide haben Vorteile und Nachteile bezüglich der Berechnung der Desolvatisierungs-Energie. Eine generelle Aussage, welche Methode verlässlichere Ergebnisse ergibt scheint nicht möglich. Um die Vorteile beider Methoden zu vereinen, wurde ein weiterer Algorithmus implementiert, der im Gegesatz zu den Flächen- (WSAS)

bzw. Strahlen-basierten (WRAY) Ansätzen ein Maß für zugängliche Volumen beschreibt. Dazu wird die Zugänglichkeits-Berechnung der WSAS-Methode mit dem Strahlen-Modell erweitert. Im Gegensatz zur Berechnung der WRAY gehen die Strahlen nicht vom Mittelpunkt des Atoms aus sondern von den Eckpunkten der Tesselierungs-Prozedur. Der Arbeitstitel der Methode lautet SWSA (*smoothed weighted solvent accessibility*).

Für jeden Eckpunkt k auf der SAS-Oberfläche eines Atoms werden die Abstände zu allen umgebenden Atomen l bestimmt. Ist der Abstand $r_{k,l}$ zu einem Atom geringer als der SAS-Radius, so verringert sich das Gewicht des Eckpunktes bei der Berechnung der Zugänglichkeit.

$$a(i,j,k) = \prod_l \left(\min \left(1, \max \left(0, \frac{r_{k,l} - r^j_{vdW}}{r_{SAS}} \right) \right) \right) \cdot f_{adev}(i,k) \tag{4.8}$$

4.3. Wechselwirkungen

4.3.1. Optimierung des Wechselwirkungs-Modells

Wechselwirkungs-Geometrien in FlexX

In FlexX wird die Qualität einer Wechselwirkung zwischen zwei Atomen i und j analog dem Böhm-Modell anhand der relativen Anordnung von sogenannten Wechselwirkungs-Geometrien ausgewertet [120]. Allen Atomen, die potenziell in der Lage sind, Wechselwirkungen auszubilden, werden ein oder mehrere Wechselwirkungs-Typen und Wechselwirkungs-Flächen zugeordnet, welche Ausschnitte einer Kugel-Oberfläche sind. Eine Wechselwirkung zwischen zwei Atomen i und j ergibt einen Energie-Beitrag, wenn deren Wechselwirkungs-Typen kompatibel zueinander sind und die jeweiligen Wechselwirkungs-Zentren auf den Wechselwirkungs-Flächen des anderen Atoms liegen (Abbildung 4.2). Abweichungen von einer idealen Geometrie werden mit einem Faktor $f_{dev}(i,j)$ bestraft.

$$f_{dev}(i,j) = f_{ldev}(\Delta d_{i,j}) \cdot f_{adev}(\Delta\alpha_i) \cdot f_{adev}(\Delta\alpha_j) \qquad \forall\, f_{dev}(i,j)\,(0 \leq f_{dev}(i,j) \leq 1) \tag{4.9}$$

$f_{dev}(i,j)$ ist der Abweichungs-Faktor für die spezifische Geometrie zwischen den wechsel-

wirkenden Atomen i und j. f_{ldev} bestraft Abweichungen Δd von einem idealen Abstand, und mit f_{adev} werden Abweichungen $\Delta\alpha$ von den entsprechenden Winkeln zwischen i und j bestraft. f_{ldev} und f_{adev} sind 1 für eine gute Geometrie und nehmen linear mit zunehmendem Abstand bzw. Winkel ab.

Die Programm-interne Repräsentation der Wechselwirkungs-Flächen erfolgt über ein lokales Koordinaten-System, das von dem entsprechenden Atom, dem die Wechselwirkung zugeordnet ist, und den benachbarten Atomen innerhalb der funktionellen Gruppe aufgespannt wird. Der Öffnungswinkel und die Direktionalität der Flächen werden in Polar-Koordinaten angegeben. Die in FlexX verwendeten Wechselwirkungs-Geometrien sind größtenteils von den in LUDI verwendeten Geometrien [14] übernommen worden, welche auf einer statistischen Auswertung von kristallographischen Daten beruhen. Seit 1992 hat die Verfügbarkeit solcher Daten stark zugenommen (vgl. Abbildung 2.4). Angesichts der größeren Bedeutung der Wechselwirkungs-Flächen in HYDE, die einen entscheidenden Einfluss auf die Berechnung der Desolvatisierung-Energie haben, erschien es in einigen Fällen als notwendig, die Geometrie-Parameter neu zu definieren.

Neudefinition der Wechselwirkungs-Flächen

Aus der ISOSTAR-Datenbank [20] wurden für verschiedene Wasserstoff-Brücken-Donoren und -Akzeptoren die Verteilungen der jeweiligen Wechselwirkungs-Partner ermittelt und deren räumliche Position in einer mol2-Datei gespeichert. Diese Daten wurden mit einem Perl-Skript ausgewertet und Verteilungs-Kurven bezüglich der Lage der Wechselwirkungs-Partner erstellt. Die Analyse wurde dabei in Bezug auf die durch die internen Koordinatensysteme gegebenen Polarkoordinaten durchgeführt, die weitere Auswertung erfolgte manuell. Die neuen Wechselwirkungs-Geometrien wurden so ausgewählt, dass die am häufigsten besetzten Bereiche um ein Atom den vollen Wechselwirkungs-Beitrag ergeben. Weniger populierte Bereiche werden durch die Strafterme f_{ldev} und f_{adev} erfasst. Für die Auswertung wurden Daten aus der CSD [5] verwendet. Bei diesen Daten ist im Gegensatz zur PDB auch die Lage der Wasserstoff-Atome experimentell bestimmt, welche für die Ausrichtung von Wasserstoff-Brückenbindungen von entscheidender Bedeutung sind. Die neuen Wechselwirkungs-Flächen wurden so definiert, dass solche Bereiche in den jeweiligen Verteilungs-Histogrammen, deren Anteil größer als 30% des maximalen Anteils ist, den vollen Energie-Beitrag ergeben.

Während sich bei den meisten Geometrien nur geringe Änderungen ergeben haben, wurde z.B. für die Carbonyl-Gruppe eine große Abweichung zwischen dem bisher verwendeten Modell und der tatsächlichen Verteilung der Wechselwirkungs-Partner festgestellt. In Abbildung 4.7 ist links das alte Modell zu sehen. Der rote Schirm repräsentiert die Fläche,

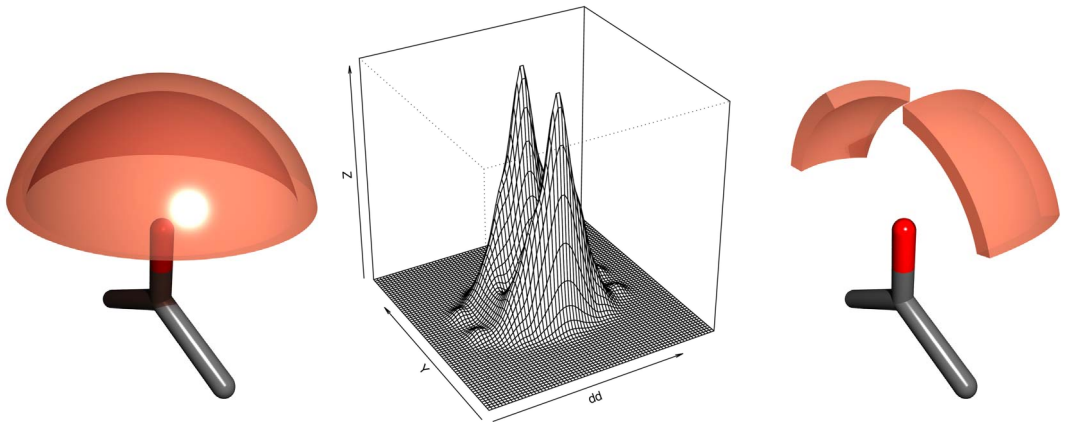


Abbildung 4.7.: Neudefinition der FlexX-Wechselwirkungs-Geometrien anhand einer statistischen Analyse über Verteilungen von Wechselwirkungs-Partnern. Die roten Schirme repräsentieren Regionen um eine Carbonyl-Gruppe, in denen ein Wasserstoff-Atom eines Wasserstoff-Brücken-Donors liegen muss, um eine ideale Wechselwirkung mit vollem Energie-Beitrag zu ergeben. Im alten Modell (links) ist die in der ISOSTAR-Datenbank [20] beobachtbare Aufteilung in mehrere bevorzugte Bereiche nicht berücksichtigt. Eine manuelle Auswertung führte zu dem neuen Modell mit zwei Wechselwirkungs-Flächen (rechts).

auf der ein wechselwirkendes Atom seinen vollen Beitrag erhält. Die statistische Verteilung in der CSD zeigt jedoch, dass Wechselwirkungen überwiegend in der Richtung der freien Elektronen-Paare auftreten. Das neue Modell (rechts) berücksichtigt diesen Umstand wesentlich besser. Dies führt insbesondere dazu, dass eine Desolvatation in Bereichen außerhalb der Carbonyl-Ebene durch HYDE nicht bestraft wird. Dies hatte ursprünglich zu Problemen bei Komplexen geführt, bei denen die Bindung eines Liganden zwischen zwei β-Faltblatt Strukturen stattfindet. Die Backbone-Carbonyl-Gruppen solcher Strukturen sind durch intramolekulare Wechselwirkungen innerhalb des Proteins abgesättigt, jedoch führte eine Verwendung des alten Modells zu einem großen Strafterm bei der Desolvatation, der nicht durch Wechselwirkungen ausgeglichen wurde. Die Verkleinerung der Wechselwirkungs-Fläche verhindert dies, so dass auch solche Strukturen korrekt bewertet werden können.

Es sei jedoch angemerkt, dass diese Definition auf manueller Basis erfolgt ist und einer endgültigen Überarbeitung bedarf. Diese sollte anhand einer automatischen Auswertung von PDB-Strukturen erfolgen. An dieser Stelle wäre zu überlegen, das in FlexX verwendete Wechselwirkungs-Modell grundlegend zu überarbeiten. Statt der diskretisierten Wechselwirkungs-Flächen wäre auch eine Verwendung von Gauß-Funktionen denkbar, die in der Quantenmechanik zur Beschreibung von Aufenthaltswahrscheinlichkeiten verwendet werden.

4.3.2. Sättigung als Funktion der Zugänglichkeit

In Abschnitt 3.3 wurde beschrieben, wie die Wechselwirkungs-Energie mit HYDE berechnet wird. Von Bedeutung ist unter anderem ein Maß für die Wechselwirkungs-Sättigung sat_i. Dieses wurde über einen Wurzel-Ausdruck (Gleichung 3.8) definiert, da es nicht möglich scheint, einen Wert für eine mittlere Sättigung als Bezugs-Größe anzugeben, ohne einen systematischen Fehler bei der Bewertung bestimmter funktioneller Gruppen zu machen.

In einem alternativen Ansatz wird die Sättigung und damit die Energie einer Wasserstoff-Brücke nur noch indirekt über die Geometrie der Wechselwirkungs-Partner berechnet. Der Energie-Beitrag eines Atoms i zu einer Wechselwirkung ist darin proportional zu der Änderung in der Zugänglichkeit Δsat^i, die durch die wechselwirkenden Atome j verursacht wird.

$$\Delta sat^i = \sum_j \frac{acc^i_{complex,j} - acc^i_{free}}{acc^i_{mean}} \qquad \forall i \forall j (f_{dev}(i,j) > 0) \tag{4.10}$$

Bei der Berechnung von $acc^i_{complex,j}$ wird nur der entsprechende Wechselwirkungs-Partner j als zusätzliches Atom berücksichtigt (vgl. Abbildung 4.8). Voraussetzung ist, dass die Atome i und j kompatibel vom Wechselwirkungs-Typ sind und gemäß ihrer relativen Anordnung Wasserstoff-Brücken zueinander ausbilden können. Je zentraler ein Atom dabei auf der gewichteten Oberfläche ist, desto größer ist die Änderung in der Zugänglichkeit und der entsprechende Energie-Beitrag durch die Wechselwirkung. Die Direktionalität einer solchen Bindung geht über den Gewichtungs-Faktor bei der Berechnung der zugänglichen Oberfläche in diese Rechnung mit ein. Das Problem der multiplen Wechselwirkungen eines Atoms und der damit verbundenen Schwierigkeit, einen Ausdruck für die mittlere Sättigung des Atoms zu finden wird dadurch auf ein geometrisches Problem reduziert. Ein Atom enthält dann den vollen Wechselwirkungs-Beitrag, wenn es aufgrund der Wechselwirkung mit anderen Atomen vollständig desolvatisiert wird.

Erweiterungen des Zugänglichkeits-Schemas

Ein Nachteil dieser Methode ist, dass die Wechselwirkungs-Flächen per Definition teilweise größere Bereiche abdecken, als durch einen Wechselwirkungs-Partner eingenommen werden können. Im Falle einer Wechselwirkung wird dadurch nicht die komplette zugängliche Oberfläche durch das Partner-Atom desolvatisiert und der resultierende Energie-

Beitrag entspricht nicht dem vollen Energie-Äquivalent des plogP-Wertes. Dennoch ist in vielen Fällen aus sterischen Gründen keine zusätzliche Wechselwirkung mit anderen Atomen möglich, der Desolvatations-Term von HYDE überwiegt somit den Wechselwirkungs-Term. Insbesondere bei Verwendung des WRAY-Algorithmus für die Berechnung der Zugänglichkeit würden sehr kleine Wasserstoff-Brücken-Energien resultieren, da eine einzelne Kugel grundsätzlich nur zu einer geringen Änderung in der Zugänglichkeit führen kann.

Um einen größeren Beitrag für die Wechselwirkungs-Energie zu erhalten, besteht die Möglichkeit, bei der Berechnung der Zugänglichkeits-Änderung weitere Atome in der unmittelbaren Nachbarschaft des Wechselwirkungs-Partners zu berücksichtigen, welche über eine maximale Anzahl von Bindungen mit diesem verbunden sind. Weiterhin möglich ist eine artifizielle Erweiterung des van der Waals Radius des jeweiligen Atoms j, um damit einen größeren Bereich abzudecken. Beide Möglichkeiten sind in Abbildung 4.9 dargestellt. Welche Einstellung letztendlich verwendet werden sollte, lässt sich nicht mit allgemeiner Gültigkeit angeben, da alle aufgeführten Methoden auch Nachteile mit sich bringen. Letztendlich kommt es hier auf die Expertise des Anwenders an und das zu bewertende System.

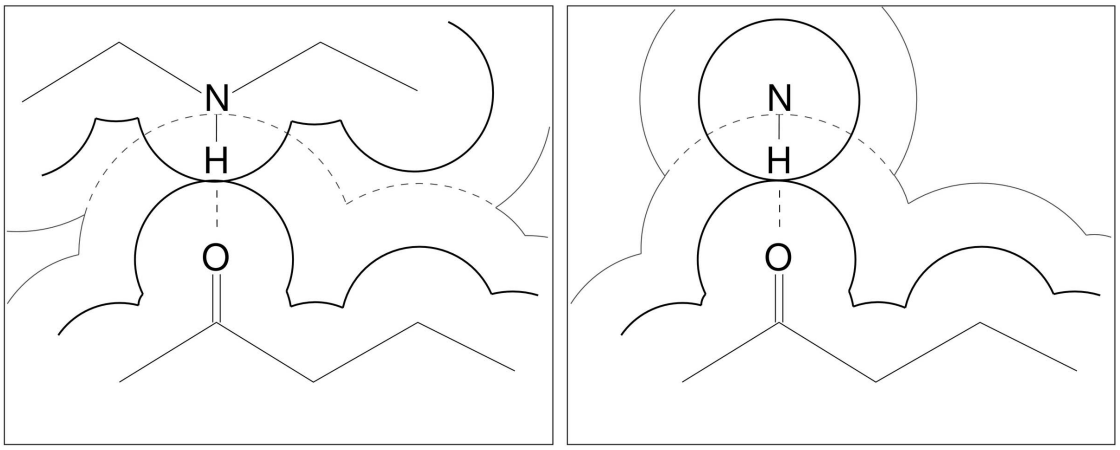


Abbildung 4.8.: Berechnung der Sättigung über die Zugänglichkeits-Änderung. Im linken Bild sei die zu bewertende Wechselwirkung zwischen einem Amin und einer Carbonyl-Gruppe dargestellt. Der Beitrag zur Wasserstoff-Brücken-Energie des Sauerstoffs ergibt sich aus der durch das Stickstoff-Atom verursachten Änderung in der zugänglichen Oberfläche (repräsentiert durch die gestrichelte Linie).

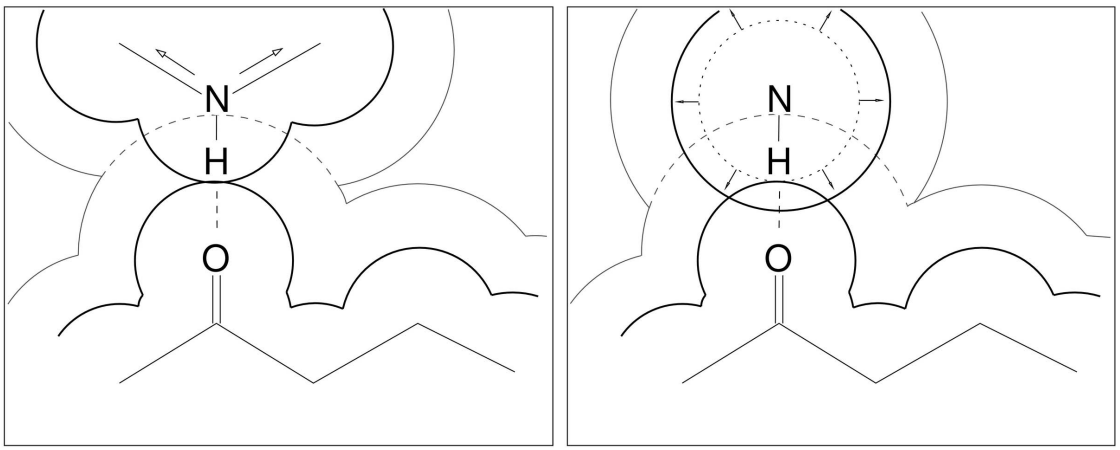


Abbildung 4.9.: Erweiterungen des Zugänglichkeits-Modells. Bei der linken Methode werden zusätzlich die benachbarten Atome von j berücksichtigt. Die rechte Methode erweitert den Radius von j um einen bestimmten Betrag.

5. Implementierung

5.1. Beschreibung der Algorithmen

5.1.1. Berechnung der Wechselwirkungs-Energie

Die Berechnung der Wechselwirkungs-Energie unterscheidet sich von der G_{match} Funktion in FlexX in der Hinsicht, dass zusätzlich zu den Wechselwirkungen zwischen Protein und Ligand auch intramolekulare Wechselwirkungen, also solche zwischen Ligand-Atomen oder denen im Protein einbezogen werden. Dies ist von essentieller Bedeutung, um eine konsistente Bewertung der Desolvatation zu ermöglichen. Auf der Seite des Liganden sind solche Wechselwirkungen seltener zu beobachten, im Protein jedoch tragen sie wesentlich zur Stabilisierung der Tertiärstruktur bei. Vor allem Wasserstoff-Brückenbindungen zwischen den Amid-Gruppen des Proteins sind ein häufig wiederkehrendes Element.

```
1  SATURATION_MATCH_ENERGY (LIGAND ATOMS L, RECEPTOR
   ATOMS R)
2  E_I ← 0
3  M_{L,R} ← ∅                        // list of matches between L and R
4  forall ligand atoms L[i] of L do
5      M_{L[i],R} ← generate_matchlist(L[i], R)
6      append_matchlist(M_{L,R}, M_{L[i],R})
7      M_{L[i],L} ← generate_ligand_matchlist(L[i], L)
8      E_I ← E_I + calc_interaction_energy(L[i], M_{L[i],R}, M_{L[i],L})
9  forall receptor atoms R[i] in matchlist M_{L,R} do
10     M_{R[i],L} ← get_matches(R[i], M_{L,R})
11     M_{R[i],R} ← recmatches[R[i]]  // annotated on molecule initialization
12     E_I ← E_I + calc_interaction_energy(R[i], M_{R[i],L}, M_{R[i],R})
13 return E_I
```

Die Bewertung der Wechselwirkungen erfolgt in zwei Schritten. Zunächst werden für alle

Atome über eine geometrische Nachbarschafts-Suche von kompatiblen Wechselwirkungs-Typen sogenannte *match*-Listen generiert. Mit diesen Informationen wird die Wechselwirkungs-Energie berechnet. Der abgebildete Algorithmus `saturation_match_energy()` beschreibt die Berechnung der Wechselwirkungs-Energie. In einer Schleife über alle Atome $L[i]$ des Liganden werden Wechselwirkungen innerhalb des Liganden und solche zum Rezeptor registriert und daraus die Wechselwirkungs-Energie von $L[i]$ berechnet. Die Suche nach Wechselwirkungen innerhalb des Liganden geschieht in einer Schleife über alle Atome und hat daher die Komplexität $O(n)$. Die Rezeptor-Atome verändern ihre Position beim Docking nicht. Aus diesem Grund ist ein indizierter Zugriff auf ihre Koordinaten möglich und die Suche kann in konstanter Zeit durchgeführt werden.

Für die Bestimmung der Wechselwirkungs-Energie auf Rezeptor-Seite lässt sich die paarweise Abhängigkeit der Wechselwirkungen zwischen Ligand und Protein ausnutzen. Es müssen nur diejenigen Rezeptor-Atome $R[i]$ untersucht werden, für die eine Wechselwirkung in der *match*-Liste $M_{L,R}$ registriert ist. Intramolekulare Wechselwirkungen im Rezeptor werden bei der Initialisierung am Atom annotiert und werden mit $O(1)$ ausgelesen. Auch die Berechnung der Energie geschieht in konstanter Zeit, damit ist die Komplexität des Algorithmus von der Größenordnung $O(n^2)$ bezüglich der Anzahl der betrachteten Atome.

5.1.2. Berechnung der Desolvatations-Energie

Die Berechnung der Desolvatations-Energie ist aufwändiger als die Berechnung der Wechselwirkungs-Energie, da für jedes platzierte Atom eine Änderung in der Zugänglichkeit berechnet werden muss, was einen erheblichen Aufwand bezüglich der Rechenzeit darstellt. Aus diesem Grund wird durch die Funktion `plogp_fragment_energy()` direkt der inkrementelle Aufbau des Liganden aus Fragmenten unterstützt. Die Energie wird nur für ein oder mehrere gegebene Fragmente $F[i : \forall i(f_{min} \leq i \leq f_{max})]$ eines Liganden berechnet. FlexX unterteilt Ligand-Moleküle durch Schnitte an rotierbaren Bindungen in mehrere Fragmente. Ausgehend von einem Basis-Fragment $F[i = 0]$ werden im inkrementellen Docking-Algorithmus nachfolgende Fragmente in unterschiedlichen Konformationen mit diesem verbunden. Dadurch ergibt sich eine Baumstruktur mit unterschiedlichen Platzierungen eines Liganden. Eigenschaften von bereits platzierten Elementen müssen nicht neu berechnet werden, sondern werden vererbt.

Bei der Berechnung der Desolvatations-Energie ist zu beachten, dass sich durch den Anbau von einem Fragment die Zugänglichkeit bereits platzierter Atome ändern kann. Dies gilt insbesondere für das Atom, an dem die Verknüpfung mit dem neuen Fragment erfolgt. Einhergehend mit dieser Änderung in der Zugänglichkeit ergibt sich auch eine

Änderung in der Desolvatations-Energie. Dies muss bei der inkrementellen Bewertung durch einen Korrektur-Wert berücksichtigt werden.

Entsprechend der Suche nach Wechselwirkungen werden für alle Ligand-Atome $L[i]$ des aktuellen Fragments zunächst Nachbarschafts-Listen von bereits platzierten Ligand- und Rezeptor-Atomen ermittelt, welche nahe genug an $L[i]$ liegen, um dessen Zugänglichkeit zu beeinflussen. Mit diesen Informationen lässt sich durch `calc_atom_accessibility()` die Zugänglichkeit im ungebundenen Zustand acc_f und im Komplex acc_c berechnen und daraus die Desolvatations-Energie E_{atom}, die am Atom $L[i]$ annotiert wird.

Alle Atome L^*, welche Lösungen lokaler Nachbarschafts-Anfragen waren und nicht zum aktuellen Fragment gehören, werden anschließend in ihrem Energie-Beitrag korrigiert. Aus dem neu berechneten Wert und der am Atom annotierten Energie wird die Differenz gebildet und daraus der korrigierte Beitrag zur Gesamt-Energie E_D berechnet.

Zuletzt erfolgt in analoger Weise die Berechnung der Desolvatations-Energie der Rezeptor-Atome R^*, welche durch die Platzierung des Fragments beeinflusst worden sind. Auch hier muss überprüft werden, ob das Atom schon durch frühere Aufrufe der Energie-Funktion einen Beitrag geliefert hat. Die berechnete Energie wird entsprechend des annotierten Wertes korrigiert. Die Berechnung von acc_f, also der Zugänglichkeit im ungebundenen Zustand, kann in konstanter Zeit durchgeführt werden, da das Protein seine Konformation nicht ändert und der entsprechende Wert bei der Initialisierung der Rezeptor-Struktur berechnet wird.

Die beschriebene Funktion kann auch für die Berechnung der gesamten Desolvatations-Energie des Komplexes verwendet werden, wenn f_{min} und f_{max} entsprechend gewählt werden. In diesem Fall entfällt die Energie-Korrektur für bereits platzierte Ligand-Atome. Die grundsätzliche Komplexität von $O(n^2)$ bezüglich der Anzahl an berücksichtigten Atomen bleibt in jedem Fall bestehen.

Rekursive Berechnung der Zugänglichkeit

Die Komplexität der Zugänglichkeits-Funktion `calc_atom_accessibility()` wurde mit $O(n)$ angegeben. Da diese Funktion einen essentiellen Einfluss auf die benötigte Rechenzeit hat, soll sie im Folgenden genauer untersucht werden. Eine ausführliche Beschreibung des Ablaufs der Funktion ist in Abschnitt 4.2.1 zu finden. Im wesentlichen werden zwölf initiale Eckpunkte I_A eines Ikosaeders auf der SAS-Oberfläche des Atoms A berechnet. Für jeden Eckpunkt wird durch den Aufruf von `calc_point_accessibility()` und

```
1  PLOGP_FRAGMENT_ENERGY (FRAGMENTS F, LIGAND ATOMS L,
   RECEPTOR ATOMS R, FIRST FRAGMENT f_min, LAST FRAGMENT f_max)
2  E_D ← 0
3  L* ← ∅                                          // ligand atom list
4  R* ← ∅                                          // receptor atom list
5  forall ligand atoms L[i] of all fragments F[f_min..f_max] do
       // score all ligand atoms in given fragments
6      N_{L[i],L} ← calc_atom_neighbours(L[i],L)                    // O(n)
7      append_atomlist(L*,N_{L[i],L})
8      N_{L[i],R} ← calc_atom_neighbours(L[i],R)                    // O(1)
9      append_atomlist(R*,N_{L[i],R})
10     acc_f ← calc_atom_accessibility(L[i], N_{L[i],L}, ∅)         // O(n)
11     acc_c ← calc_atom_accessibility(L[i], N_{L[i],L}, N_{L[i],R})
12     E_atom ← calc_desolvation_energy(L[i], acc_c, acc_f)         // O(1)
13     E_D ← E_D + E_atom
14     set_atom_energy(L[i], E_atom)
15 forall ligand atoms L[i] of F[j : j < f_min] do
       // correct already placed ligand atoms
16     if L[i] in L* then
17         N_{L[i],L} ← calc_atom_neighbours(L[i],L)
18         N_{L[i],R} ← calc_atom_neighbours(L[i],R)
19         acc_f ← calc_atom_accessibility(L[i], N_{L[i],L}, ∅)
20         acc_c ← calc_atom_accessibility(L[i], N_{L[i],L}, N_{L[i],R})
21         E_atom ← calc_desolvation_energy(L[i], acc_c, acc_f)
22         E_D ← E_D − energy[L[i]] + E_atom
23         set_atom_energy(L[i], E_atom)
24 forall receptor atoms R[i] in R* do
       // score receptor atoms affected by newly placed ligand atoms
25     N_{R[i],L} ← calc_atom_neighbours(R[i],L)
26     N_{R[i],R} ← calc_atom_neighbours(R[i],R)
27     acc_f ← access[R[i]]          // annotated on molecule initialization
28     acc_c ← calc_atom_accessibility(R[i], N_{R[i],L}, N_{R[i],R})
29     E_atom ← calc_desolvation_energy(R[i], acc_c, acc_f)
30     E_D ← E_D − energy[R[i]] + E_atom
31     set_atom_energy(R[i], E_atom)
32 return E_D
```

`ia_deviation_penalty()` eine Gewichtung bezüglich der übergebenen Nachbarschafts-Listen N_L und N_R und bezüglich der Wechselwirkungs-Geometrie von A durchgeführt.

Ist die Mindestanzahl an Iterationen WSAS_MINLEVEL erreicht, wird für alle durch benachbarte Eckpunkte I_A gebildeten regelmäßigen, kleinsten Dreiecke T überprüft, ob die berechneten Gewichte sich maximal um WSAS_EPS unterscheiden. In diesem Fall wird der Beitrag des jeweiligen Dreiecks zur Zugänglichkeit *acc* entsprechend Gleichung 4.4 berechnet. Andernfalls erfolgt für das Dreieck der rekursive Funktions-Aufruf von `accessibility_patch_approximation()`.

```
1  CALC_ATOM_ACCESSIBILITY (ATOM A, LIGAND ATOMS N_L,
   RECEPTOR ATOMS N_R)
2  acc ← 0
3  level ← 0
4  I_A ← calc_icosaeder_vertices(A)
5  for i ← 1 to 12 do
       // determine weighting factor for each vertex
6      access[I_A[i]] ← calc_point_accessibility(I_A[i], A, N_L, N_R)   // O(n)
7      if is_polar(A) then
8          weight[I_A[i]] ← access[I_A[i]] * ia_deviation_penalty(I_A[i], A)
9      else
           // hydrophobic atoms have no directional preference
10         weight[I_A[i]] ← access[I_A[i]]
11 forall smallest triangles T[i : 1 ≤ i ≤ 20] = (v_1[i], v_2[i], v_3[i]) of vertices I_A do
12     if (level ≥ WSAS_MINLEVEL) and (max_deviation(v_1[i], v_2[i], v_3[i])
       ≤ WSAS_EPS) then
13         acc ← acc + calc_triangle_surface(T[i], level)
14     else
15         acc ← acc + accessibility_patch_approximation(A, N_L, N_R, T[i],
           level + 1)
16 return acc
```

In `accessibility_patch_approximation()` wird zunächst überprüft, ob die maximale Anzahl an Iterationen WSAS_MAXLEVEL erreicht ist. Ist dies der Fall, wird der Beitrag des Dreiecks zur Zugänglichkeit abhängig von der Anzahl zugänglicher Eckpunkte approximiert. Wenn der maximale Rekursions-Level noch nicht erreicht ist, erfolgt eine Verfeinerung der Rechnung durch eine Aufteilung des Dreiecks T in vier neue Dreiecke T^*. Zu diesem Zweck werden drei zusätzliche Eckpunkte auf der Kugelfläche berechnet, welche jeweils den Abschnitt eines Großkreises zwischen zwei Eckpunkten von T halbieren. Für diese neuen Eckpunkte erfolgt wieder die Berechnung von Zugänglichkeit

```
1  ACCESSIBILITY_PATCH_APPROXIMATION (ATOM A, LIGAND
   ATOMS N_L, RECEPTOR ATOMS N_R, TRIANGLE OF VERTICES T = (v_1, v_2, v_3),
   RECURSION DEPTH level)
2  acc ← 0
3  if level ≥ WSAS_MAXLEVEL then
4      approximate_triangle_surface(T, level)
5  else
       // calculate 3 additional vertices which partition T equally
6      T*[i : i ← 1..4] = (v*_1[i], v*_2[i], v*_3[i]) ← subdivide_triangle(T)
7      forall new vertices in T* do
8          CALCULATE ACCESSIBILITY AND WEIGHTS
9      for i ← 1 to 4 do
10         if (level ≥ WSAS_MINLEVEL) and (max_deviation(v*_1[i],v*_2[i],v*_3[i])
           ≤ WSAS_EPS) then
11             acc ← acc + calc_triangle_surface(T*[i], level)
12         else
13             acc ← acc + accessibility_patch_approximation(A, N_L, N_R,
               T*[i], level + 1)
14 return acc
```

und Gewichtungs-Faktoren. Anschließend wird für alle vier Dreiecke überprüft, ob die Abbruchbedingungen der Rekursion erfüllt sind. Dementsprechend wird der Beitrag der Dreiecke berechnet oder es erfolgt ein erneuter Aufruf der rekursiven Funktion.

Die Laufzeit wird im wesentlichen durch die Häufigkeit der Berechnung von Zugänglichkeit und Gewicht eines Eckpunktes bestimmt. Bezüglich der Anzahl der Atome in den Nachbarschaftslisten eines Atoms geschieht dies mit der Komplexität $O(n)$. Wird eine maximale Rekursionstiefe von k angenommen, verhält sich die maximale Laufzeit $T(k,n)$ entsprechend der Gleichung

$$O(T(k,n)) = \begin{cases} 12 \cdot n & : k = 0 \\ (12 + 20 \cdot 4^{k-1} \cdot 3) \cdot n & : k \geq 1 \end{cases} \tag{5.1}$$

Es ist eine exponentielle Abhängigkeit der Laufzeit von der Rekursionstiefe gegeben. Dennoch kann die Komplexität insgesamt mit $O(n)$ angegeben werden, da die maximale Rekursionstiefe durch WSAS_MAXLEVEL als Konstante in die Rechnung eingeht und somit einen Faktor darstellt.

Parametrisierung

Der Algorithmus zur Bestimung der Zugänglichkeit wird über drei Parametern gesteuert, welche als Abbruch-Kriteria der Rekursions-Schleife dienen. Die Auswahl der Parameter erfolgte über eine Berechnung der zugänglichen Oberfläche von 200 Wirkstoff-ähnlichen Molekülen aus dem FlexX Test-Datensatz. Durch eine Test-Reihe mit 165 unterschiedlichen Parameter-Kombinationen wurde der in Tabelle 5.1 aufgeführte Parameter-Satz bestimmt. Gegenüber der aufwändigsten Methode, bei der die maximale Anzahl an Funktions-Aufrufen erfolgt, wird mit diesen Parametern eine mittlere Abweichung der berechneten Zugänglichkeit von 1.5% bei nur 0.4% der benötigten Rechenzeit erzielt. Dies entspricht bezüglich der Berechnung der Desolvatations-Energie auch bei großen Molekülen einer Abweichung von weniger als 1 kJ mol^{-1} und liegt damit in einem für die Anwendbarkeit akzeptablen Bereich.

WSAS_MINLEVEL	0
WSAS_MAXLEVEL	2
SAS_EPS	0.01

Tabelle 5.1.: Standard-Parameter für die rekursive Berechnung der Zugänglichkeit.

5.2. Laufzeitanalyse

Der Fokus bei der Entwicklung von HYDE lag in erster Linie darin, die Vorhersage-Ergebnisse im strukturbasierten Wirkstoffentwurf zu verbessern. Damit verbundene Einschränkungen hinsichtlich der Laufzeit der Anwendung sind tolerierbar, solange eine Anwendbarkeit in der Praxis noch gegeben ist. Tabelle 5.2 listet die mittlere Laufzeit für die Bewertung des Protein-Ligand Komplexes 4dfr [18] unter Verwendung der verschiedenen Möglichkeiten zur Berechnung der Zugänglichkeit auf. Der Ligand Methotrexat besteht aus 66 Schweratomen, die in der Bindetasche der Dihydrofolatreduktase direkten Kontakt zu ca. 120 Rezeptor-Atomen haben. Dieser Komplex eignet sich zur Analyse der verwendeten Algorithmen, da verschiedene Atom-Typen und Arten von zwischenmolekularen Wechselwirkungen in ihm repräsentiert sind. Die mittlere Laufzeit wurde durch 1000 Aufrufe der Funktion `dock_entry_energy()` ermittelt, welche die einzelnen Bewertungs-Funktionen aufruft. Für WSAS_MINLEVEL und WSAS_EPS wurden Standard-Parameter entsprechend Tabelle 5.1 verwendet. Die maximale Rekursionstiefe WSAS_MAXLEVEL wurde variiert.

Im Vergleich zur FlexX Scoring-Funktion wird für den Aufruf von HYDE unabhän-

	WSAS_MAXLEVEL	dock_entry_energy()	plogp_fragment_energy()
FlexX		2.67	
	2	182.65	178.81
WSAS	1	118.16	113.55
	0	25.27	20.65
	2	3146.44	3137.96
WRAY	1	995.18	989.98
	0	175.51	171.54
	2	1146.34	1141.50
SWSA	1	289.14	285.36
	0	56.57	52.55

Tabelle 5.2.: Durchschnittliche Rechenzeit für die Bewertung von 4dfr in Millisekunden. Die Analyse wurde auf einer Intel Xeon 2.66GHz CPU unter einem Linux 2.6.8 Betriebs-System durchgeführt (Compiler-Version: gcc 3.4.3).

gig von der verwendeten Methode eine deutlich verlängerte Rechenzeit benötigt. Die Auswertung der Original-Funktion dauert durchschnittlich 2.67 ms. Die Berechnung der Wechselwirkungs-Energie G_{match} inklusive dem Erstellen der match-Liste und die Auswertung der paarweisen Kontakte G_{lipo}, G_{ambig} und G_{clash} zwischen Protein und Ligand haben einen etwa gleichgroßen Anteil an der Rechenzeit (nicht in der Tabelle dargestellt). Im Vergleich dazu benötigt der WSAS-Algorithmus unter Verwendung der Standard-Parameter die 68-fache Laufzeit. Der SWSA-Algorithmus benötigt die 429-fache Zeit und eine Verwendung der WRAY-Funktion ergibt einen Faktor von 1178.

Der Anstieg in der benötigten Rechenzeit bei Verwendung von HYDE ist auf die Berechnung der Desolvatations-Energie zurückzuführen. Im vorangegangenen Kapitel wurde erläutert, dass die Komplexität der verwendeten Algorithmen bezüglich der Anzahl an Atomen von der Größenordnung $O(n^2)$ ist. Diesbezüglich unterscheiden sie sich nicht von den Algorithmen der Original Scoring-Funktion. In Tabelle 5.2 wird die ebenfalls diskutierte exponentielle Abhängigkeit von der maximalen Rekursionstiefe deutlich. Vor allem bei den Algorithmen WRAY und SWSA ist ein merklicher Anstieg der Laufzeit bei größer werdendem WSAS_MAXLEVEL zu beobachten. Beim WSAS Algorithmus ist dieser Unterschied nicht so deutlich. Der WSAS Algorithmus klassifiziert einen Punkt gerade in engen Bereichen auf der Moleküloberfläche schneller als unzugänglich (dies wurde in Abschnitt 4.2.3 als Nachteil diskutiert). Damit verbunden ist ein schnellerer Abbruch der Rekursions-Schleife, wenn benachbarte Punkte als unzugänglich eingestuft sind. Weiterhin muss nicht der Abstand zu jedem benachbarten Atom berechnet werden. Es genügt, wenn ein Atom zu nahe gelegen ist, um einen Punkt als unzugänglich zu bewerten. Dies ist bei den beiden anderen Algorithmen nicht möglich, da sich die Zugänglichkeit aus

dem Abstand zu allen benachbarten Atomen ergibt.

Für alle drei Algorithmen gilt, dass der wesentliche Anteil der Rechenzeit für die Berechnung von Abständen und Winkeln benötigt wird. Zweiteres zur Ermittlung des Geometrie-bedingten Gewichtungs-Faktors f_{dev}. Diese ohnehin sehr aufwändigen Operationen werden duch den rekursiven Algorithmus besonders häufig aufgerufen, so dass es zu dem beobachteten starken Anstieg in der Rechenzeit kommt.

5.3. Alternativen

Durch die Verwendung von HYDE ist eine sorgfältigere Auswahl von möglichen Inhibitoren für weitere Forschungen zu einem frühen Zeitpunkt der Untersuchungen möglich, es lässt sich Zeit in anschließenden Schritten des Wirkstoffentwurfs einsparen. Dadurch ist der vermehrte Rechenaufwand für eine Neubewertung von einer limitierten Anzahl an Ligand-Platzierungen tolerierbar. Zudem ist nach wie vor ein stetiger Anstieg der zur Verfügung stehenden Rechenkapazitäten gegeben, was einen Teil der zusätzlich benötigten Rechenzeit kompensieren kann. Zukünftigere Entwicklungen an der HYDE Funktion sollten jedoch auch eine Beschleunigung der Berechnungen bei gleichbleibender Qualität der Ergebnisse ermöglichen. Insbesondere im Hinblick auf eine Bewertung von Teil-Lösungen schon während des Docking-Vorgangs ist eine schnellere Berechenbarkeit der HYDE-Funktion wünschenswert.

Die einfachste denkbare Möglichkeit zur Beschleunigung des Verfahrens wäre eine flexible Verwendung der Rekursions-Parameter. So würde es für den Komplex-Aufbau genügen, die Berechnungen mit verringerter Genauigkeit durchzuführen. Erst bei der abschließenden Bewertung des fertig aufgebauten Protein-Ligand Komplexes würden die (langsameren) Standard-Parameter verwendet werden. Dies stellt allerdings keine optimale Lösung dar. Falsch bewertete Teil-Lösungen könnten durch zu große Abweichungen in der berechneten Energie aufgrund der *k-greedy* Strategie der Docking-Routine frühzeitig aussortiert werden. Die korrekte Lösung des Docking-Problems könnte auf diese Weise verloren gehen. Wünschenswert wäre demnach eine schnellere Möglichkeit zur Berechnung der Oberfläche eines Moleküls bei gleich bleibender Genauigkeit.

Aufgrund der zentralen Bedeutung dieses Themas bei der Computer-basierten Beschreibung von molekularen Vorgängen existiert bereits eine Reihe von Algorithmen, welche sowohl analytische als auch numerische Lösungen für die effiziente Berechnung von Oberflächen bieten [43]. Exakte Lösungen wurden erstmals durch Connolly und Richmond formuliert [27, 123]. Auf Grundlage des Satzes von Gauß-Bonnet lässt sich die Oberfläche

über den Zusammenhang zwischen Krümmung und Euler-Charakteristik einer Fläche berechnen. Weiterentwicklungen dieser Methode erzielen eine Beschleuningung der Berechnung durch Umformulierung des Problems, so dass weniger rechenintensive Operationen benötigt werden [36], oder durch eine schnelle Identifikation von unzugänglichen Atomen im Inneren eines Moleküls [115]. Zu dem gleichen Zweck werden von Fraczkiewicz und Braun in der GETAREA Methode Voronoi-Diagramme verwendet [45]. In einem neueren analytischen Ansatz wird die zugängliche Oberfläche eines Atoms über eine Projektion des Moleküls auf eine Ebene ermittelt [61]. Für Fragestellungen, welche die Desolvatation von Biomolekülen betreffen, ist vor allem auch die Änderung von Oberflächen durch die Überlappung von Kugeln wichtig [52, 116, 9]. Auch hier ist die Verwendung von Voronoi-Diagrammen und korrespondierenden Delauny-Triangulierungen hilfreich [97].

Demgegenüber stehen approximative Lösungen, die durch eine Diskretisierung des dreidimensionalen Raums eine schnellere Abschätzung der Oberfläche ermöglichen. Basierend auf der Arbeit von Shrake *et al.* [136] erfolgt eine Repräsentation der Atome entweder durch gleichmässig verteilte Punkte auf der zugänglichen Oberfläche [90, 21] oder durch eine Triangulierung ähnlich der in FlexX verwendeten Methode [28, 37]. Bis zu 50 mal schneller in der Berechnung der Zugänglichkeit als die GETAREA Methode ist eine approximative Funktion von Guvench *et al.* [58], welche die zugängliche Oberfläche eines Atoms über ein Polynom vierter Ordnung in Abhängigkeit des Abstandes der umgebenden Atome beschreibt. Die Koeffizienten der Formel wurden für unterschiedliche Atom-Typen durch eine Regressions-Analyse anhand von exakt bestimmten Werten ermittelt, ebenso wie in Ref. [158]. Rychkov [127] verwendet eine rekursive Methode zur paarweisen Vereinigung von benachbarten Atomen und deren Eigenschaften, die keiner Parametrisierung bedarf. Zudem wurden verschiedene Möglichkeiten verwirklicht, die Besonderheiten in der Topologie von Makromolekülen durch eine passende Repräsentation zu berücksichtigen, etwa durch Verwendung von α-Shapes [34, 92] oder Gauß-Funktionen [159].

Trotz der Vielzahl der veröffentlichten Algorithmen zur Berechnung von zugänglichen Oberflächen und ihrer jeweiligen Vorzüge ist ein Vergleich hinsichtlich ihrer Verwendbarkeit in der HYDE Funktion schwierig. Angaben, welche die Geschwindigkeit betreffen, mit der eine Oberfläche berechnet werden kann, beziehen sich meist auf unterschiedliche Test-Systeme. Zudem ist die Auswirkung der leistungsfähigeren Hardware moderner Systeme auf die Rechenzeit nicht zu vernachlässigen, was einen Vergleich noch schwieriger macht. Viele der Verbesserungen an den vorgestellten Methoden wurden hinsichtlich der schnelleren Identifizierung von unzugänglichen Atomen ohne Einfluss auf die Oberfläche eingeführt. Dies ist für einen Einsatz von HYDE, bei dem nur eine kleine Auswahl von Atomen im Protein-Ligand Interface bewertet wird, von untergeordneter Bedeutung. Zudem erscheint es auf den ersten Blick schwierig, die unterschiedliche Gewichtung der Oberflächen-Anteile in eine der vorgestellten Methoden zu integrieren.

Neben einer Veränderung des gesamten Algorithmus böte sich auch die Möglichkeit, den bestehenden Algorithmus hinsichtlich einer schnelleren Berechenbarkeit zu modifizieren. Entsprechende Datenstrukturen und Konzepte müssten rechenintensive Vorgänge optimal unterstützen. Zum Beispiel wird für den als starr behandelten Rezeptor bereits eine indizierte Datenstruktur verwendet, welche einen schnellen Zugriff auf Atom-Koordinaten in einem gegebenen Bereich erlaubt (*box hashing*) [120]. Eine ähnliche Repräsentation wäre auch auf der Seite des Liganden sinnvoll. Über eine Distanz-Matrix ließe sich eine mehrfache Berechnung von Abständen zwischen denselben Atomen vermeiden, diese Matrix könnte auf später angebaute Fragmente erweitert werden und durch Vererbung ihrer Eigenschaften den inkrementellen Aufbau-Prozess unterstützen. Für eine effektive Berechnung der Abstände von Atomen zu Punkten auf der Oberfläche ließe sich unter anderem die Dreiecksungleichung ausnutzen.

Sehr zeitaufwändig ist auch die Berechnung des Gewichtungsfaktors f_{dev}, der für jeden betrachteten Punkt ermittelt wird. Da sich dieser Faktor bezüglich der internen Geometrie eines Moleküls nicht verändert, wäre es sinnvoll, ihn bereits bei der Initialisierung des Moleküls in einer hierarchischen Datenstruktur ähnlich einem Min-max-Octree [129] zu hinterlegen. Generell lassen sich viele Parallelen zu Problemen in der Kollisions-Erkennung ziehen, in denen solche Datenstrukturen verwendet werden. Gerade bei Anwendungen im Echtzeit-Bereich, etwa aufwändigen 3-D Spielen ist eine sehr zeiteffiziente Verwaltung verschiedener Objekte von großer Bedeutung und eine Übertragung der dort verwendeten Methoden auf das Problem der effizienten Beschreibung von Oberflächen-Eigenschaften sollte zu guten Resultaten führen.

6. Validierung

Eine Scoring-Funktion wie HYDE kann als analytisches Modell zur Beschreibung des Systems Protein/Ligand interpretiert werden. Die Beschreibung von elementaren Interaktionen zwischen Molekülen in Kapitel 3 durch nur zwei grundlegende Energie-Terme ist eine weitgehende Idealisierung der komplexen Vorgänge zugunsten einer besseren Verständlichkeit und Interpretierbarkeit. Ein allgemeiner Beweis für die Richtigkeit eines solchen Modells ist formal nicht möglich, stattdessen muss geprüft werden, inwiefern das Modell geeignet ist, das Realsystem in ausreichender Güte abzubilden (konzeptuelle Prüfung). Weiterhin muss untersucht werden, ob die diskrete Umsetzung des Modells dem gedanklichen Konzept gerecht wird (Modellverifikation) und ob das Modell brauchbare Aussagen liefert, welche Rückschlüsse auf das reale System erlauben (operationelle Modellvalidierung).

Die Entwicklung des gedanklichen Modells, also die Herleitung der Abhängigkeit der intermolekularen Wechselwirkungen von den Eigenschaften des umgebenden Lösungsmittels (Abschnitt 2.5), war nicht primär Bestandteil der Doktorarbeit und wird im Folgenden als valide angenommen. Ein Vergleich der auf Basis dieses Modells berechneten Werte mit experimentell ermittelten Größen zeigt eine gute Übereinstimmung [87].

Auch die konzeptionelle Umsetzung soll an dieser Stelle nicht weiter diskutiert werden. Formell ist die Herleitung der Scoring-Funktion durch die Verwendung des Faktors f_{sat} und den mathematischen Zusammenhang zwischen Desolvatations- und Wechselwirkungs-Energie (Gl. 2.19) sehr nahe an der Theorie. Durch die konsistente Verwendung der richtungsabhängigen Größen, auch bei der Berechnung der Zugänglichkeit, sind die Besonderheiten der Desolvatation von polaren Gruppen gut wiedergegeben. Dies zeigt sich unter anderem in dem in Abschnitt 3.4 angeführten Beispiel beim Austausch eines Wasserstoff-Brücken-Akzeptors durch ein hydrophobes Atom oder ein polares mit Donor-Funktionalität.

Für die operationelle Validierung einer neuen Scoring-Funktion muss man sich vor allem mit drei Fragestellungen auseinandersetzen, die von essentieller Bedeutung für die Anwendbarkeit im strukturbasierten Wirkstoffdesign sind: (1) Ist der berechnete Score ein Maß für die tatsächliche Bindungs-Affinität, (2) kann die Scoring-Funktion unter

verschiedenen Konformationen eines Liganden in der Bindetasche eines Proteins diejenige bestimmen, die am ähnlichsten zum wirklichen Bindungs-Modus ist, und (3) ist die Scoring-Funktion in der Lage, zwischen wirklichen Liganden und solchen Molekülen zu diskrimieren, die keine nachweisbare Affinität zum betrachteten Protein aufweisen?

Prinzipiell verbirgt sich hinter diesen Fragen die Forderung, dass eine Scoring-Funktion die physikalisch-chemischen Besonderheiten eines Protein-Ligand Komplexes berücksichtigen und die entscheidenden Einflüsse auf die Bindungs-Affinität quantitativ richtig beschreiben können muss. Diese Fragestellungen in der praktischen Anwendung zu beantworten, ist jedoch mit Schwierigkeiten verbunden. Genauere Untersuchungen sind nur an Einzelfällen möglich, und Ergebnisse, die für ein bestimmtes Protein richtig sind, besitzen nicht genügend generelle Aussagekraft für alle weiteren möglichen Szenarien, wie zum Beispiel andere Protein-Familien.

Zudem hängt die Bewertung von Protein-Ligand Komplexen umittelbar von der Qualität der zugrunde liegenden Daten ab. Öffentlich zugängliche Datensätze, welche für entsprechende Vergleichs-Studien herangezogen werden, müssen für die Bewertung durch eine Scoring-Funktion meist manuell aufbereitet und in ein bestimmtes Format gebracht werden. Damit fließt das Wissen des Anwenders in den Bewertungs-Prozess mit ein. Entsprechend erklären sich auch die Resultate von vielen Vergleichs-Studien zu existierenden Scoring-Methoden, bei denen Scoring-Funktionen in verschiedenen Szenarien unterschiedlich gut ab abschneiden [13, 42, 75, 96, 110, 112, 114, 156, 155], teilweise mit gravierender Diskrepanz.

Zur Validierung der HYDE Funktion werden im Folgenden drei Test-Szenarien diskutiert. Es wird geprüft, ob es möglich ist, durch Verwendung des HYDE-Scores bekanntes Wissen zu reproduzieren und publizierte Inhibitoren in einem zufälligen Datensatz wiederzufinden. Weiterhin wird untersucht, ob der HYDE-Score ein Maß für die tatsächliche Bindungs-Affinität eines Protein-Ligand Komplexes ist und zuletzt wird die Anwendbarkeit des Modells in einem realen Test-Szenario anhand von experimentellen Ergebnissen geschildert.

6.1. Anreicherung aktiver Verbindungen

6.1.1. Enrichment-Faktoren und Enrichment-Plots

Analog zu der tatsächlichen Suche nach aktiven Verbindungen in einem experimentellen Screening wird bei einer Anreicherungs-Studie, einem sog. *Enrichment*, überprüft,

ob es mittels einer Scoring-Funktion möglich ist, aus einem Datensatz mit verschiedenen Inhibitor-ähnlichen Molekülen diejenigen auszuwählen, für die eine Affinität zu dem untersuchten Protein nachgewiesen ist. In einem nach dem berechneten Score sortierten Datensatz sollten die wirklichen Inhibitoren auf den oberen Rängen zu finden sein. Für ein solches Experiment werden mit einem Docking-Programm verschiedene Konformationen aller Liganden in der Bindetasche eines Proteins erzeugt und mit der Scoring-Funktion bewertet. Die Lösung mit dem höchsten Score wird als wahrscheinlichste Konformation eines Liganden ausgewählt. Die Anreicherung bei i Prozent eines Datensatzes mit N_n^{tot} Molekülen wird als der Enrichment-Faktor EF_i berechnet.

$$EF_i = \frac{N_i^{act}}{N_n^{act}} \times \frac{N_n^{tot}}{N_i^{tot}}. \tag{6.1}$$

EF_i ergibt sich aus dem Verhältnis der in der Unterauswahl gefundenen aktiven Verbindungen N_i^{act} bezüglich der Anzahl an Aktiven im Datensatz N_n^{act} und der jeweiligen Gesamtanzahl an Verbindungen N_n^{tot} und N_i^{tot}.

Der Enrichment-Faktor hängt allerdings von der Größe des betrachteten Datensatzes ab und auch von dem Verhältnis zwischen Inhibitoren und inaktiven Verbindungen in diesem. Seifert [134] hat zudem gezeigt, dass der Enrichment-Faktor nicht als alleiniges Maß für die Validierung in einer Enrichment-Studie geeignet ist. Selbst bei einer Übereinstimmung des jeweiligen Mittelwertes bei der Bewertung von Aktiven und Inaktiven führt ein Unterschied in der Standard-Abweichung zu einer Anreicherung oder Abreicherung auf den oberen Rängen. Gerade für den Vergleich unterschiedlicher Methoden ist er nur eingeschränkt zu gebrauchen.

Eine generellere Aussage lässt sich aus der Betrachtung des zugehörigen Enrichment-Plots ableiten, bei dem die Anzahl der gefundenen aktiven Verbindungen bzgl. des nach Score sortierten Datensatzes aufgetragen wird. Ein ideales Enrichment würde zu einem steilen Anstieg der Kurve führen. Diese Auftragung ist ähnlich zu dem in der Statistik verwendeten ROC-Plot (*receiver operating characteristic*), der die Abhängigkeit von Sensibilität und Sensitivität darstellt. Ein ROC-Plot wird zur Optimierung eines Grenzwertes verwendet, bei dem mit der größten Wahrscheinlichkeit eine korrekte Klassifikation erfolgt. Demgegenüber ist es für eine Anwendung in der Wirkstoff-Forschung vor allem von Interesse, dass auf den obersten Rängen nur aktive Verbindungen gefunden werden. Durch ein virtuelles Screening soll vor allem der zu untersuchende Datensatz reduziert werden, dabei ist es eher hinzunehmen, wenn einige interessante Verbindungen verloren gehen, als wenn durch zu viele falsch positive Hits nachfolgende (aufwändigere) Untersuchungen zu einem negativen Ergebnis kommen.

Bei der Deutung eines Enrichment-Plots sollte einem jedoch stets bewusst sein, dass damit der gesamte Docking-Vorgang bewertet wird, der aus der Aufbereitung der Strukturen, der Generierung und Bewertung von Protein-Ligand Komplexen und dem Vergleich unterschiedlicher Liganden besteht. Ein schlechtes Abschneiden in solch einer Studie ist nicht zwangsläufig auf die Scoring-Funktion zurückzuführen. Ebenso ist auch ein gutes Abschneiden nicht notwendigerweise durch diese bedingt. Es konnte gezeigt werden, dass gute Enrichment-Ergebnisse sogar dann erzielt werden können, wenn die am besten bewertete Konformation eines Liganden nicht der physikalisch richtigen Lösung entspricht [91]. Teilweise genügt eine hohe Bewertung eines charakteristischen Fragmentes, welches in allen Inhibitoren enthalten ist, um diese auf den oberen Rängen im Enrichment zu platzieren. Gerade bei Scoring-Funktionen, die keine Straf-Terme verwenden, wäre es in solche einem Fall egal, ob durch weitere Fragmente der Inhibitoren sinnvolle Wechselwirkungen ausgebildet werden können oder nicht. Das Enrichment wäre damit zwar korrekt (die richtigen Moleküle werden gefunden), die Art und Weise, auf die es berechnet wurde, jedoch nicht zwangsläufig. Eine wirklich Aussage über die Qualität einer Scoring-Funktion kann letztendlich nur die visuelle Inspektion der Ergebnisse und der alltägliche Einsatz bringen.

6.1.2. Validierung am DUD-Datensatz

Für eine Validierung der HYDE Scoring Funktion wurde der kürzlich veröffentlichte DUD (*database of useful decoys*) Datensatz ausgewählt [68]. Dieser Datensatz besteht aus 40 unterschiedlichen Proteinen mit pharmakologischer Relevanz und insgesamt 2950 Liganden mit bekannter Bindungs-Affinität zu diesen. Für jeden dieser Liganden enthält der Datensatz bis zu 36 sogenannte *decoy* Moleküle. Diese wurden so ausgewählt, dass sie ähnliche physikalische Parameter wie die aktiven Moleküle haben, sich in ihrer Topologie jedoch unterscheiden. Auf diese Weise soll sichergestellt werden, dass die Decoy-Verbindungen keine Affinität zum Protein aufweisen, es in einem Enrichment Experiment jedoch nicht zu einer künstlichen Anreicherung aufgrund der physikalischen Eigenschaften kommt. Entscheidende Parameter bei der Erzeugung der Decoys waren unter anderem das Molekulargewicht, die Anzahl drehbarer Bindungen und der logP-Wert, aber auch die An- oder Abwesenheit von funktionellen Gruppen.

In einer initialen Untersuchung wurde die Bindetasche über einen Radius von 6.5Å um die Schwer-Atome des natürlichen Liganden festgelegt. Die Ausrichtung von rotierbaren Hydroxyl-Gruppen und die Protonierung der Rezeptor-Atome erfolgte mit FlexX Standard-Einstellungen (Version 2.1.0). Alle Verbindungen wurden unter Verwendung der Original Scoring-Funktion gedockt. Die besten 100 Lösungen jedes Liganden wurden im mol2 Format gespeichert und in einem zweiten Schritt mit HYDE unter Verwendung

des WRAY Algorithmus neu bewertet. Basierend auf der am besten bewerteten Pose jedes Liganden wurde für alle Proteine ein Enrichment-Plot berechnet (siehe Anhang C).

Für viele Proteine wurde ein gutes Resultat mit einer hohen Anreicherung von Aktiven unter den am besten bewerteten Verbindungen erzielt. Im Vergleich mit den Ergebnissen der FlexX Scoring-Funktion konnte in 13 Fällen das Ergebnis verbessert werden. Allerdings zeigt HYDE bei 15 Verbindungen auch ein schlechteres Enrichment-Verhalten, bei 12 Proteinen ist kein großer Unterschied festzustellen. Wie schon weiter oben angedeutet, lässt sich aus den berechneten Ergebnissen nicht ohne weiteres folgern, dass HYDE besser oder schlechter als Scoring Funktion geeignet ist als andere Funktionen. Die Gründe für ein verschlechtertes Abschneiden bei einem Enrichment können vielfältig sein. Vor allem ist eine große Abhängigkeit von den mit FlexX generierten Konformationen des Liganden gegeben.

Aus diesem Grund wurden zur weiteren Validierung exemplarisch drei Proteine für genauere Untersuchungen ausgewählt. Bei diesen Proteinen lieferte HYDE sowohl gute Ergebnisse (Estrogen Rezeptor) als auch nicht zufriedenstellende (Thrombin, p38 MAP Kinase). Die Bindetasche wurde durch einen erweiterten Radius von 8Å festgelegt und die jeweils 200 besten Resultate für die Bewertung mit HYDE gespeichert. Zudem wurden mit PROTOPLEX [111] für jeden Liganden alle sinnvollen Tautomere und Protonierungszustände erzeugt und separat gedockt. Die berechneten Konformationen wurden anschließend mit MOLMEC unter Verwendung eines erweiterten AMBER Kraftfeldes bewertet und Lösungen mit einer Konformations-Energie von mehr als 60 kJ mol^{-1} verworfen. Für die nachfolgende Auswertung wurde nur das am besten bewertete Tautomer eines Liganden verwendet.

6.1.3. Estrogen-Rezeptor

Estrogen-Rezeptoren gehören zur Klasse der nukleären Hormonrezeptoren und sind entscheidend an der Entwicklung des Organismus beteiligt. Durch eine Komplexbildung mit Estrogenen (weibliche Sexualhormone aus der Klasse der Steroidhormone) wird eine Dimerisierung bewirkt und eine Stabilisierung einer Protein-Konformation, in der der Estrogen-Rezeptor an unterschiedliche Regionen der DNA binden kann. Dadurch werden bestimmte Gene in verstärktem Maße exprimiert und das Wachstum der Zelle angeregt.

Die Bindetasche des Estrogen-Rezeptors ist eine größtenteils hydrophobe Kavität im Inneren des Proteins. Der verwendete Datensatz besteht aus 2570 Decoys und 67 Liganden, die eine agonistische Wirkung zeigen und die aktive Konformation des Proteins stabilisie-

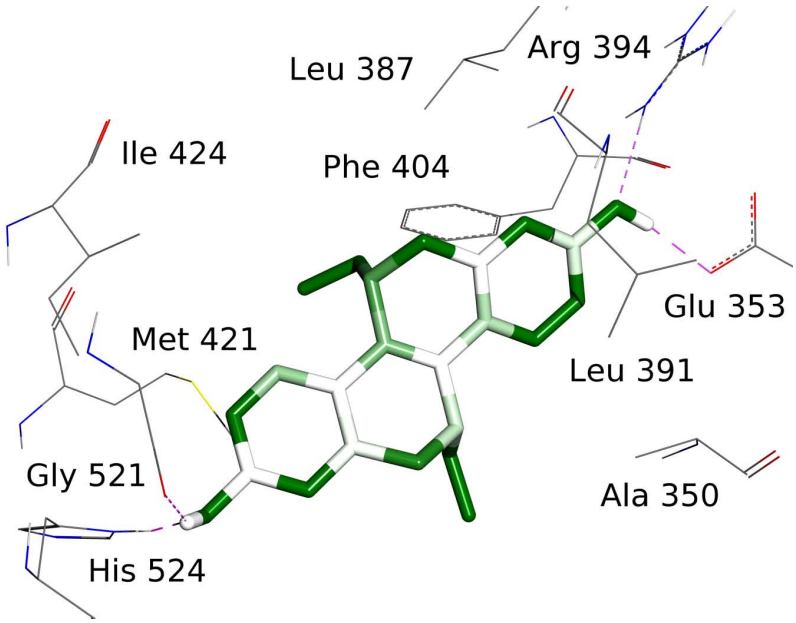


Abbildung 6.1.: HYDE-Bewertung des co-kristallisierten Liganden mit dem Estrogen-Rezeptor 1l2i [135] in Farb-Codierung. Sowohl die Hydroxyl-Gruppen als auch die Kohlenstoff-Atome des ausgedehnten Ringsystems tragen stabilisierend zur Bindungs-Affintät bei.

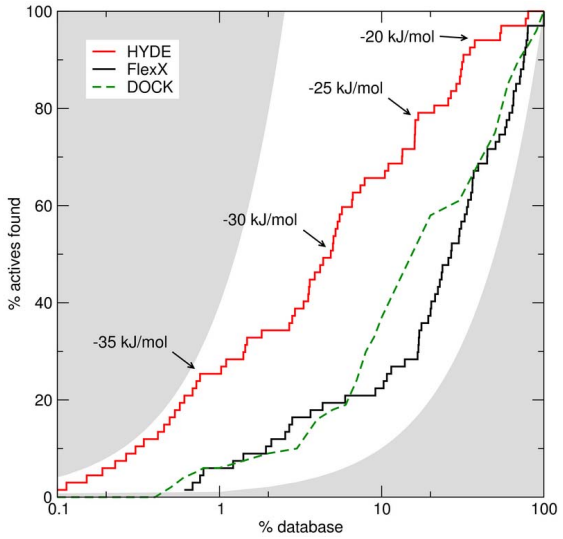


Abbildung 6.2.: Die Abbildung zeigt den berechneten Enrichment-Plot für den Estrogen-Rezeptor 1l2i. Die Anreicherungs-Kurve von FlexX ist in schwarz dargestellt, die Ergebnisse der Nachbewertung mit HYDE in rot. Die grüne Linie repräsentiert die aus der DUD-Publikation entnommenen Ergebnisse, die unter Verwendung des DOCK-Programmes erzielt worden sind [68]. Die grauen Bereiche markieren die maximal erzielbare Anreicherung und eine zufällige Verteilung. Sind die Aktiven auf den höchsten Rängen des sortierten Datensatzes zu finden, würde eine entsprechende Auftragung am oberen Rand des weißen Bereiches verlaufen. Wenn der berechnete Score keine Diskriminierung zwischen Aktiven und Inaktiven zulässt, verläuft die Linie am unteren Rand.

ren. Diese Moleküle sind in ihrer Struktur den Steroiden sehr ähnlich und haben typischerweise ein ausgedehntes hydrophobes Gerüst mit 2 Hydroxylgruppen in einem Abstand von 12Å, welche essentielle Wasserstoff-Brücken mit dem Rezeptor ausbilden können. Die Wechselwirkungen mit dem co-kristallisierten Liganden der verwendeten Protein-Struktur 1l2i [135] sind in Abbildung 6.1 schematisch dargestellt. In der farbcodierten Darstellung des HYDE-Scores ist klar ersichtlich, dass sowohl die Hydroxyl-Gruppen als auch die hydrophoben Bereiche des Liganden stabilisierend zur Bindungs-Affinität beitragen. Der berechnete HYDE-Score dieser Verbindung von -43 kJ mol^{-1} entspricht der experimentell ermittelten Affinität von 10 nM. Mit insgesamt -39 kJ mol^{-1} trägt die Desolvatation hydrophober Atome sehr stark zur berechneten Energie bei. Die Wasserstoff-Brücken haben in der Summe nur einen geringen Anteil, wenn man die Bindungs-Energie und die Bestrafung für die Desolvatation zusammenrechnet.

Die Ergebnisse des Enrichments sind in den Abbildungen 6.2 und 6.3 dargestellt. Gut zu erkennen ist die frühe Anreicherung der aktiven Verbindungen im Datensatz bei Verwen-

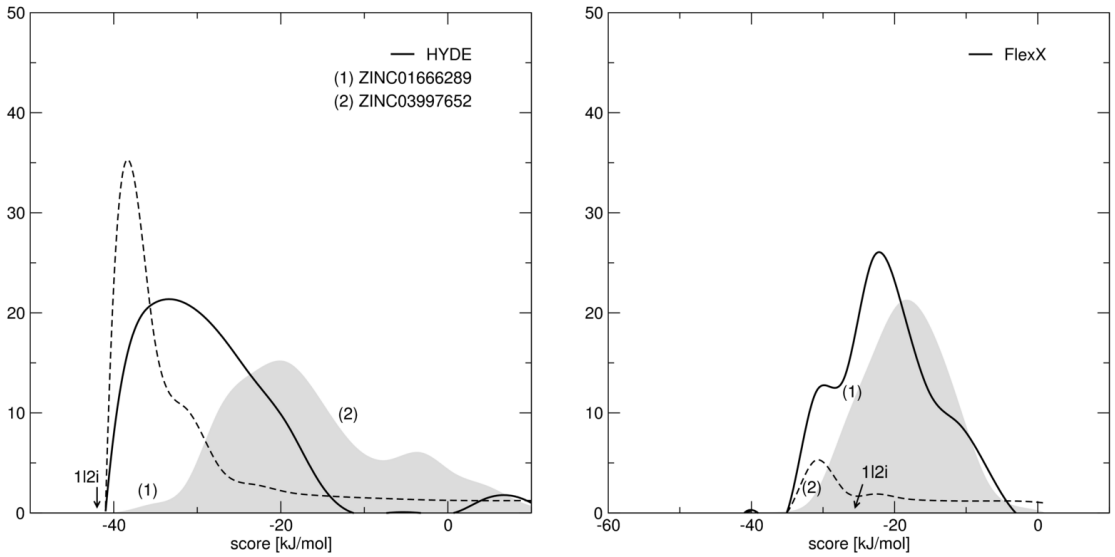
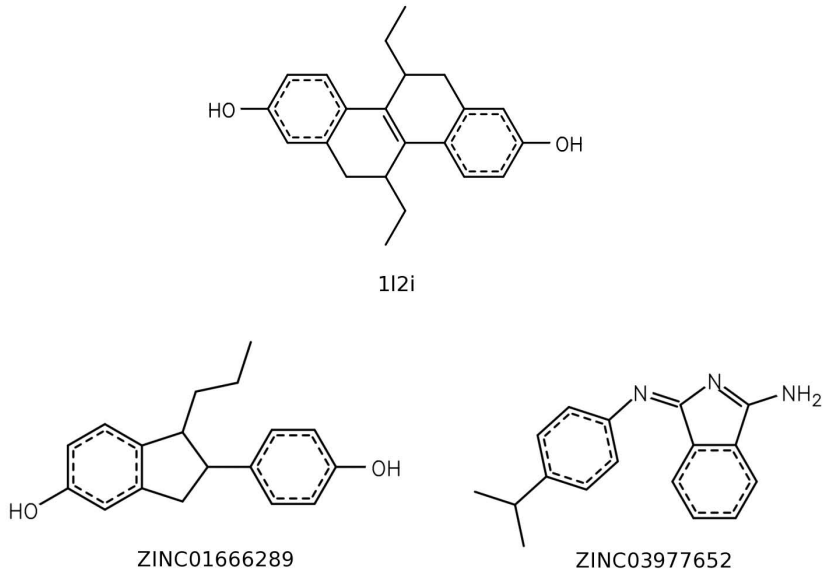


Abbildung 6.3.: Die Diagramme zeigen geglättete Verteilungs-Kurven für die berechneten Werte, aufgeteilt in aktive Verbindungen (schwarze Linie) und Decoys (grauer Bereich). Im linken Diagramm sind die Bewertungen durch HYDE dargestellt, rechts die FlexX-Ergebnisse. Die gestrichelte Linie repräsentiert den Enrichment-Faktor (Gl. 6.1) in Abhängigkeit vom Score.



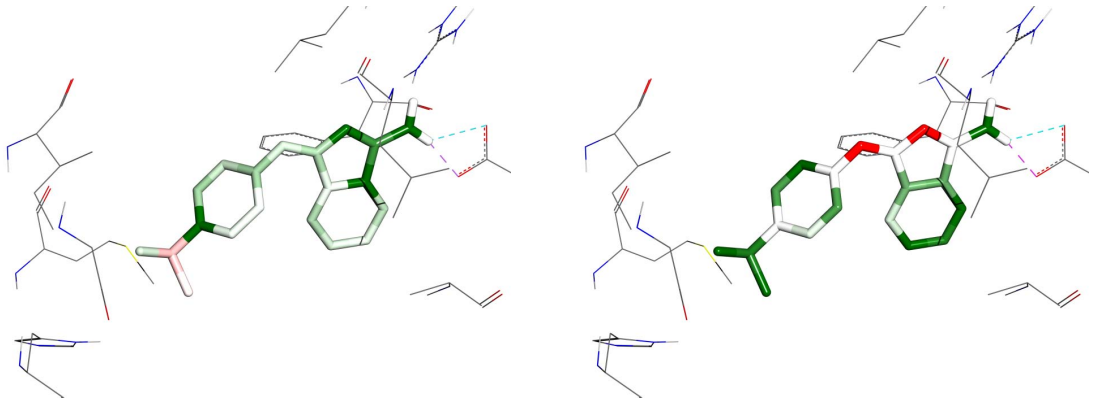


Abbildung 6.4.: Beiträge der einzelnen Atome zur Bewertung von Decoy ZINC03977652 durch FlexX (links) und HYDE (rechts), dargestellt in Farb-Codierung. Die von FlexX als stabilisierend eingestuften Stickstoff-Atome werden im HYDE-Score bestraft.

dung des HYDE-Scores. Bis zu einem Wert von ca. -35 kJ mol^{-1} verläuft die rote Linie im Enrichment-Plot nahezu optimal am oberen Rand des weißen Bereichs, erst dann werden die ersten Decoy-Verbindungen mit hoher Wertung gefunden und es kommt es zu einem kleinen Einbruch im Kurvenverlauf. Eine visuelle Inspektion der Docking-Ergebnisse ergab, dass die überwiegende Mehrheit der Aktiven mit FlexX gut in der Bindetasche platziert werden konnte. Nur zwei Verbindungen konnten nicht ins aktive Zentrum gedockt werden. Diese wurden dementsprechend schlecht durch HYDE bewertet. Sie sind für den kleinen Anstieg in der Score-Distribution ab 0 kJ mol^{-1} verantwortlich. Die Decoys wurden deutlich schlechter bewertet als die Aktiven. Dies ist darauf zurückzuführen, dass sie entweder nicht in der Bindetasche platziert werden konnten oder ungesättigte polare Funktionen zu einer Bestrafung im Energie-Term führen.

Im Gegensatz dazu sind in den durch die FlexX-Scoring-Funktion hoch bewerteten Resultaten viele Decoys zu finden, die einige gute Wechselwirkungen zum Protein ausbilden aber ungesättigte polare Atome haben, die durch den G_{ambig} Term sogar belohnt werden. Abbildung 6.4 zeigt die individuellen Beiträge zur Bewertung der Decoy-Verbindung `ZINC03977652`, die von FlexX mit insgesamt -33 kJ mol^{-1} als Binder klassifiziert wird. Zwei Stickstoff-Atome, die als Wasserstoff-Brücken-Akzeptoren fungieren, werden in der Bindetasche vergraben. Dies wird durch die HYDE-Funktion erkannt und entsprechend bestraft.

Eine genauere Analyse der Ergebnisse legt sogar nahe, dass sich unter den als Decoy klassifizierten Verbindungen einige Substanzen befinden, die sehr wahrscheinlich eine *in vitro* Affinität zum Estrogen-Rezeptor aufweisen. Die Art und Weise, wie diese Moleküle generiert worden sind, sollte sicherstellen, dass es zu keiner ungewollten Anreicherung

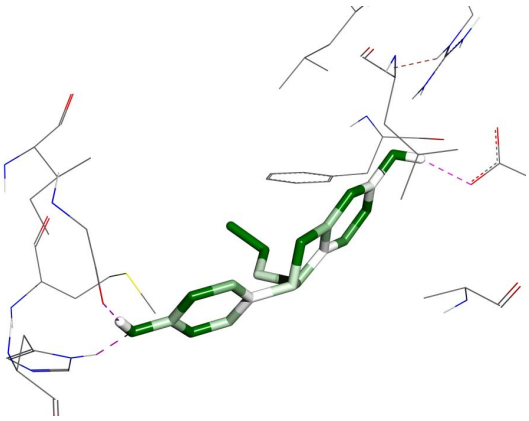


Abbildung 6.5: Bewertung des vermutlich falsch klassifizierten Nicht-Binders ZINC01666289 durch HYDE. Das Wechselwirkungs-Profil entspricht dem des Original-Inhibitors.

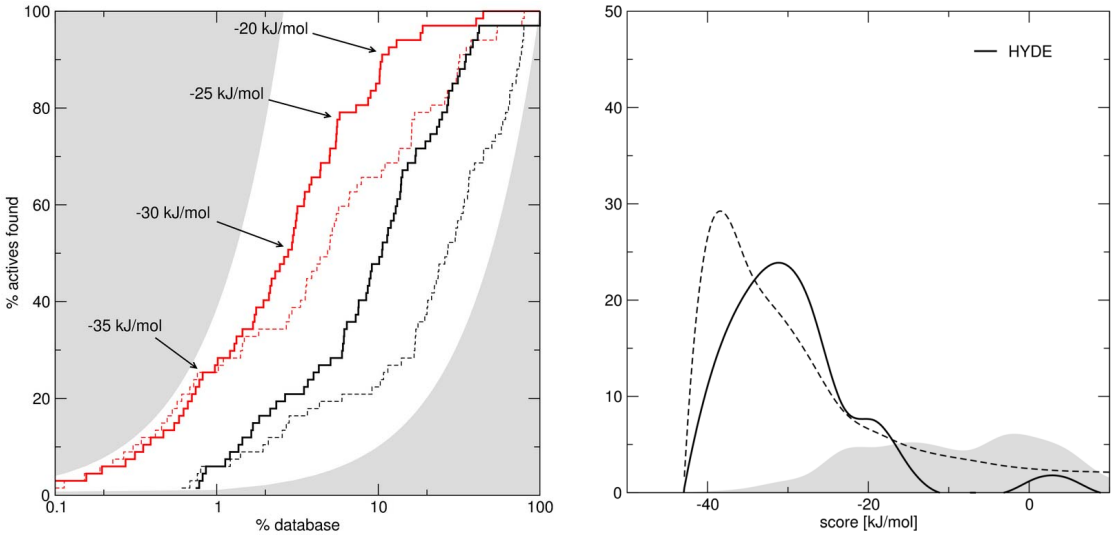


Abbildung 6.6.: Anreicherung Estrogen-Rezeptor. Anstatt der Decoys aus dem DUD-Datensatz wurden 2500 Verbindungen aus einer *in-house* Datenbank verwendet. Im linken Bild ist der Enrichment-Plot dargestellt, im rechten Bild die entsprechende Verteilung des HYDE-Scores.

aufgrund von physikalisch-chemischen Parametern kommt. Eine ähnliche Vorgehensweise wird aber überlicherweise auch bei der Suche nach neuen Liganden verwendet, um über die Eigenschaften von bekannten Inibitoren neue potenzielle Leitstrukturen zu generieren. Als Beispiel sei die Decoy-Verbindung `ZINC01666289` angeführt, die durch HYDE mit -35 kJ mol^{-1} als bindend bewertet wird. Abbildung 6.5 zeigt, dass sie ähnliche Wechselwirkungen zum Protein ausbildet wie der Original Inhibitor in 1l2i.

Um den Einfluss des Decoy-Datensatzes auf die Anreicherungs-Kurve zu verdeutlichen wurden aus einer *in-house* Datenbank 2500 Moleküle ausgewählt, die bezüglich der phyiskalisch-chemischen Eigenschaften ebenfalls in einem ähnlichen Bereich liegen, allerdings eine größere topologische Diversität aufweisen. Deutlich zu erkennen in Abbildung 6.6 ist die Verbesserung im Enrichment sowohl für FlexX als auch für HYDE und die insgesamt schlechtere Bewertung der Decoys. Unabhängig von dieser Änderung ist das mit HYDE berechnete Enrichment wesentlich besser als das mit FlexX berechnete.

6.1.4. Thrombin

Thrombin ist das entscheidende Protein bei der plasmatischen Blutgerinnung. Nach enzymatischer Spaltung der inaktiven Vorstufe Protothrombin über die Gerinnungskaskade katalysiert Thrombin die Spaltung von Fibrinogen. Es entsteht Fibrin, welches polymerisiert und einen netzartigen Wundverschluss bildet. Als Serinprotease ist die Bindetasche charakterisiert durch das katalytisch wirksame Serin `SER 195`, ein negativ geladenes Aspartat `Asp 189` in der ansonsten hydrophoben Spezifitäts-Tasche und weitere, durch eine β-Faltblatt-Struktur abgetrennte, hydrophobe Subtaschen. Bei der enzymatischen Spaltung wird das Serin `SER 195` durch einen benachbarten Histidin-Rest stark polarisiert und zu einem nukleophilen Angriff auf die Amid-Bindung des Substrats befähigt, welcher zur Hydrolyse bzw. Spaltung desselben führt. Eine Inhibition erfolgt üblicherweise durch das Binden eines Liganden direkt im katalytischen Zentrum. Von entscheidender Bedeutung ist ein spezifisches Wechselwirkungs-Muster mit `Asp 189`, `Gly219` und einem konservierten Wasser-Molekül. Den meisten Inhibitoren ist ein Benzamidin-Gerüst gemein, welches diese Wasserstoff-Brücken in der S1 Bindetasche in idealer Geometrie ausbilden kann. Da Benzamidin-Derivate aufgrund ihrer positiven Ladung ungünstige pharmakokinetische Eigenschaften besitzen, wird bei der Erforschung neuer Thrombin-Inhibitoren vor allem Wert auf einen Austausch dieser Gruppe durch ein ungeladenes Molekül-Fragment gelegt. Vielversprechend sind hier unter anderem Chlor-substituierte Phenyle, welche anstatt des Benzamidins in der S1-Tasche platziert werden können [60].

Nach einem unbefriedigendem Enrichment-Ergebnis für Thrombin im ersten Versuch, haben wir uns entschieden, für weitere Untersuchungen nicht die im DUD enthaltene

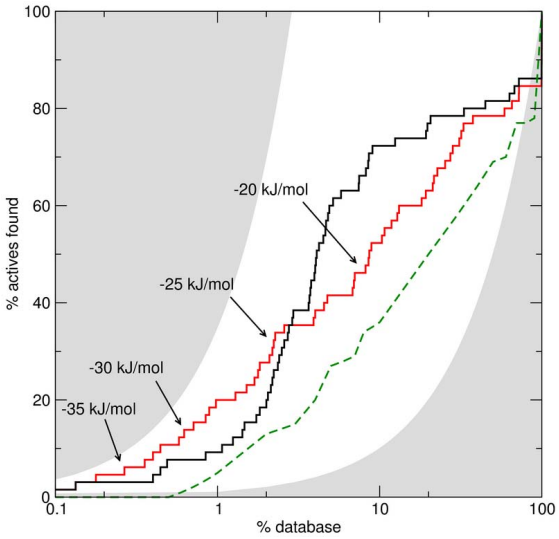


Abbildung 6.7: Berechnete Anreicherungs-Kurven für den Thrombin-Datensatz unter Verwendung einer alternativen PDB-Struktur.

PDB-Struktur 1ba9 [83] zu verwenden. Diese weist einige ungewöhnliche Geometrien in der Bindetasche auf, die auf eine kovalente Bindung des co-kristallisierten Liganden zurückzuführen sind, und es ist nicht möglich, diesen mit FlexX in seine ursprüngliche Position zu docken. Stattdessen wurde die Struktur 1k22 [33] ausgewählt, die hinsichtlich der strukturellen Integrität besser geeignet erscheint. Die Struktur des Protein-Inhibitor Komplexes konnte mit FlexX gut reproduziert werden und weist einen HYDE Score von -42 kJ mol^{-1} auf, was mit dem experimentell bestimmten Wert von ΔG = -48 kJ mol^{-1} [33] gut übereinstimmt.

Der verwendete Datensatz besteht aus 72 Thrombin-Inhibitoren und 2184 Decoy-Verbindungen. Beim Vergleich des Kurvenverlaufs in Abbildung 6.7 mit der entsprechenden Grafik im Anhang C fällt auf, dass es zu einem deutlich verbesserten Enrichment-Ergebnis bei Verwendung von HYDE kommt, während sich die FlexX-Kurven kaum unterscheiden. Neben der Auswahl eines geeigneteren Rezeptors spielt hier vor allem auch die sorgfältigere Überprüfung der Liganden eine Rolle. 13 Amidinobenzamidazol-Derivate lagen im DUD-Datensatz in einer falschen Protonierung vor. Die Amidino-Gruppe war durch eine Diammonium-Funktion ersetzt. Aufgrund der damit verbundenen sp^3-Hybridisierung der Stickstoff-Atome konnten die essentiellen Wechselwirkungen zum Thrombin nicht mehr reproduziert werden und es kam richtigerweise zu einer Abwertung dieser Verbindungen durch HYDE. Weiterhin auffällig war der Inhibitor `ZINC03834158`. Dieses Lacton-Derivat bindet nicht in der annotierten Form an Thrombin sondern erst nach einer Ringöffnung [24]. Zudem liegt auch bei diesem Molekül ein falscher Protonierungs-Zustand vor. Die Einstufung durch HYDE als nicht-bindend ist somit korrekt.

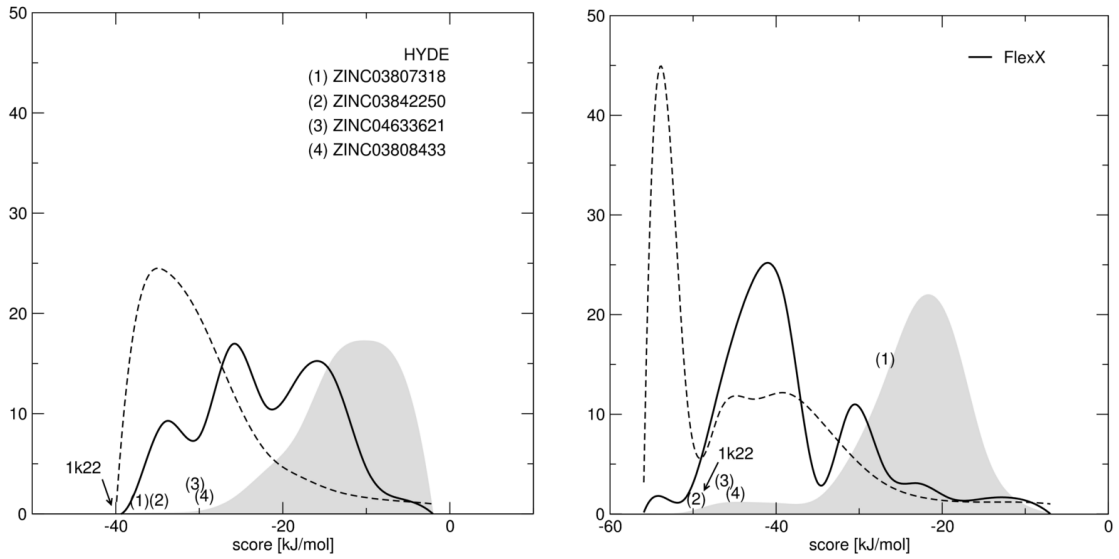
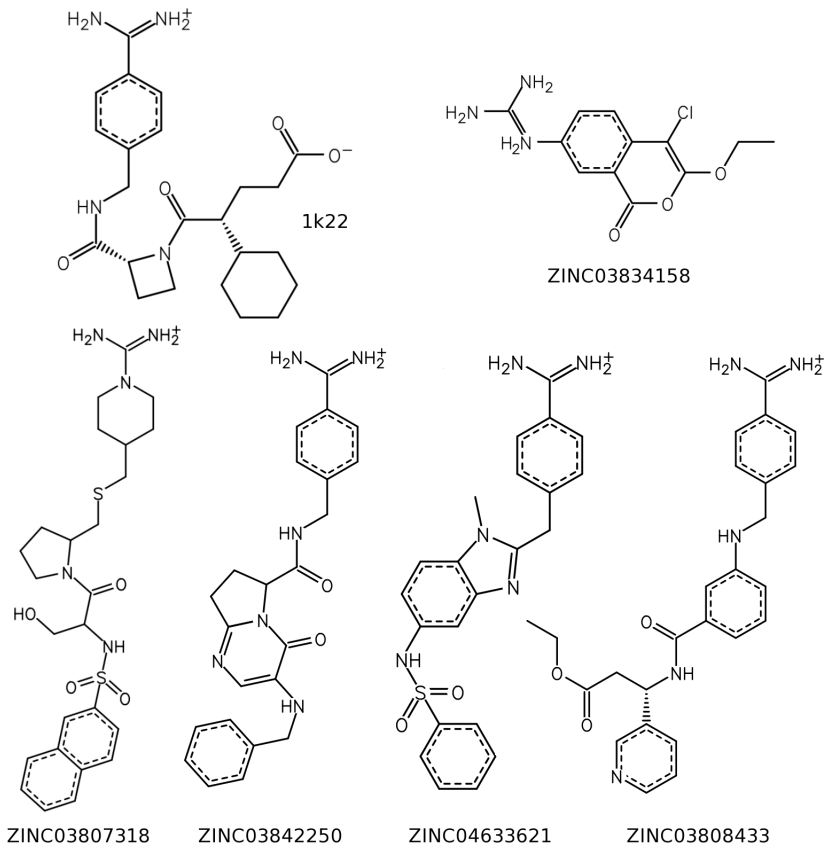


Abbildung 6.8.: Score-Distribution Thrombin (links: HYDE, rechts: FlexX).



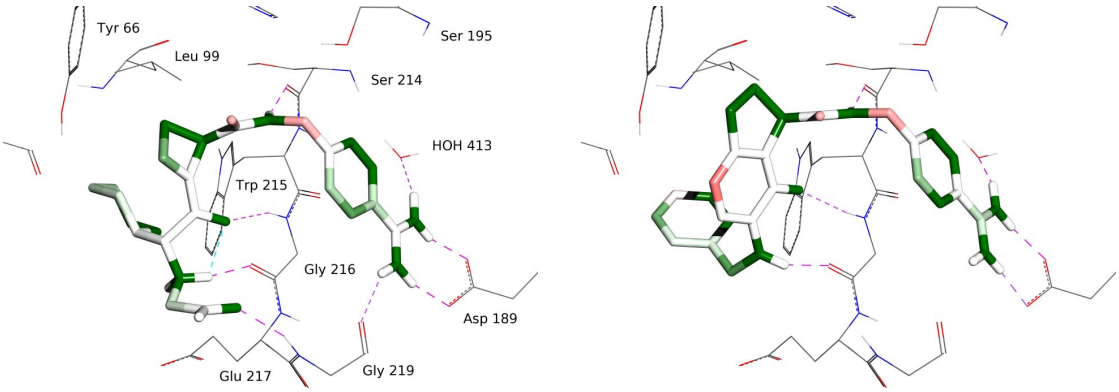


Abbildung 6.9.: Vergleich des co-kristallisierten Liganden in 1k22 (links) mit der gedockten Decoy-Verbindung ZINC03842250 (rechts). Einfärbung der Atome gemäß der Bewertung mit HYDE.

Die verminderte Anreicherungsrate ab einem HYDE-Score von etwas -25 kJ mol^{-1} (Abbildung 6.7) lässt sich durch schlecht bewertete Aktive erklären, die größtenteils auf die Protein-Aufbereitung bzw. den Docking-Algorithmus zurückzuführen sind. Zum Beispiel kommt aufgrund einer falschen Orientierung der Hydroxyl-Gruppe von `SER 195` eine Wasserstoff-Brücke mit dem Inhibitor `ZINC03831937` nicht zustande und dies wird mit etwa 10 kJ mol^{-1} durch HYDE bestraft. Außerdem werden einige Verbindungen so platziert, dass die Wechselwirkungen im Benzamidin-Fragment eine ideale Geomtrie haben, dafür aber andere Wasserstoff-Brücken-Funktionen ungesättigt bleiben bzw. nur schwache Wechselwirkungen ausbilden. Dies ist eine Folge des inkrementellen Docking-Algorithmus, der zunächst das stark bindende Benzamidin-Fragment möglichst ideal platziert und danach den Liganden fragmentweise in der Bindetasche rekonstruiert.

Zusätzlich zu diesen schlecht bewerteten Verbindungen gibt es auch eine Reihe von Decoy-Verbindungen, die einen sehr guten HYDE Score erzielen und somit das Enrichment-Ergebnis ebenfalls verschlechtern. Eine Literatur-Recherche ergab, dass viele dieser falsch Positiven zu patentierten Thrombin-Inhibitor-Klassen gehören. Alleine unter den zehn am besten bewerteten Decoys befinden sich sechs bekannte Inhibitoren (Tabelle 6.1). `ZINC03842250` und der co-kristallisierte Ligand aus 1k22 sind in Abbildung 6.9 dargestellt. Beide Verbindungen bilden das gleiche Wasserstoff-Brücken-Netzwerk zum Protein aus und die hydropbhoben Bereiche der Bindetasche werden von beiden Molekülen gut ausgefüllt.

Im Fall von Thrombin lässt sich der Einbruch in der Anreicherung-Kurve ab einem HYDE-Score von etwa -25 kJ mol^{-1} also durch mehrere zusammenwirkende Faktoren begründen. Der Datensatz mit den Inhibitoren enthält solche Verbindungen, die nicht

als aktiv einzustufen sind, und unter den inaktiven Verbindungen finden sich viele, für die eine Aktivität als Thrombin-Inhibitor nachgewiesen ist. Weiterhin konnten durch den Docking-Algorithmus nicht für alle Verbindungen die richtigen Konformationen in der Bindetasche erzeugt werden und eine dementsprechend schlechte Bewertung durch die Scoring-Funktion ist gerechtfertigt. Festzuhalten bleibt, dass durch Verwendung von HYDE diejenigen Moleküle ausgewählt werden konnten, für die eine Aktivität wahrscheinlich ist. Falsch negative Verbindungen sind im Sinne der Scoring-Funktion nicht falsch bewertet worden.

ZINC03807318	ZINC03807319	ZINC03842250
ZINC03838231	ZINC04633621	ZINC03808433

Tabelle 6.1.: Sechs patentierte Thrombin-Inhibitoren aus Ref. [54, 78, 124, 145], die unter den zehn am besten bewerteten Decoys zu finden sind.

6.1.5. p38 MAP Kinase

Die p38 Mitogen-aktivierten Proteinkinasen gehören zur Familie der Serin/Threonin-Proteinkinasen. Über die Phosphorylierung von spezifischen Proteinen bewirken sie als Teil einer Signalkaskade eine Reaktion der Zelle auf externe Stress-Signale wie Hitzeschock, Strahlung und osmotischen Schock. Sie haben darüber hinaus Bedeutung für Zelldifferenzierung, Zellwachstum und Apoptose. Für die pharmakologische Forschung stellen Kinasen ein sehr interessantes und gut erforschtes Target dar, da sie in vielen wichtigen biochemischen Signalwegen relevante Funktionen innehaben. Essentiell für die Funktion einer Kinase ist ein Adenosintriphosphat (ATP), welches durch Spaltung einer Phosphoanhydrid-Bindung zu Adenosindiphosphat (ADP) umgesetzt wird. Die abgespaltene Phosphat-Gruppe wird auf ein anderes Protein übertragen. Da dieses ATP-Molekül von allen Kinasen gebunden werden muss, ist die entsprechende Region strukturell stark konserviert und eignet sich gut als Bindestelle für einen Inhibitor. Durch ein gutes Wechselwirkungs-Profil mit den Amiden des Protein-Backbones in der sogenannten Hinge-Region ergeben sich hohe Bindungs-Affinitäten die zu einer Verdrängung des ATP führen. Selektivität bezüglich einer bestimmten Kinase lässt sich vor allem durch Komplementarität zu den benachbarten hydrophoben Regionen erzielen.

Der für dieses Target verwendete DUD-Datensatz besteht aus 440 Inhibitoren und 8387 Decoys. Eine erste Analyse der Enrichment-Resultate ergab keine erkennbare Anreicherung. Die HYDE-Kurve in Abbildung 6.10 ist nahezu deckungsgleich mit einer zufälligen Verteilung von Aktiven und Inaktiven. Die Verteilungs-Kurve der berechneten Bewertungen (Abbildung 6.11) zeigt, dass die Aktiven und die Decoys in demselben Werte-Bereich

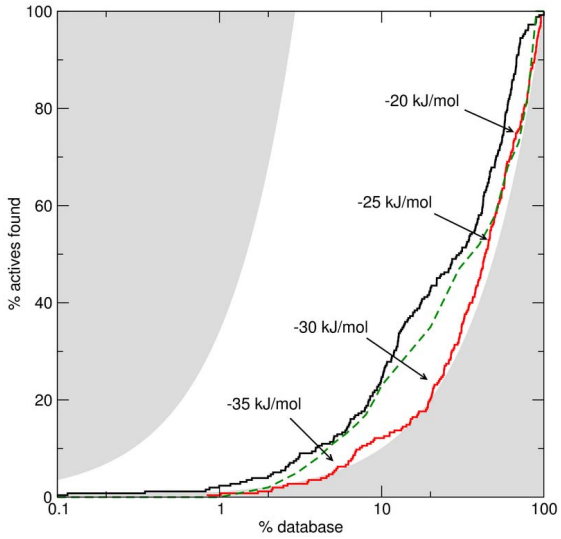


Abbildung 6.10: Enrichment-Plot für den Kinase Datensatz (1kv2).

liegen und somit eine Unterscheidung zwischen diesen Klassen nicht möglich ist.

Im Gegensatz zu den üblichen Kinase-Inhibitoren, bindet die im DUD-Datensatz verwendete Protein-Struktur 1kv2 den Inhibitor nicht direkt in der ATP-Bindetasche sondern in einer gewissen Entfernung dazu. Gemäß der Publikation von Pargellis *et al.* [107] bewirkt die Bildung des Komplexes mit dem Liganden BIRB796 eine konformationelle Änderung im Protein, ATP passt nicht mehr in die Bindetasche und die Kinase ist in ihrer Funktion inhibiert. Ein strukturelle Überlagerung mit einer anderen p38 MAP Kinase 1a9u [157] verdeutlicht das Ausmaß dieser konformationellen Änderung. Der Inhibitor SB203580 bindet in charakteristischer Weise in die ATP-Bindetasche, dargestellt durch die graue Struktur in Abbildung 6.12. Eine Schleife der in rot dargestellte DUD-Struktur ist in der ATP-Region um bis zu 5Å verschoben. Eine Bindung ist an dieser Stelle aus sterischen Gründen nicht mehr möglich.

Eine Analyse der p38 MAP Kinase-Inhibitoren im DUD-Datensatz ergab zunächst, dass unter den 440 Verbindungen 185 enthalten sind, die sich von den anderen nur in der Protonierung oder der genauen Lage der Doppelbindungen unterscheiden. Es ist unwahrscheinlich, dass ein Ligand in unterschiedlichen Protonierungszuständen oder tautomeren Formen an dasselbe Protein bindet. Dies kann aber nicht der alleinige Grund für das unbefriedigende Enrichment sein, da die verwendete Aufbereitungs-Prozedur die Generierung von solchen Formen beinhaltet und nur die am besten bewertete Struktur berücksichtigt wurde. Desweiteren enthält der Datensatz jedoch drei spezifische Molekül-Klassen, die sich in ihrer Struktur stark unterscheiden (Abbildung 6.14). Den größten Anteil hat die Klasse der Triarylpyrrol-Verbindungen (TA), die mit insgesamt 189 Strukturen im berei-

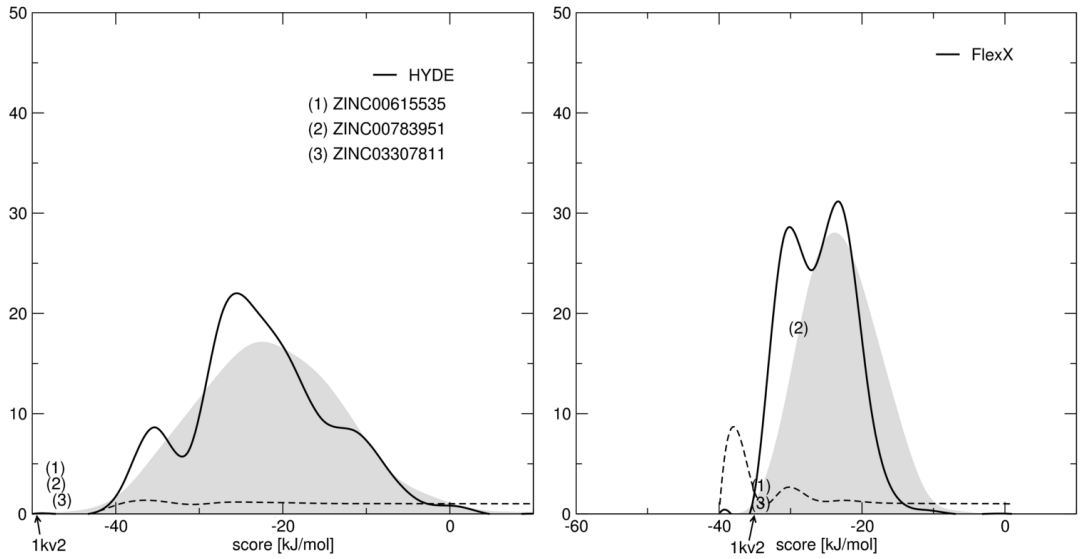
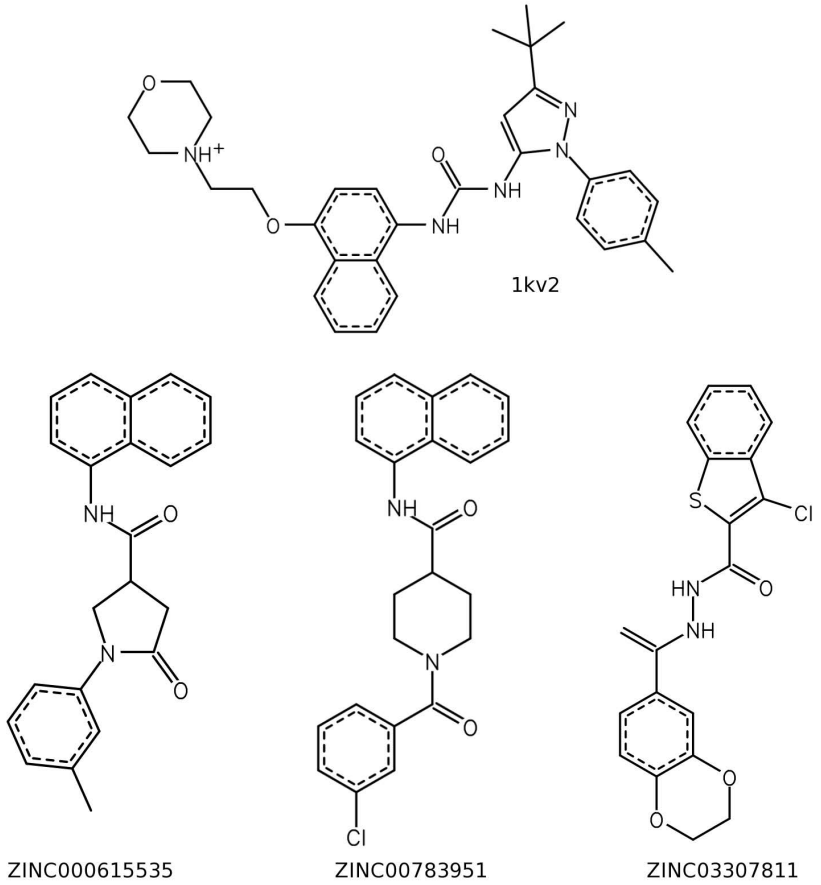


Abbildung 6.11.: Score-Verteilungs-Kurven p38 MAP Kinase.



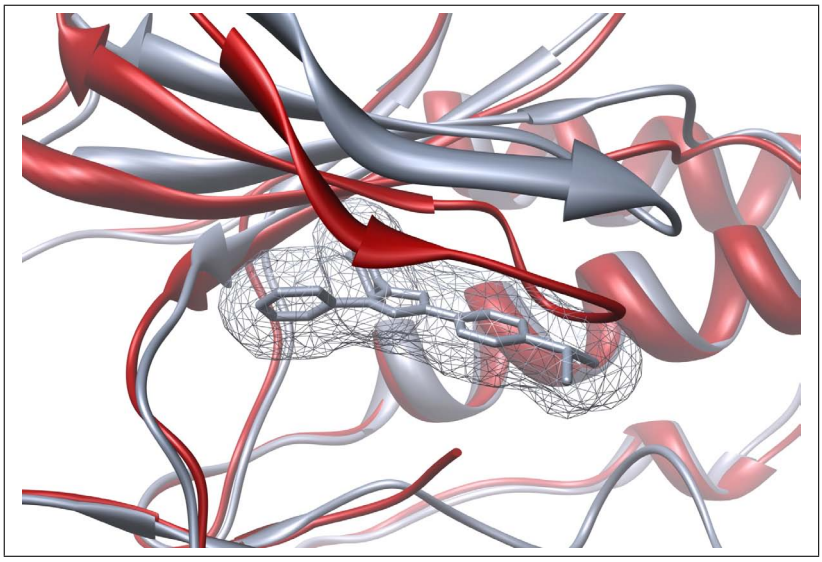


Abbildung 6.12: Überlagerung der PDB Strukturen 1a9u (grau) und 1kv2 (rot). Eine konformationelle Änderung in der ATP-Region von 1kv2 führt zu einer sterischen Überlappung mit dem Liganden von 1a9u.

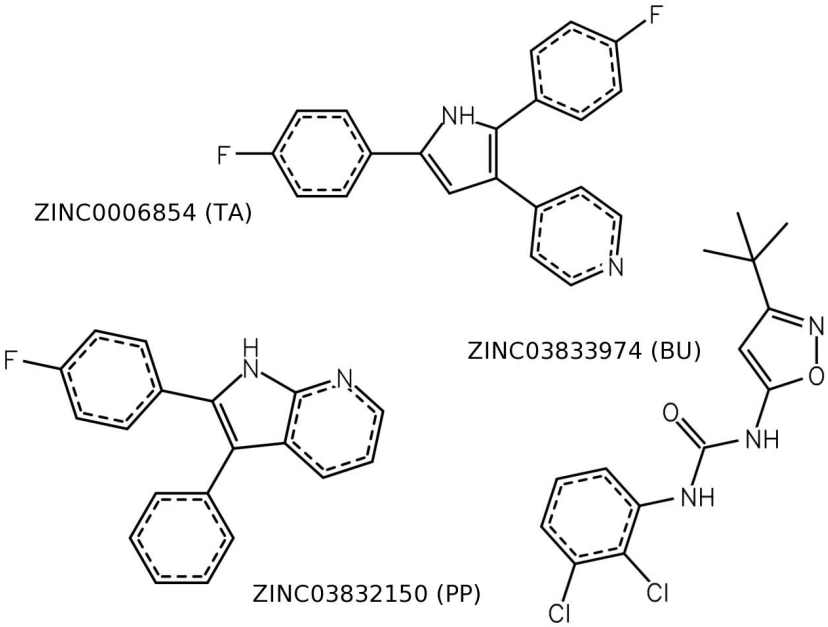


Abbildung 6.13.: Beispiele für die im p38 MAP Datensatz enthaltenen Sub-Klassen.

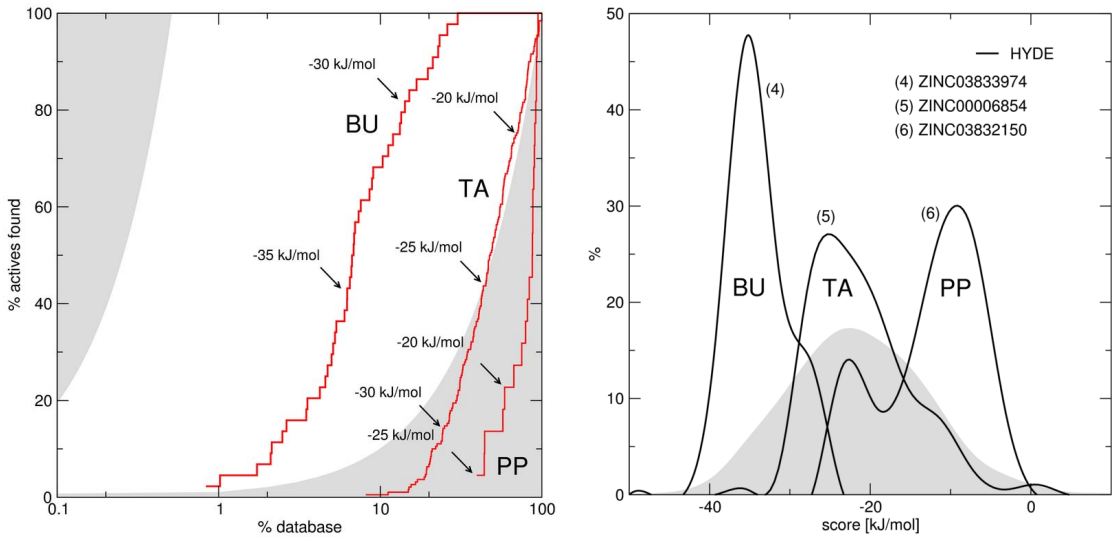


Abbildung 6.14.: Enrichment im p38 Datensatz bezüglich verschiedener Molekülklassen. Dargestellt sind individuelle Anreicherungen für Biarylharnstoffe (BU) in rot, Triarylpyrrol-Verbindungen (TA) in grün und die Pyrrolopyridine (PP) in blau.

nigten Datensatz repräsentiert sind. Der in Abbildung 6.12 gezeigte Inhibitor SB203580 ist ein typisches Beispiel für diese Klasse und es ist bekannt, dass solche Strukturen in der ATP-Region binden [148, 32]. Auch für die zweite Klasse der 22 Pyrrolopyridine (PP) wurde eine Bindung in der ATP-Region nachgewiesen [7]. Übrig bleiben 44 Verbindungen, die als Biarylharnstoff-Derivate (BU) einzustufen sind und eine ähnliche Struktur wie der Original-Inhibitor aufweisen. Bedingt durch die konformationelle Änderung des Proteins ist anzunehmen, dass nur die letztgenannte Klasse in der Lage ist, vernünftige Wechselwirkungen zu der DUD-Proteinstruktur auszubilden. Das Enrichment-Experiment wurde mit dem bereinigten Datensatz wiederholt und es wurden jeweils nur die Verbindungen einer Molekül-Klasse als aktiv eingestuft.

Abbildung 6.14 zeigt eine deutliche Anreicherung der BUs. Alle Verbindungen dieser Klasse erhalten einen Score besser als -25 kJ mol^{-1}, der die BUs als Binder klassifiziert. Im Gegensatz dazu werden die PPs durch HYDE sehr schlecht bewertet, vor allem die Dehydratation von polaren Atomen führt zu einem schlechten Score, der nicht durch den hydrophoben Effekt ausgeglichen werden kann. Dazwischen liegen die TA Verbindungen, für die eine unspezifische Bindung vorhergesagt wird. Diese Verbindungen können teilweise die hydrophoben Bereiche der Protein-Struktur ausfüllen, durch das Fehlen von Wasserstoff-Brücken-Funktionen scheint eine spezifische Bindung in dieser Bindetasche unwahrscheinlich.

Weiterhin auffällig ist die hohe Bewertung, die einige der Decoy-Verbindungen erzielen. Eine Literatur-Recherche hat ergeben, dass viele dieser Verbindungen zu Klassen gehören, die als Kinase-Inhibitoren patentiert sind [77, 50]. Da die beim Docking verwendete allosterische Bindetasche viele über unterschiedliche Protein-Klassen konservierte Aminosäuren enthält, ist es wahrscheinlich, dass diese Moleküle auch Inhibitoren der p38 MAP Kinase sind.

ZINC00615535	ZINC01072327	ZINC00788124	ZINC00290115
ZINC00615522	ZINC00432687	ZINC00843792	ZINC01103953
ZINC00035049	ZINC03307811	ZINC01780257	ZINC02748340
ZINC03296097	ZINC02751904	ZINC00124860	ZINC03855198
ZINC03284746	ZINC03852818	ZINC02761165	ZINC02754843

Tabelle 6.2.: Im p38 Decoy-Datensatz enthaltene Verbindungen, die als Kinase-Inhibitoren patentierten Struktur-Klassen angehören [77, 50].

6.2. Vorhersage von Bindungs-Affinitäten

Neben der korrekten Einstufung einer chemischen Verbindung als Binder oder Nicht-Binder ist die quantitative Vorhersage von Bindungs-Affinitäten ein häufig verwendetes Kriterium bei der Validierung von Scoring-Funktionen. Der PDBbind Datensatz [150, 151] ist eine Sammlung von PDB-Komplexen mit experimentell bestimmten Bindungs-Affinitäten. Bei der Auswahl der Strukturen wurde auf gewisse Mindest-Anforderungen wie Auflösung, Bindungsmodi oder Vorhandensein von Co-Faktoren geachtet. Der negative dekadische Logarithmus $-logK_D$ der Dissoziations-Konstanten der Komplexe nimmt einen großen Wertebereich von 0.60 bis 13.96 ein.

In einer umfangreichen Studie von Wang *et al.* [155] wurden 14 aktuelle Scoring-Funktionen anhand dieses Datensatzes miteinander verglichen. In einer Regressionanalyse wurden die experimentell bestimmten Werte y_i in einen linearen Zusammenhang zu den Ergebnissen einer Scoring-Funktion x_i gesetzt.

$$y_i = ax_i + b \tag{6.2}$$

Als Maß für die Qualität der Scoring-Funktion wurden der Korrelations-Koeffizient (nach Pearson) R_p, die Standardabweichung SD und der mittlere Fehler ME berechnet:

$$R_p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{6.3}$$

$$SD = \sqrt{\sum_i [y_i - (ax_i + b)]^2 / (N-2)} \tag{6.4}$$

$$ME = \sum_i \mid y_i - (ax_i + b) \mid / N \tag{6.5}$$

In Tabelle 6.3 sind die Ergebnisse dieser Studie abgebildet. N bezeichnet die Anzahl der in der Regression verwendeten Protein-Ligand Komplexe. Da einige Scoring-Funktionen sterische Überlappungen zwischen Protein und Ligand bestrafen und dies zu einer Verfälschung der Ergebnisse führen würde, wurden jeweils nur solche Komplexe in der Regressions-Analyse berücksichtigt, für die eine Bindungs-Konstante $-logK_D$ größer Null berechnet wurde. Die besten der in dieser Studie betrachteten Scoring-Funktionen erreichen bezüglich des experimentellen pK_D Wertes einen Korrelations-Koeffizienten von maximal 0.5 bei einer Standardabweichung von 2 $\log K_D$ Einheiten.

Zur Validierung von HYDE wurde dieses Experiment mit unterschiedlichen Funktions-Parametern wiederholt. Dazu wurde in FlexX die Bindetasche der Proteine durch einen Abstand von 8Å zu den Atomen des jeweiligen Referenz-Liganden definiert. Für die Ausrichtung der frei rotierbaren Hydroxyl-Gruppen des Rezeptors sowie dessen Protonierung wurden Standard-Parameter verwendet. Die Protonierung des Liganden wurde aus der PDBbind übernommen. Bewertet wurde die Konformation des Referenz-Liganden, dabei erfolgte eine automatische Ausrichtung der rotierbaren Hydroxyl-Gruppen des Liganden unter Optimierung der verwendeten Scoring-Funktion. Die Ergebnisse dieser Bewertung sind in Tabelle 6.4 aufgelistet.

Der Funktions-Parameter *hyde_ access_ mode* bezeichnet den verwendeten Algorithmus für die Berechnung der Zugänglichkeit. Mit *hyde_ desolvation_ mode* ist die genaue Formel zur Berechnung der Desolvatations-Energie angegeben. Modus 3 entspricht Gleichung 3.3 mit der Normierung der Zugänglichkeitsänderung auf die mittlere Zugänglichkeit $acc(mean)$ eines Atom-Typs. Bei Wahl von Modus 4 wird stattdessen die Zugänglichkeit des jeweiligen Atoms im ungebundenen Zustand $acc(free)$ im Nenner verwendet. Weiterhin wurde die Formel für die Berechnung der Wechselwirkungs-Energie variiert (*hyde_ sat_ mode*). Neben der Verwendung des Wurzel-Terms *dev_ root* (Gl. 3.8) wurde die Wechselwirkungs-Energie auch anhand der Änderung in der Zugänglichkeit berechnet. Dabei wurde ein erweiterter Radius von 1.2Å sowie eine Berücksichtigung von Atomen über 2 Bindungen hinweg verwendet (vgl. Abschnitt 4.3.2). Zum Vergleich

Scoring Funktion	N	R_p	SD	ME	a	b
X-Score::HPScore	800	0.514	1.89	1.47	0.71	2.03
X-Score::HMScore	800	0.566	1.82	1.42	0.92	1.18
X-Score::HSScore	800	0.506	1.90	1.48	0.93	1.24
DrugScore::Pair	800	0.473	1.94	1.51	4.9×10^{-6}	4.10
DrugScore::Surf	800	0.463	1.95	1.53	7.2×10^{-5}	4.48
DrugScore::Pair/Surf	800	0.476	1.94	1.50	4.7×10^{-6}	4.09
Sybyl::D-Score	800	0.322	2.09	1.67	9.7×10^{-3}	5.00
Sybyl::PMF-Score	785	0.147	2.16	1.74	6.4×10^{-3}	5.92
Sybyl::G-Score	800	0.443	1.98	1.56	9.1×10^{-3}	4.34
Sybyl::ChemScore	797	0.499	1.91	1.50	9.1×10^{-2}	3.90
Sybyl::F-Score	732	0.141	2.19	1.77	2.1×10^{-2}	6.06
Cerius2::LigScore	717	0.406	2.00	1.57	0.79	4.63
Cerius2::PLP1	800	0.458	1.96	1.52	2.3×10^{-2}	4.09
Cerius2::PLP2	800	0.455	1.96	1.53	2.6×10^{-2}	3.93
Cerius2::PMF	795	0.253	2.13	1.71	1.1×10^{-2}	5.37
Cerius2::LUDI1	790	0.334	2.08	1.66	2.6×10^{-3}	4.88
Cerius2::LUDI2	799	0.379	2.04	1.62	4.2×10^{-3}	4.28
Cerius2::LUDI3	800	0.331	2.08	1.67	3.2×10^{-3}	4.68
GOLD::GoldScore	694	0.285	2.16	1.72	2.4×10^{-2}	5.33
GOLD::GoldScore_opt	772	0.365	2.06	1.63	3.0×10^{-2}	4.70
GOLD::ChemScore	741	0.423	2.00	1.56	8.5×10^{-2}	4.65
GOLD::ChemScore_opt	762	0.449	1.96	1.52	8.6×10^{-2}	4.41
HINT	800	0.330	2.08	1.65	0.20	6.36

Tabelle 6.3.: Korrelations-Parameter von 14 Scoring-Funktionen aus Wang *et al.* [155].

HYDE Parameter[a]	N	R_p	SD	ME	a	b
WSAS/3/dev_root	800	0.138	2.18	1.76	0.08	6.17
WSAS/4/dev_root	800	0.368	2.05	1.64	0.14	5.93
WRAY/3/dev_root	800	0.095	2.20	1.77	0.06	6.21
WRAY/4/dev_root	800	0.184	2.17	1.75	0.11	6.03
SWSA/3/dev_root	800	0.064	2.20	1.78	0.03	6.33
SWSA/4/dev_root	800	0.214	2.15	1.74	0.09	6.19
WSAS/3/acc	800	0.207	2.16	1.73	0.14	6.22
WSAS/4/acc	800	0.372	2.05	1.63	0.13	5.90
WRAY/3/acc	800	0.176	2.17	1.75	0.13	6.10
WRAY/4/acc	800	0.222	2.15	1.73	0.14	6.01
SWSA/3/acc	800	0.120	2.19	1.77	0.07	6.33
SWSA/4/acc	800	0.247	2.14	1.73	0.10	6.18
F-Score	800	0.151	2.18	1.76	0.10	6.04

Tabelle 6.4.: HYDE-Ergebnisse, alle 800 PDBbind-Komplexe. [a]Parameter: *hyde_access_mode / hyde_desolvation_mode / hyde_sat_mode.*

sind auch Ergebnisse der FlexX-Scoring Funktion in der Tabelle aufgeführt. Abweichend von den Original-Einstellungen wurden die für HYDE umdefinierten Wechselwirkungs-Geometrien verwendet.

Gegenüber der FlexX Scoring-Funktion wurden mit HYDE deutlich verbesserte Resultate erzielt. Jedoch liegen die Korrelations-Parameter ebenso deutlich unter denen der besten Scoring-Funktionen in der Wang-Studie. Dazu sei aber angemerkt, dass zumindest die X-Score Funktion, die in der Studie am besten abschneidet, anhand der Daten aus der PDBbind Datenbank kalibriert worden ist. Relativiert wird das Abschneiden der Scoring-Funktionen in dieser Studie auch dadurch, dass rein additive Funktionen üblicherweise stark mit der Größe des bewerteten Moleküls korrelieren und der Korrelationskoeffizient für den einfachen Zusammenhang zwischen Affinität und dem Molekulargewicht der betrachteten Liganden mit 0.4 in dem gleichen Wertebereich liegt. Ein einfaches Abzählen der Atome, welches in diesem Fall also zu einer vergleichbaren Korrelation führen würde, kann jedoch nicht ausreichend sein, um die Wechselwirkungen zwischen Protein und Ligand ausreichend zu quantifizieren. Dies funktioniert nur bei Komplexen mit einem hinreichend abgesättigten Wechselwirkungs-Profil. Es wäre jedoch nicht möglich zwischen verschiedenen Konformationen eines Liganden in der Bindetasche zu diskriminieren, und auch die spezifische Zusammensetzung (zum Beispiel der Austausch einer Hydroxyl- gegen eine Methyl-Gruppe) eines Liganden hätte keine Auswirkung auf den berechneten Score.

Bei der genaueren Analyse der HYDE-Ergebnisse überrascht das durchgängig bessere Korrelations-Verhalten (bei ansonsten gleichen Parametern) des $hyde\_access\_mode$ 4 im Vergleich zu dem Original-Term der Scoring-Funktion (Modus 3). Durch die Verwendung von $acc(free)$ im Nenner des Desolvatations-Terms wird nicht mehr zwischen exponierten und unzugänglichen Atomen unterschieden. Alle Atome können einen Beitrag entsprechend des plogP-Wertes leisten, sofern sie bei der Komplex-Bildung vollständig desolvatisiert werden. Insbesondere wirkt sich dies auf Protein-Atome aus, deren Zugänglichkeit erheblich vom $acc(mean)$ Wert abweichen kann. Bei solchen Atomen führt die Verwendung der mittleren Zugänglichkeit gerade in engen Bindetaschen mit vielen Wasserstoff-Brücken-Funktionen zu einer Überbewertung der Energie-Beiträge, da dem Wechselwirkungs-Beitrag quasi keine Bestrafung durch die Desolvatation gegenübersteht.

Deutlich bemerkbar macht sich dieser Umstand auch bei Metall-Ionen im aktiven Zentrum des Proteins. Zusätzlich zu der durch den kleinen vdW-Radius des Atoms bedingten sehr geringen Zugänglichkeit ist bei Metallen eine große Varianz des Koordinierungs-Zustandes gegeben. Die Anzahl der durch das Protein besetzten Koordinierungs-Stellen varriert ebenso wie die Koordinierungs-Geometrie [133]. Dadurch lässt die Zugänglichkeit von Metallen nicht mehr durch eine Normalverteilung beschreiben und die Definition einer mittleren Zugänglichkeit führt zwangsweise zu Fehlern bei der Energie-Berechnung. Für ein besseres Verständnis der Abweichungen im Korrelations-Verhalten wurden die Regressions-Parameter ein zweites Mal bestimmt (Tabelle 6.5). Diesmal wurden nur solche Komplexe berücksichtigt, in denen keine Metall-Ionen im aktiven Zentrum vorkommen. Durch diese Maßnahme ergeben sich deutlich verbesserte Werte, die generellen Tendenzen zwischen den verschiedenen HYDE-Parametern bleiben aber bestehen.

Auffällig ist weiterhin die verbesserte Korrelation zwischen berechneter und experimenteller Affinität bei Verwendung der Zugänglichkeit acc im Wechselwirkungs-Term. Dies wurde diskutiert als eine Maßnahme, den ungewissen Referenz-Zustand für eine funktionelle Gruppe zu definieren, welcher angibt, wie viele Wechselwirkungen ein Atom machen kann. Tatsächlich wird dadurch auch die Überbewertung sehr vergrabener Atome abgemildert, da der systematische Fehler bei der unzureichenden Berechnung der Zugänglichkeit in unzugänglichen Regionen durch eine wiederholte Verwendung dieses Ausdrucks abgeschwächt wird.

Betrachtet man die Auftragung der mit HYDE berechneten Werte gegen die experimentellen Affinitäten in Abbildung 6.15 (umgerechnet in kJ/mol), ist klar ersichtlich, dass die (statistisch) schwache Korrelation vor allem mit der zu schlechten Bewertung von vielen Komplexen zusammenhängt, was zu einer breiten Streuung der Werte führt. Im vorangegangenen Kapitel wurde bereits diskutiert, dass die Qualität der Bewertung eines Protein-Ligand Komplexes unmittelbar von der Qualität der zugrunde liegenden

HYDE Parameter	N	R_p	SD	ME	a	b
WSAS/3/dev_root	594	0.246	2.09	1.68	0.15	5.94
WSAS/4/dev_root	594	0.444	1.94	1.54	0.16	5.77
WRAY/3/dev_root	594	0.218	2.11	1.71	0.14	5.86
WRAY/4/dev_root	594	0.297	2.06	1.68	0.18	5.73
SWSA/3/dev_root	594	0.205	2.11	1.71	0.10	5.99
SWSA/4/dev_root	594	0.300	2.06	1.68	0.12	6.03
WSAS/3/acc	594	0.285	2.07	1.65	0.20	6.10
WSAS/4/acc	594	0.453	1.93	1.52	0.16	5.74
WRAY/3/acc	594	0.296	2.06	1.67	0.23	5.81
WRAY/4/acc	594	0.352	2.02	1.63	0.22	5.71
SWSA/3/acc	594	0.242	2.10	1.70	0.14	6.15
SWSA/4/acc	594	0.348	2.03	1.64	0.14	6.01
F-Score	594	0.202	2.12	1.71	0.14	5.93

Tabelle 6.5.: HYDE-Ergebnisse. Nur Komplexe ohne Metall-Zentrum in der Bindetasche.

Daten abhängt. Während die für das Enrichment berechneten Konformationen des Liganden immerhin mit demselben Wechselwirkungs-Modell erzeugt worden sind, das auch bei der Bewertung mit HYDE verwendet wird, kann ein PDB-Komplex mit beliebig abweichenden Geometrien optimiert worden sein, je nach verwendetem Kraftfeld. Dies führt dazu, dass Wechselwirkungen nicht erkannt werden und es zu starken (und im Sinne der Scoring-Funktion korrekten) Straf-Beiträgen für die Desolvatisierung polarer Atome kommt. Neben der schwankenden Qualität der verwendeten Daten kommt zusätzlich die Frage der Daten-Aufbereitung ins Spiel. Wie viel Wissen darf man bei einer Vergleichs-Studie in das System stecken, ohne den Aspekt der Objektivität und damit der Vergleichbarkeit zu verlieren?

Als ein extremes Beispiel für eine schlecht bewertete Verbindung ist der PDB Komplex 1bjr [137] in der bekannten Farbcodierung angeführt (Abbildung 6.16). Es handelt sich um einen Komplex der Proteinkinase K mit einem durch Proteolyse erzeugten Lactoferrin-Fragment. Dieses Dekapeptid inhibiert die Proteinkinase K, die Gleichgewichtskonstante ist mit $2.1 \cdot 10^{-8}$ angegeben [137]. Das als Wasserstoff-Brücken-Mediator fungierende Wasser `WAT 554` ist in der PDBbind-Struktur nicht enthalten. Weiterhin geben die Abstände zwischen Protein- und Ligand-Atomen, die teilweise geringer als 1.5Å sind (und damit eher im Bereich von kovalenten Bindungen liegen) deutliche Hinweise darauf, dass bei der Aufklärung der Struktur Fehler gemacht worden sind. Es werden keine Wasserstoff-Brücken zwischen Protein und Ligand erkannt. Ähnliches ist bei vielen anderen Komplexen zu beobachten. HYDE identifiziert fehlerhafte Strukturen mit einem schlechten Score.

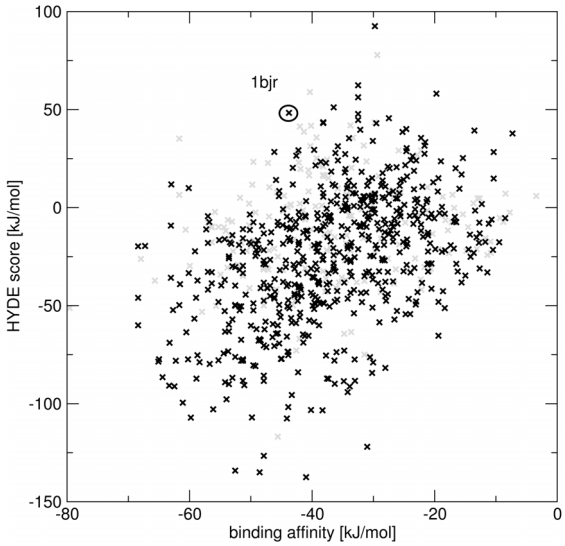


Abbildung 6.15.: Berechnete HYDE-Affinitäten im Vergleich mit den experimentell bestimmten Werten. Aufgetragen sind die Ergebnisse für die HYDE Parameter WSAS/4/dev_root. Die in grau dargestellten Verbindungen enthalten ein Metall-Zentrum.

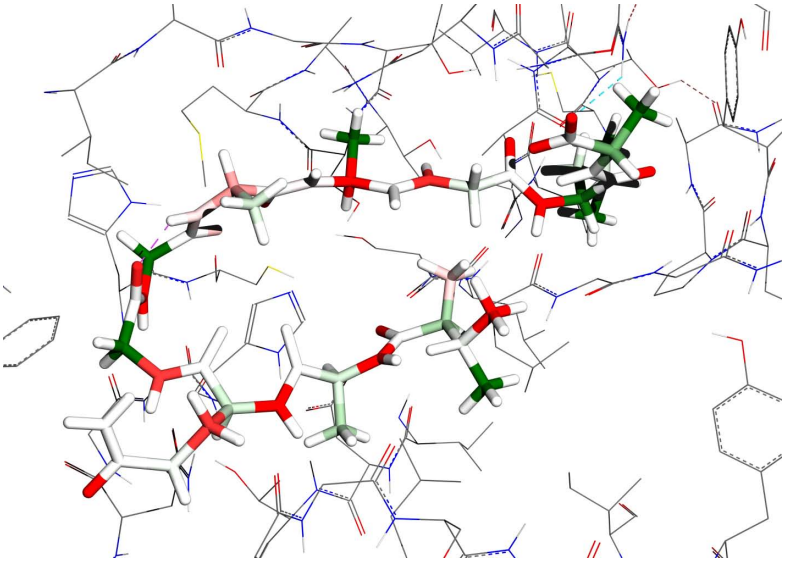


Abbildung 6.16.: Farbcodierte Darstellung von 1bjr[137].

6.3. Virtuelles Screening mit experimenteller Validierung

Da die aufgeführten theoretischen Validierungs-Methoden eine große Zahl von Nachteilen und Einschränkungen mit sich bringen, ist es von besonderem Interesse zu prüfen, ob HYDE in einem experimentellen Szenario einsetzbar ist und dort gute Resultate erzielt. In einer ersten Studie von Bayer CropScience wurde ein virtuelles Screening an ein aktuellen Target durchgeführt. Für dieses Target lag eine Kristall-Struktur von guter Qualität vor, die Protonierung und Ausrichtung von rotierbaren Gruppen erfolgte unter großer Sorgfalt manuell.

Mit FlexX wurden unter Verwendung der Original Scoring-Funktion 4.5 Millionen Verbindungen einer *in house* Datenbank gedockt. Die höchstbewerteten Posen jeder Verbindung wurden gespeichert und in einem zweiten Schritt mit HYDE neu bewertet. In diesem Bewertungs-Schritt erhielten nur 2200 Verbindungen einen Score, der auf eine mögliche Affinität zum betrachteten Rezeptor schließen ließ. Unter verschiedenen Gesichtspunkten wie Verfügbarkeit und ADME-Kriterien wurden von diesen Verbindungen 172 ausgewählt und in einem experimentellen Setup bezüglich einer Inhibition des Proteins untersucht. Dabei konnten für zwei Drittel der untersuchten Verbindungen Inhibitions-Konstanten im mikro- sowie im nanomolaren Bereich nachgewiesen werden. Dies ist eine bemerkenswert hohe Treffer-Rate, die nur in wenigen veröffentlichten Studien über erfolgreiche Anwendungen von virtuellen Screening-Methoden erreicht wird [85].

pIC_{50}	*Anzahl*	*Prozent*
alle	172	100%
≥ 4.3	116	67%
≥ 5.5	76	44%
≥ 6.5	33	19%
≥ 7.5	20	12%

Tabelle 6.6.: Inhibitions-Konstanten der im virtuellen Screening identifizierten Verbindungen.

6.4. Diskussion der Ergebnisse

Bewertet man die Ergebnisse aus Kapitel 6 nur anhand der produzierten Kennzahlen, stellt sich die Frage, warum HYDE eine Verbesserung gegenüber existierenden Methoden darstellen soll. Bei einer wesentlich längeren Rechenzeit sind bei den durchgeführten Studien keine verbesserten Resultate im Vergleich zu anderen Methoden zu erzielen. Wie

jedoch mehrfach erwähnt, können solche Studien kein endgültiger Nachweis der Korrektheit des verwendeten Modells sein, weder im positiven Sinne noch im negativen. Zu viele Faktoren wirken auf das Ergebnis ein. Selbst extra für den Zweck des Vergleichs unterschiedlicher Methoden publizierte Datensätze wie der DUD [68] weisen zu viele Fehler auf, als dass ein aussagekräftiger Vergleich noch gegeben wäre. Die zu beantwortende Frage ist: Macht die Scoring-Funktion das, was von ihr erwartet wird, und wie sind abweichende Ergebnisse zu erklären?

Dazu muss man sich die Zielsetzung bei der Entwicklung der neuen Scoring-Funktion in Erinnerung rufen. Durch die ganzheitliche Modellierung von Desolvatations-Effekten und die explizite Berücksichtigung von destabilisierenden Beiträgen sollte die Rate an falsch positiven Hits im virtuellen Screening deutlich vermindert werden. Dass die Desolvatation von polaren Atomen sich zunächst ungünstig auf die Bindungs-Affinität auswirkt, ist hinreichend bekannt. Gleiches gilt für den großen Einfluss des Lösungsmittels auf die Komplexbildung. Dennoch ist eine entsprechende Umsetzung in einer Scoring-Funktion bisher nur in Form von Hilfstermen gegeben, die eher eine Art Filter-Funktion darstellen als eine korrekte Beschreibung der freien Energie (vgl. Abschnitt 2.4.3). Dieser Widerspruch zwischen Theorie und Modell scheint vor allem darin begründet, dass ein objektiver Vergleich von unterschiedlichen Methoden schwierig ist und sich deshalb oft auf einfach berechen- und darstellbare Zahlen verlassen wird. Zahlen, welche umso schlechter erscheinen, je intoleranter die verwendete Scoring-Funktion sich gegenüber Ungenauigkeiten in den zu bewertenden Strukturen verhält.

Eine Scoring-Funktion, die ungünstige Konstellationen im Protein-Ligand Komplex bestrafen soll, muss zwangsläufig auch solche Konstellationen bestrafen, die aufgrund von Fehlern in den zugrunde liegenden Daten und schlechter Aufbereitung als ungünstig erscheinen. Wenn ein Abstand zwischen einem Wasserstoff-Brücken-Donor und einem Akzeptor zu groß ist, dann ist es nicht Schuld der Scoring-Funktion, wenn dies zu einer Strafe führt. Eine größere Toleranz bezüglich des Abstandes würde zwangsläufig zu einer vermehrten Rate an falsch positiven Ergebnissen führen, was in Diskrepanz zu dem Ziel bei der Entwicklung der Scoring-Funktion steht.

Der entscheidende Unterschied zwischen HYDE und herkömmlichen Scoring-Funktionen ist die Verlässlichkeit des berechneten Scores in dem Sinne, dass nur solche Liganden eine gute Bewertung erzielen, die tatsächlich als Inhibitoren in Frage kommen. Eine Korrelation des Scores mit der Größe des Liganden bzw. die Möglichkeit eine gute Bewertung zu erhalten ist nur dann gegeben, wenn es zu keiner Bestrafung von einzelnen Atomen kommt, also das Wasserstoff-Brücken-System zwischen Ligand und Rezeptor komplementär ist. Eine schlechte Bewertung einer aktiven Verbindung ist auch mit HYDE nicht ausgeschlossen, wenn durch den Docking-Algorithmus nicht die korrekte Lösung gefunden werden konnte. Dieses kann vielfältige Gründ haben, alleine der Umstand, dass die

Flexibilität des Rezeptors üblicherweise nicht in Betracht gezogen wird, kann schon zu der entscheidenden Diskrepanz in einem Winkel oder Abstand führen.

Wie im vorangegangenen Kapitel beschrieben wurde, konnte bei einer genaueren Untersuchung der Enrichment-Ergebnisse in allen betrachteten Fällen gezeigt werden, dass HYDE in korrekter Weise funktioniert. Am Beispiel des Estrogen-Rezeptors konnte gezeigt werden, welche grundlegende Bedeutung die Berücksichtigung der Desolvatation polarer Funktionen für eine korrekte Bewertung von Protein-Ligand Komplexen hat. Während durch die FlexX Scoring-Funktion (als typischer Repräsentant einer empirischen Funktion ohne Strafterme) viele Moleküle mit ungesättigten Wasserstoff-Brücken-Funktionen eine gute Bewertung erhielten, wurden diese Moleküle durch die Bewertung mit HYDE als falsch positive Hits identifiziert. Im Fall der Thrombin-Inhibitoren ließ sich das zunächst unbefriedigende Ergebnis durch eine fehlerhafte Protein-Struktur und viele inkorrekt protonierte Liganden erklären. Zudem konnte für eigentlich als aktiv klassifizierte Moleküle keine korrekte Pose in der Bindetasche erzeugt werden, so dass eine Bewertung als Nicht-Binder korrekt war. Abschließend wurde am Beispiel der p38 MAP Kinase deutlich, dass selbst die zufällige Verteilung der als aktiv gekennzeichneten Verbindungen im sortierten Datensatz nicht unbegründet war. Durch HYDE konnten drei Klassen von Inhibitoren identifiziert werden, von denen nur eine kompatibel zur verwendeten Protein-Struktur war. Die entsprechenden Moleküle wurden mit HYDE erkannt und erzielten eine gute Bewertung.

Die Verlässlichkeit eines guten HYDE-Scores zeigt sich unter anderem darin, dass für alle untersuchten Beispiele unter den mit einem guten Score versehenen Decoy-Verbindungen zahlreiche Moleküle zu finden sind, die als Inhibitor patentierten Verbindungs-Klassen angehören. Die Generierung der Decoys im DUD Datensatz, die eigentlich verhindern sollte, dass es zu einer artifiziellen Anreicherung aufgrund von Unterschieden in den physikalisch-chemischen Eigenschaften kommt, hat offenbar auch viele aktive Verbindungen reproduziert.

Allen Beispielen gemeinsam ist trotz der unterschiedlichen Charakteristika der Bindetaschen ein Grenzwert von ca. -25 kJ mol^{-1}, über dem bei den verwendeten Einstellungen keine aktiven Konformationen mehr zu erwarten waren. Da dieser Wert nicht von dem Zielprotein abhängt, lässt sich HYDE auch bei unbekannten Systemen einsetzen. Liegen die bei einem Screening erzielten Ergebnisse oberhalb dieses Grenzwertes, sind mit hoher Wahrscheinlichkeit auch auf den oberen Rängen keine aktiven Verbindungen zu finden.

Bei der Korrelation des HYDE-Scores mit publizierten Bindungs-Affinitäten zeigt sich dementsprechend, dass eine Abweichung bevorzugt in Richtung eines schlechteren Wertes erfolgt. Aus genannten Gründen werden einige Wechselwirkungen nicht erkannt und Komplexe werden entsprechend schlechter bewertet. Der umgekehrte Fall, dass eine Ver-

bindung zu gut bewertet wird, ist weitaus seltener zu beobachten. Solche Fälle konnten auf eine Überbewertung von Wasserstoff-Brücken in unzugänglichen Bindetaschen zurückgeführt werden, was die Frage aufwirft, ob die stärkere Berücksichtigung solcher Gruppen aufgrund der geringeren Desolvatations-Kosten überhaupt korrekt modelliert wird, und, noch entscheidender, ob dies zu einem Fehler bei einem Screening-Experiment führt.

Es wurde gezeigt, dass bei Verwendung dieses Modells eine geringere Übereinstimmung mit experimentellen Bindungs-Affinitäten erzielt wird als bei derjenigen Variante, die den Desolvatations-Term nicht bezüglich der mittleren Zugänglichkeit skaliert. Dies ist vor allem auf die Schwierigkeit zurückzuführen, die Zugänglichkeit in schwer zugänglichen Regionen hinreichend gut für diese Verwendung zu modellieren (siehe Abschnitt 4.2.3). Gerade bei einem Vergleich von Proteinen mit sehr unterschiedlichen Oberflächen-Charakteristika führt dies zu einem großen resultierenden Fehler.

Bei dem Einsatz in einem virtuellen Screening jedoch ändern sich die Charakteristika der Bindetasche nicht. An dieser Stelle ist es richtig und explizit erforderlich zwischen Atomen zu differenzieren, die unterschiedlich gut solvatisiert sind. Wasserstoff-Brücken in engen Bindetaschen fallen stärker ins Gewicht als solche in gut zugänglichen Bereichen. Der Fehler bei der Bewertung ist nun systematischer Natur und, wie gezeigt werden konnte, gering genug, um die Anwendbarkeit von HYDE in einem Screening-Szenario zu gewährleisten.

Dazu sei vor allem auf die experimentellen Ergebnisse hingewiesen. Diese belegen nicht nur eindrucksvoll, dass HYDE in der Lage ist, unter mehreren Millionen Moleküle mit großer Sicherheit solche zu bestimmen, die tatsächlich eine Affinität zum betrachteten Rezeptor aufweisen. Zusätzlich konnte auch die technische Anwendbarkeit von HYDE in einem industriellen Setup belegt werden. Trotz des erhöhten Rechenaufwands im Vergleich zu herkömmlichen Funktionen konnten die Rechnungen in einem tolerierbaren Zeitrahmen durchgeführt werden.

6.5. Anwendbarkeit von HYDE

Wichtig für eine korrekte Bewertung eines Protein-Ligand Komplexes ist an erster Stelle ein kompatibles Wechselwirkungs-Modell bei der Generierung der Eingabe-Daten. Die Aufklärung der dreidimensionalen Struktur eines Proteins erfolgt in der Regel durch das iterative Anpassen eines Modells an eine berechnete Elektronendichte. Für diesen Zweck wird eine Energie-Funktion verwendet, in den meisten Fällen ein klassisches Kraftfeld.

Wenn solch ein Kraftfeld eine Wasserstoff-Brücke zwischen zwei Atomen erkennt, heißt das nicht zwangsläufig, dass auch HYDE diese Konstellation auf diese Weise interpretiert. Für einen Modeller ist es einfach, aufgrund seiner Erfahrung eine gegebene Anordnung von Atomen als Wechselwirkung zu deuten. HYDE ist auf die korrekte Modellierung durch das in Abschnitt 4.3.1 beschriebene Wechselwirkungs-Modell angewiesen.

In vielen Fällen ist dies gegebeben, da die verwendeten Wechselwirkungs-Geometrien von Kristall-Strukturen kleiner Moleküle abgeleitet wurden. Mit relativ geringem Aufwand sollte sich dieses Modell weiter verfeinern lassen. Nicht zuletzt belegen dies die guten Resultate der wissensbasierten DrugScore^{CSD} Funktion, die anhand von geometrischen Eigenschaften strukturell aufgeklärter Komplexe parametrisiert worden ist [149]. Größere Abweichungen vom idealen Modell müssen jedoch immer zu einer Bestrafung führen, da eine zu große Toleranz die Diskriminierung zwischen Bindern und Nicht-Bindern verhindern würde.

Weitere grundlegende Voraussetzungen für die Anwendbarkeit von HYDE sind die richtige Ausrichtung von rotierbaren Wasserstoff-Brücken-Funktionen und eine korrekte Protonierung der Atome. (SYBYL-)Atomtypen müssen in konsistenter Weise zugewiesen werden, und es darf zu keiner übermäßigen sterischen Überlappung von Protein und Ligand kommen. Ebenso wird eine niedrige Konformations-Energie für die Bewertung mit HYDE vorausgesetzt. Ausdrücke dieser Art sind bewusst nicht Teil der HYDE Scoring-Funktion, da sie unabhängig vom Desolvatations- und Wechselwirkungs-Beitrag zu einem Verwerfen der Pose führen können. Selbst bei einem perfekten Wasserstoff-Brücken-Muster kann ein Ligand nicht binden, wenn er zu groß für die Bindetasche ist. Dementsprechend sollten diese Fälle als zusätzlicher Schritt bei der Aufbereitung oder in einem Filter-Prozess berücksichtigt werden. Sie leisten jedoch keinen energetischen Beitrag, der Teil einer Scoring-Funktion sein sollte.

Solange der Ursprung der zu bewertenden Strukturen bekannt und man sich deren Plausibilität sicher ist, sind Terme für sterische und konformationelle Barrieren also zu vernachlässigen. Bei der Bewertung von Strukturen, die mit unbekannten Programmen erzeugt und optimiert worden sind, sollten solche Ausschlusskriterien jedoch grundsätzlich überprüft werden. Ebenso wäre eine lokale Optimierung der Struktur hinsichtlich des Wechselwirkungs-Modells wünschenswert.

In sehr vereinfachender Weise werden durch HYDE geladene Gruppen und Metalle dargestellt. Diese werden in ihrem Bindungs-Verhalten analog zu Wasserstoff-Brücken-Funktionen behandelt, obwohl die physikalischen Grundlagen teilweise anderer Natur sind. Da keine plogP-Werte für solche funktionellen Gruppen berechnet werden können, wurden diese von korrespondierenden, ungeladenen Atomen übernommen, für Metall-Ionen wurde der Beitrag geschätzt.

Für eine erste Näherung ist eine solche Vorgehensweise sicherlich angebracht. Im Prinzip bedeutet auch für solche Gruppen eine Desolvatation zunächst einen energetischen Nachteil, da sie im ungebundenen Zustand stets von Gegen-Ionen umgekehrter Polarität umgeben sind. Dies lässt sich mit dem verwendeten Modell weitgehend richtig beschreiben. Im Speziellen ist aber zum Beispiel die Frage eines Referenz-Zustandes für die Zugänglichkeit zu klären, und auch wenn es aufgrund dieser Umstände bei einem virtuellen Screening nur zu einem systematischen Fehler kommt, sollte man sich der damit verbundenen Problematik stets bewusst sein.

Weiterhin ist die vereinfachende Annahme, dass jedes Atom unabhängig vom Wechselwirkungs-Partner mit einem konstanten Wert gemäß seines partiellen logP-Wertes zu einer Bindung beiträgt, nicht in allen Fällen zutreffend. Zum Beispiel ist damit nicht zu erklären, dass es durch den Austausch eines einzigen Atoms mit einer isosteren Gruppe zu einer großen Änderung in der Affinität kommen kann, obwohl gleichartgige Wechselwirkungen ausgebildet werden. Bei einer geeigneten Kombination von Wasserstoff-Brücken-Donor und -Akzeptor sind durch Resonanz-Effekte und teilweise kovalente Bindungen große Wechselwirkungs-Energien möglich. In solchen Ausnahme-Fällen ist das Ganze mehr als die Summe der Einzelbeiträge und HYDE kann diesen Umstand nicht mehr korrekt beschreiben.

7. Zusammenfassung und Ausblick

7.1. Zusammenfassung

Im Rahmen dieser Arbeit wurde eine neue Scoring-Funktion HYDE für die Bewertung von Protein-Ligand Komplexen entwickelt. HYDE beruht auf einer neuen Theorie, derzufolge der Beitrag von intermolekularen Wechselwirkungen zur Bindungs-Affinität zwischen zwei Molekülen unmittelbar mit den Eigenschaften des umgebenden Lösungsmittels zusammenhängt und auf deren Basis berechenbar ist. Mit nur zwei Termen a) für die Desolvatation und b) für diskrete Wechselwirkungen (Wasserstoff-Brücken) werden alle entscheidenden Einflüsse auf die Bindungs-Affinität in HYDE berücksichtigt. Diese Terme beschreiben sowohl die stabilisierenden Beiträge durch nichtbindende Wechselwirkungen und den hydrophoben Effekt als auch destabilisierende Beiträge, welche auf die Desolvatation polarer Atome zurückzuführen sind.

Die wichtigsten Deskriptoren, welche in der HYDE-Funktion Verwendung finden, sind zum einen von experimentell bestimmten Oktanol/Wasser Verteilungskoeffizienten abgeleitete Solvatations-Parameter, welche die Affinität der Atome zum Lösungsmittel beschreiben, sowie ein Ausdruck für die gewichtete Lösungsmittel-zugängliche Oberfläche. Der Zugänglichkeits-Deskriptor berücksichtigt in einer konsistenten Weise die bevorzugte Ausrichtung von Wasserstoff-Brückenbindungen zu den Lösungsmittel-Molekülen. Zu diesem Zweck wurde die ursprüngliche SAS-Definition mit dem FlexX Wechselwirkungs-Modell sinnvoll und in dieser Art und Weise neuartig erweitert.

Im Gegensatz zu anderen empirischen Scoring-Funktionen wurde HYDE nicht anhand von experimentell ermittelten Bindungs-Affinitäten strukturell aufgeklärter Komplexe parametrisiert. Die Verwendung von Verteilungs-Koeffizienten ermöglicht es, auch solche Konstellationen sinnvoll zu bewerten, die in Protein-Ligand Komplexen üblicherweise nicht vorkommen und deshalb bei der klassischen Parametrisierung keine Berücksichtigung finden können. Dies sind vor allem die destabilisierenden Beiträge, die einen entscheidenden Anteil am HYDE-Score ausmachen. Aufgrund der physikalisch konsequenten Herleitung der Funktions-Terme liegen die vorhergesagten Werte in einem Bereich, der mit gemessenen Affinitäten sehr gut übereinstimmt.

Gegenüber der FlexX Scoring-Funktion benötigt HYDE ein Vielfaches an Rechenzeit, die prinzipielle Komplexität des neuen Algorithmus ist jedoch von der gleichen Größenordnung. Eine Anwendbarkeit im industriellen Screening ist gegeben und Möglichkeiten zur Beschleunigung des Verfahrens sind offensichtlich. Aufgrund der expliziten Berücksichtigung destabilisierender Beiträge zur Bindungs-Affinität ist HYDE in der Lage wesentlich besser zwischen aktiven Verbindungen und Inaktiven zu diskriminieren. Die für die meisten Scoring-Funktionen typische große lineare Abhängigkeit zwischen dem berechneten Score und dem Molekular-Gewicht ist bei HYDE nicht zu beobachten. Von energetisch günstigen Konformationen eines Liganden in der Bindetasche abweichende Geometrien werden erkannt und entsprechend schlecht bewertet. Dies konnte durch theoretische Berechnungen von Anreicherungs-Kurven und erste experimentelle Ergebnisse belegt werden.

HYDE stellt einen großen Fortschritt für den strukturbasierten Entwurf von Liganden im Wirkstoffentwurf dar. Durch die Fähigkeit, falsch positive Hits in einem frühen Stadium der Forschung zu identifizieren, ermöglicht der Einsatz dieser Scoring-Funktion eine zielgerichtetere Entwicklung und eine sinnvollere Verwendung der zur Verfügung stehenden Kapazitäten. Im Besonderen ermöglicht HYDE auch die Bewertung für den Erfolg eines virtuellen Screenings, wenn nur wenig Vorwissen über das untersuchte Protein vorhanden ist. Durch die Berücksichtigung lokaler Eigenschaften der Bindetasche erfolgt eine implizite Adaption der Funktion. Die gegenseitige Gewichtung von Desolvatations- und Wechselwirkungs-Beiträgen hängt direkt von der Zugänglichkeit einzelner Atome zum Lösungsmittel ab. Der erzielte Score kann als absolutes Maß für die Bindungs-Affinität verwendet werden.

7.2. Ausblick

Im Rahmen der Validierung und der genaueren Analyse der Funktion wurden die Grenzen der Anwendbarkeit aufgezeigt, aber auch erste Vorschläge für zukünftige Verbesserungen gemacht. In Abschnitt 5.3 wurden prinzipielle Möglichkeiten angeführt, die Funktion schneller und effizienter berechenbar zu machen. Eine schnellere Anwendbarkeit von HYDE würde unter anderem ermöglichen, die Funktion schon während des Docking-Vorgangs zu verwenden und damit die Generierung sinnvoller Lösungen durch den *k-greedy* Algorithmus zu unterstützen. Wichtig wäre für diesen Zweck neben der reinen Geschwindigkeit auch eine direkte Unterstützung der Scoring-Funktion, etwa durch eine genauere Abschätzung der maximal erzielbaren Energie bzw. einen zielgerichteten Aufbau, der nicht nur (wie es derzeit der Fall ist) auf eine Maximierung der gemachten Wasserstoff-Brücken abzielt, sondern auch destabilisierende Effekte berücksichtigt.

Vor einer technischen Verbesserung der Scoring-Funktion sollte zunächst das theoretische Modell weiter verfeinert und generalisiert werden. In einem konsekutiven Prozess sollte zuerst das zugrunde liegende Wechselwirkungs-Modell überarbeitet werden. Darauf basierend sollten weitere Methoden entwickelt werden, die Zugänglichkeit von Atomen zu berechnen, wobei der besondere Fokus auf der korrekten Modellierung schwer zugänglicher Regionen läge. Abschließend muss mit dem neuen Modell eine erneute Parametrisierung vorgenommen werden. Im Rahmen der Parametrisierung sollten auch Möglichkeiten erwogen werden, die besondere Rolle von Wechselwirkungen zwischen geladenen Atomen besser zu berücksichtigen.

Für die Überarbeitung des Wechselwirkungs-Modells bieten sich im Grunde zwei Methoden an. Die bevorzugte Ausrichtung von Wasserstoff-Brücken könnte anhand einer Analyse aktueller Struktur-Informationen erfolgen, ebenso denkbar wäre eine physikalischere Methode, welche aufgrund von Ladungs-Verteilungen vorherrschende elektrostatische Einflüsse integriert, etwa in Form eines Kraftfeldes. Beide Methoden würden Vor- und Nachteile mit sich bringen. Das generelle Ziel dieses Vorhabens wäre die bessere Identifizierung von Wechselwirkungen in den zu bewertenden Komplexen.

Dies wirkt sich natürlich essentiell auf die gewichtete Berechnung der Zugänglichkeit aus. Hier sollten insbesondere Anstrengungen unternommen werden, die Besonderheiten von schwer zugänglichen Regionen besser abzubilden. Wichtig wäre auch eine gesonderte Beurteilung der zur Normierung dienenden mittleren Zugänglichkeit von Protein-Atomen. Für Metall-Ionen in der Bindetasche bietet sich die Kombination mit einem kürzlich veröffentlichten Modell zur Vorhersage von Koordinations-Geometrien in Metalloproteinen aufgrund struktureller Eigenschaften an [133].

Da bei der Parametrisierung die Zugänglichkeit der Atom-Typen berücksichtigt wird, muss eine solche Analyse nach jeder Veränderung an dem Modell wiederholt werden. Auch hier gibt es Möglichkeiten, die Vorhersagekraft von HYDE weiter zu verbessern, etwa durch eine direkte Verwendung von experimentellen Desolvatations-Energien anstatt der Verteilungs-Koeffizienten oder durch Einsatz von aktuelleren Parametrisierungs-Methoden wie OPLS, neuronalen Netzwerken oder maschinellen Lernverfahren.

Neben der Verwendung von HYDE für das virtuelle Screening im herkömmlichen Sinne bietet sich ein Einsatz von HYDE bei allen Fragestellungen an, welche eine effiziente Berechnung von intermolekularen Wechselwirkungen erfordern. Einhergehend mit einer experimentellen Validierung der Scoring-Funktion im alltäglichen praktischen Einsatz ist eine Kombination mit anderen in der Arbeitsgruppe entwickelten Methoden angedacht, um Synergie-Effekte zu nutzen und die Effizienz von strukturbasierten Methoden im Wirkstoffdesign weiter zu verbessern.

Literaturverzeichnis

[1] http://www.syrres.com/esc/physprop.htm.

[2] R&D myths: The case against the drug industry's R&D scare card. Technical report, Public Citizen, 2001.

[3] J. Åvist, C. Medina, and J. E. Samuelson. A new method for predicting binding affinity in computer-aided drug design. *Prot. Eng.*, 7:385–391, 1994.

[4] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM - a new method for protein modeling and design: Applications to docking and structure predictions from the distorted native conformation. *J. Comput. Chem.*, 15(5):488–506, 1994.

[5] F. H. Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.*, 58:380–388, 2002.

[6] P. Aloy, E. Querol, F. X. Aviles, and M. J. Sternberg. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, 311(2):395–408, 2001.

[7] D. R. Anderson, M. J. Meyers, W. F. Vernier, M. W. Mahoney, R. G. Kurumbail, N. Caspers, G. I. Poda, J. F. Schindler, D. B. Reitz, and R. J. Mourey. Pyrrolopyridine inhibitors of mitogen-activated protein kinase-activated protein kinase 2 (mk-2). *J. Med. Chem.*, 50(11):2647–2654, 2007.

[8] I. Antes, C. Merkwirth, and T. Lengauer. POEM: Parameter optimization using ensemble methods: application to target specific scoring functions. *J. Chem. Inf. Model.*, 45(5):1291–302, 2005.

[9] J. D. Augspurger and H. A. Scheraga. An efficient, differentiable hydration potential for peptides and proteins. *J. Comput. Chem.*, 17(13):1549–1558, 1996.

[10] J. C. Baber, W. A. Shirley, Y. Gao, and M. Feher. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.*, 46(1):277–88, 2006.

[11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[12] D. L. Beveridge and F. M. Dicapua. Free-energy via molecular simulation - applications to chemical and biomolecular systems. *Ann. Rev. Biophys.Biophys. Chem.*, 18:431–492, 1989.

[13] C. Bissantz, G. Folkers, and D. Rogman. Protein based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *J. Med. Chem.*, 43(25):4759–4767, 2000.

[14] H. J. Böhm. The computer-program LUDI - a new method for the denovo design of enzyme-inhibitors. *J. Comput. Aided Mol. Des.*, 6(1):61–78, 1992.

[15] H.-J. Böhm. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Des.*, 6:593–606, 1992.

[16] H.-J. Böhm. 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. Aid. Mol. Des.*, 8:243–256, 1994.

[17] H.-J. Böhm. Prediction of binding constants of protein ligands: A fast method for the priorization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided Mol. Des.*, 12:309–323, 1998.

[18] J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut. Crystal-structures of escherichia-coli and lactobacillus-casei dihydrofolate-reductase refined at 1.7 a resolution .1. general features and binding of methotrexate. *J. Biol. Chem.*, 257(22):3650–3662, 1982.

[19] P. Broto, G. Moreau, and C. Vandycke. Molecular structures: perception, autocorrelation, descriptor and sar studies. *Eur. J. Med. Chem.*, 19(1):71–78, 1984.

[20] I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor, and M. L. Verdonk. Isostar: A library of information about nonbonded interactions. *J. Comput. Aided Mol. Des.*, 11:525–537, 1997.

[21] C. Bystroff. MASKER: improved solvent-excluded molecular surface area estimations using boolean masks. *Prot. Eng.*, 15(12):959–965, 2002.

[22] P. S. Charifson, J. J. Corkerz, M. A. Murcko, and W. P. Walters. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.*, 42:5100–5109, 1999.

[23] X. Chen, Z. L. Ji, D. G. Zhi, and Y. Z. Chen. CLiBE: a database of computed ligand binding energy for ligand-receptor complexes. *Comp. Chem.*, 26(6):661–666, 2002.

[24] M. M. Chow, E. F. Meyer, W. Bode, C. M. Kam, R. Radhakrishnan, J. Vijayalakshmi, and J. C. Powers. The 2.2-a resolution x-ray crystal-structure of the complex of trypsin inhibited by 4-chloro-3-ethoxy-7-guanidinoisocoumarin - a proposed model of the thrombin inhibitor complex. *J. Am. Chem. Soc.*, 112(21):7783–7789, 1990.

[25] R. D. Clark, A. Strizhev, J. M. Leonard, C. F. Blake, and J. B. Matthew. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.*, 20(4):281–295, 2002.

[26] H. Claußen, C. Buning, M. Rarey, and T. Lengauer. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, 308(2):377–395, 2001.

[27] M. L. Connolly. Analytical molecular-surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.

[28] M. L. Connolly. Molecular-surface triangulation. *J. Appl. Cryst.*, 18:499–505, 1985.

[29] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[30] J. Degen and M. Rarey. FlexNovo: structure-based searching in large fragment spaces. *ChemMedChem*, 1(8):854–68, 2006.

[31] R. S. DeWitte and E. I. Shakhnovich. 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, 1996.

[32] D. J. Diller and R. X. Li. Kinases, homology models, and high throughput docking. *J. Med. Chem.*, 46(22):4638–4647, 2003.

[33] F. Dullweber, M. T. Stubbs, D. Musil, J. Sturzebecher, and G. Klebe. Factorising ligand affinity: A combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *J. Mol. Biol.*, 313(3):593–614, 2001.

[34] H. Edelsbrunner. The union of balls and its dual shape. *Disc. Comput. Geom.*, 13(3-4):415–440, 1995.

[35] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(16):199–203, 1986.

[36] F. Eisenhaber and P. Argos. Improved strategy in analytic surface calculation for molecular-systems - handling of singularities and computational-efficiency. *J. Comput. Chem.*, 14(11):1272–1280, 1993.

[37] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf. The double cubic lattice method - efficient approaches to numerical-integration of surface-area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.*, 16(3):273–284, 1995.

[38] M. D. Eldridge, C. W. Murraz, T. R. Auton, G. V. Paolini, and R. P. Mee. 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.*, 11:425–445, 1997.

[39] T. J. A. Ewing and I. D. Kuntz. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.*, 18(9):1175–1189, 1997.

[40] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, 15:411–428, 2001.

[41] M. Feher, E. Deretey, and S. Roy. BHB: A simple knowledge-based scoring function to improve the efficiency of database screening. *J. Chem. Inf. Comput. Sci.*, 43(4):1316–1327, 2003.

[42] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, and C. L. Brooks III. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.*, 47:3032–3047, 2004.

[43] R. W. Finn and L. Kavraki. Computational approaches to drug design. *Algorithmica*, 25:347–371, 1999.

[44] A. Flemming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. *Br. J. Exp. Pathol.*, 10(31):226–236, 1929.

[45] R. Fraczkiewicz and W. Braun. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macormolecules. *J. Comput. Chem.*, 19(3):319–333, 1998.

[46] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–49, 2004.

[47] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.*, 49(21):6177–6196, 2006.

[48] P. Geladi and B. R. Kowalski. Partial least-squares regression - a tutorial. *Anal. Chim. Acta*, 185:1–17, 1986.

[49] D. K. Gelhaar, G. M. Verkhivker, R. P. A., C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.*, 2(5):317–324, 1995.

[50] R. Gericke, N. Beier, O. Poeschke, L. Burgdorf, H. Drosdat, and F. Lang. Preparation of acylhydrazones as modulators of glucocorticoid inducible kinase, 2005. PCT Int Appl. WO 2004-EP10398 20040916.

[51] A. K. Ghose and G. M. Crippen. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.*, 7(4):565–577, 1986.

[52] K. D. Gibson and H. A. Scheraga. Surface-area of the intersection of 3 spheres with unequal radii - a simplified analytical formula. *Mol. Phys.*, 64(4):641–644, 1988.

[53] F. Giordanetto, S. Cotesta, C. Catana, J.-Y. Trosset, A. Vulpetti, P. F. W. Stouten, and R. T. Kroemer. Novel scoring functions comprising QXP, SASA, and protein side-chain entropy terms. *J. Chem. Inf. Comp. Sci.*, 44:882–893, 2004.

[54] P. W. Glunz. Preparation of pyrrolo[1,2-a]pyriminidones as coagulation cascade inhibitors, 2004. U.S. Pat. Appl. US 2004006065.

[55] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337–356, 2000.

[56] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley, Longman, Reading, MA, 1989.

[57] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28:849–857, 1985.

[58] O. Guvench and C. L. Brooks III. Efficient approximate all-atom solvent accessible surface area method parameterized for folded and denatured protein conformations. *J. Comput. Chem.*, 25(8):1005–14, 2004.

[59] C. Hansch, A. Leo, and D. Hoekman. *Exploring QSAR. Hydrophobic, electronic, and steric constants.* American Chemical Society, Washington, DC, 1995.

[60] M. J. Hartshorn, C. W. Murray, A. Cleasby, M. Frederickson, I. J. Tickle, and H. Jhoti. Fragment-based lead discovery using x-ray crystallography. *J. Med. Chem.*, 48(2):403–413, 2005.

[61] S. Hayryan, C. K. Hu, J. Skrivanek, E. Hayryan, and I. Pokorny. A new analytical method for computing solvent-accessible surface area of macromolecules and its gradients. *J. Comput. Chem.*, 26(4):334–343, 2005.

[62] R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118:3959–3969, 1996.

[63] M. Hendlich, A. Bergner, J. Günther, and G. Klebe. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, 326:607–620, 2003.

[64] M. Hendlich, F. Rippmann, and G. Barnickel. LIGSITE: automatic and efficient

detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, 15(6):359–63, 389, 1997.

[65] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.

[66] T. Hou, W. Zhang, Q. Huang, and X. Xu. An extended aqueous solvation model based on atom-weighted solvent accessible surface areas: SAWSA v2.0 model. *J. Mol. Model.*, 11:26–40, 2005.

[67] T. J. Hou and X. J. Xu. ADME evaluation in drug discovery. 2. prediction of partition coefficients by atom-additive approach based on atom-weighted solvent accessible surface areas. *J. Chem. Inf. Comp. Sci.*, 43(3):1058–1067, 2003.

[68] N. Huang, B. K. Shoichet, and J. J. Irwin. Benchmarking sets for molecular docking. *J. Med. Chem.*, 49:6789–6801, 2006.

[69] S. Y. Huang and X. Zou. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. derivation of interaction potentials. *J. Comput. Chem.*, 27(15):1866–75, 2006.

[70] A. V. Ishchenko and E. I. Shakhnovich. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J. Med. Chem.*, 45:2770–2780, 2002.

[71] A. N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided Mol. Des.*, 10:427–440, 1996.

[72] G. Jones and P. Willet. Docking small-molecule ligands into active sites. *Curr. Opin. Biotech.*, 6:652–656, 1995.

[73] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.

[74] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[75] E. Kellenberger, J. Rodrigo, P. Muller, and D. Rognan. Comparative evaluation of

eight docking tools for docking and virtual screening accuracy. *Proteins*, 57:225–242, 2004.

[76] G. E. Kellogg, J. C. Burnett, and D. J. Abraham. Very empirical treatment of solvation and entropy: a force field derived from $\mathrm{LogP}_{O/W}$. *J. Comput. Aid. Mol. Des.*, 15:381–393, 2001.

[77] T.-S. Kim, J.-C. Harmange, S. Booker, N. D'Angelo, C. Dominguez, I. M. Fellows, L. Liu, A. Tasker, S. Bellon, T. S. Harvey, M. Lee, J. Germain, V. F. Patel, and J. L. Kim. Preparation of bis-aryl kinase inhibitors, 2007. PCT Int. Appl. US 20070054903.

[78] S. D. Kimball, S. E. Hall, and W. F. Lau. Guanidyl- or amidinyl-substituted hererocyclic thrombin inhibitors, 1994. Eur. Pat. Appl. EP623595.

[79] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery methods and applications. *Nat. Rev.*, 3:935–949, 2004.

[80] P. A. Kollman. Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417, 1993.

[81] E. T. Kool. Hydrogen bonding, base stacking, and steric effects in dna replication. *Annu. Rev. Biophys. Biomol. Struct.*, 30:1–22, 2001.

[82] A. Krammer, P. D. Kirchhoff, X. Jiang, C. M. Venkatachalam, and M. Waldmann. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.*, 23:395–407, 2005.

[83] R. Krishnan, E. Zhang, K. Hakansson, R. K. Arni, A. Tulinsky, M. S. Lim-Wilby, O. E. Levy, J. E. Semple, and T. K. Brunck. Highly selective mechanism-based thrombin inhibitors: structures of thrombin and trypsin inhibited with rigid peptidyl aldehydes. *Biochemistry*, 37(35):12094–12103, 1998.

[84] E. M. Krovat, T. Steindl, and T. Langer. Recent advances in docking and scoring. *Cur. Comp. Aided Drug Des.*, 1(1):93–102, 2005.

[85] H. Kubinyi. Success stories of computer-aided design. In S. Ekins, editor, *Computer applications in pharmaceutical research and development*. Wiley & Sons, New York, 2006.

[86] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, 1982.

[87] G. Lange, I. Reulecke, M. Rarey, and R. Klein. Impact of satisfied and unsatisfied H-bond functions within the water network on the intermolecular interaction in aqueous solutions. *PhysChemPhys*, 2007. submitted.

[88] R. A. Laskowski, J. M. Thornton, C. Humblet, and J. Singh. X-site: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. *J. Mol. Biol.*, 259(1):175–201, 1996.

[89] G. Lauri and P. A. Bartlett. Caveat - a program to facilitate the design of organic-molecules. *J. Comput. Aided Mol. Des.*, 8(1):51–66, 1994.

[90] S. M. Le Grand and K. M. Merz. Rapid approximation to molecular-surface area via the use of boolean logic and look-up tables. *J. Comput. Chem.*, 14(3):349–352, 1993.

[91] A. R. Leach, B. K. Shoichet, and C. E. Peishoff. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.*, 49(20):5851–5855, 2006.

[92] J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape. *Proteins*, 33(1):1–17, 1998.

[93] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeny. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.*, 23:3–25, 1997.

[94] P. Maass, T. Schulz-Gasch, M. Stahl, and M. Rarey. Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.*, 47(2):390–399, 2007.

[95] D. MacKerell, Jr., D. Bashford, M. Bellot, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[96] P. M. Marsden, D. Puvanendrampillai, J. B. O. Mitchell, and R. C. Glen. Predicting protein-ligand binding affinities: a low scoring game? *Org. Biomol. Chem.*, 2:3267–3273, 2004.

[97] B. J. McConkey, V. Sobolev, and M. Edelman. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained voronoi procedure. *Bioinf.*, 18(10):1365–73, 2002.

[98] J. B. O. Mitchell, R. A. Laskowski, A. Alex, and J. M. Thornton. BLEEP - potential of mean force describing protein-ligand interactions: I. generating potential. *J. Comput. Chem.*, 20(11):1165–1176, 1999.

[99] D. C. Montgomery. *Design and analysis of experiments.* John Wiley & Sons, New York, 4 edition, 1997.

[100] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662, 1998.

[101] I. Muegge. PMF scoring revisited. *J. Med. Chem.*, 49(20):5895–902, 2006.

[102] I. Muegge and Y. C. Martin. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.*, 42(5):791–804, 1999.

[103] A. E. Muryshev, D. N. Tarasov, A. V. Butygin, O. Y. Butygina, A. B. Aleksandrov, and S. M. Nikitin. A novel scoring function for molecular dockling. *J. Comput. Aided Mol. Des.*, 17:597–605, 2003.

[104] A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu, and S. Hirono. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *J. Chem. Inf. Model.*, 46(1):380–91, 2006.

[105] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameters sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[106] V. D. Ozrin, M. V. Subbotin, and S. M. Nikitin. PLASS: Protein-ligand affinity statistical score - a knowledge-based force-field model of interaction derived from the pdb. *J. Comput. Aided Mol. Des.*, 18:261–270, 2004.

[107] C. Pargellis, L. Tong, L. Churchill, P. F. Cirillo, T. Gilmore, A. G. Graham, P. M. Grob, E. R. Hickey, N. Moss, S. Pav, and J. Regan. Inhibition of p38 map kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.*, 9(4):268–272, 2002.

[108] R. Parthasarathi and V. Subramanian. Characterization of hydrogen bonding: from van der waals interactions to covalency. In S. J. Grabowski, editor, *Hydrogen bonding - new insights*, volume 3 of *Challenges and advances in Comp. Chem. and physics*. Springer, Dordrecht, 2006.

[109] L. Pauling. *The nature of the chemical bond*. Cornell University Press, New York, 1960.

[110] D. A. Pearlman and P. S. Charifson. Are free energy calculations useful in practice? a comparison with rapid scoring functions for the p38 map kinase protein system. *J. Med. Chem.*, 44:3417–3423, 2001.

[111] R. S. Pearlman. System and method for providing a canonical structural representation of chemical compounds, 2005. U.S. Pat. Appl. Publ. US 2005125210.

[112] C. Pérez and A. R. Ortiz. Evaluation of docking functions for protein-ligand docking. *J. Med. Chem.*, 44:3768–3785, 2001.

[113] E. Perola and P. S. Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.*, 47:2499–2510, 2004.

[114] E. Perola, W. P. Walters, and P. S. Charifson. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, 56:235–249, 2004.

[115] G. Perrot, B. Cheng, K. D. Gibson, J. Vila, K. A. Palmer, A. Nayeem, B. Maigret, and H. A. Scheraga. Mseed - a program for the rapid analytical determination of accessible surface-areas and their derivatives. *J. Comput. Chem.*, 13(1):1–11, 1992.

[116] M. Petitjean. On the analytical calculation of van-der-waals surfaces and volumes - some numerical aspects. *J. Comput. Chem.*, 15(5):507–523, 1994.

[117] T. A. Pham and A. N. Jain. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.*, 49(20):5856–5868, 2006.

[118] A. C. Pierce, G. Rao, and G. W. Bemis. BREED: Generating novel inhibitors

through hybridization of known ligands. application to CDK2, P38, and HIV protease. *J. Med. Chem.*, 47(11):2768–2775, 2004.

[119] M. Rarey, J. Degen, and I. Reulecke. Docking and scoring for structure-based drug design. In T. Lengauer, editor, *Bioinformatics - from genomes to therapies*, volume 2. Wiley-VCH, Weinheim, 2007.

[120] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.

[121] I. Reulecke, G. Lange, J. Albrecht, R. Klein, and M. Rarey. Towards an integrated description of hydrogen bonding and dehydration: Reducing false positives in virtual screening using the HYDE scoring function. *ChemMedChem*, 2007. submitted.

[122] F. M. Richards. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bio.*, 6:151–176, 1977.

[123] T. J. Richmond. Solvent accessible surface-area and excluded volume in proteins - analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.*, 178(1):63–89, 1984.

[124] U. Ries, N. Hauel, H. Mihm, G. Priepke, K. Binder, J. M. Stassen, W. Wienen, and R. Zimmermann. Preparation of five-membered, benzo-condensed heterocycles as antithrombotics. *PCT Int. Appl. WO 9940072*, 1999.

[125] O. Roche, R. Kiyama, and C. L. Brooks, III. Ligand-protein database: Linking protein-ligand complex structures to binding data. *J. Med. Chem.*, 44:3592–3598, 2001.

[126] D. Rognan, S. L. Laeumoeller, A. Holm, S. Buus, and V. Tschinke. 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, 1999.

[127] G. Rychkov and M. Petukhov. Joint neighbors approximation of macromolecular solvent accessible surface area. *J. Comput. Chem.*, 28(12):1974–1989, 2007.

[128] J. Sadowski and J. Gasteiger. From atoms and bonds to 3-dimensional atomic coordinates: automatic model builders. *Chem. Rev.*, 93(7):2567–2581, 1993.

[129] H. Samet. *Spatial Data Structures: Quadtree, Octrees and Other Hierarchical Methods.* Addison Wesley, 1989.

[130] H. J. Savage, C. J. Elliott, C. M. Freeman, and J. L. Finney. Lost hydrogen-bonds and buried surface-area - rationalizing stability in globular proteins. *J. Chem. Soc. Faraday Trans.*, 89(15):2609–2617, 1993.

[131] G. Schneider and U. Fechner. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery*, 4(8):649–663, 2005.

[132] T. Schulz-Gasch and M. Stahl. Scoring functions for protein ligand interactions: A critical perspective. *Drug Discov. Today*, 1(3):231–239, 2004.

[133] B. Seebeck, R. I., A. Kämper, and M. Rarey. Modeling of metal interaction geometries for protein-ligand docking. *Proteins*, 2007. DOI 10.1002/prot.21818.

[134] M. H. Seifert. Assessing the discriminatory power of scoring functions for virtual screening. *J. Chem. Inf. Model.*, 46(3):1456–1465, 2006.

[135] A. K. Shiau, D. Barstad, J. T. Radek, M. J. Meyers, K. W. Nettles, B. S. Katzenellenbogen, J. A. Katzenellenbogen, D. A. Agard, and G. L. Greene. Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism. *Nat. Struct. Biol.*, 9(5):359–364, 2002.

[136] A. Shrake and J. A. Rupley. Environment and exposure to solvent of protein atoms - lysozyme and insulin. *J. Mol. Biol.*, 79(2):351–371, 1973.

[137] T. P. Singh, S. Sharma, S. Karthikeyan, C. Betzel, and K. L. Bhatia. Crystal structure of a complex formed between proteolytically-generated lactoferrin fragment and proteinase K. *Proteins*, 33:30–38, 1998.

[138] M. J. Sippl. 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(4):859–883, 1990.

[139] A. K. Soper. Neutron scattering studies of solvent structure in systems of chemical and biological importance. *Faraday Disc.*, 103:41–58, 1996.

[140] E. Specker, J. Bottcher, H. Lilie, A. Heine, A. Schoop, G. Muller, N. Griebenow, and G. Klebe. An old target revisited: Two new privileged skeletons and an unexpected

binding mode for HIV-protease inhibitors. *Angew. Chem. Int. Ed.*, 44(20):3140–3144, 2005.

[141] M. Stahl. Modifications of the scoring funtion in FlexX for virtual screening applications. *Perspect. Drug Disc. Des.*, 20:83–98, 2000.

[142] M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, 44:1035–1042, 2001.

[143] K. Stierand and M. Rarey. From modeling to medicinal chemistry: Automatic generation of two-dimensional complex diagrams. *ChemMedChem*, 2(6):853–860, 2007.

[144] Y. Takamatsu and A. Itai. A new method for predicting binding free energy between receptor and ligand. *Proteins*, 33:62–73, 1998.

[145] F. S. Tjoeng and M. V. Toth. Preparation of N-acyl β-amino acid derivatives as platelet aggregration inhibitors, 1994. PCT Int. Appl. WO 9420457.

[146] N. P. Todorov and P. M. Dean. Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design. *J. Comput. Aided Mol. Des.*, 11(2):175–192, 1997.

[147] N. P. Todorov and P. M. Dean. A branch-and-bound method for optimal atom-type assignment in de novo ligand design. *J. Comput. Aided Mol. Des.*, 12(4):335–349, 1998.

[148] P. Traxler and P. Furet. Strategies toward the design of novel and selective protein tyrosine kinase inhibitors. *Pharmacol. Therapeut.*, 82(2-3):195–206, 1999.

[149] H. F. Velec, H. Gohlke, and G. Klebe. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.*, 48(20):6296–6303, 2005.

[150] R. Wang, X. Fang, Y. Lu, and S. Wang. The PDBind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, 47:2977–2980, 2004.

[151] R. Wang, X. Fang, Y. Lu, C. Y. Yang, and S. Wang. The PDBbind database: methodologies and updates. *J. Med. Chem.*, 48(12):4111–4119, 2005.

[152] R. Wang, Y. Gao, and L. Lai. Calculating partition coefficient by atom-additive method. *Perspec. Drug Discov.*, 19:47–66, 2000.

[153] R. Wang, L. Lai, and S. Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aid. Mol. Des.*, 16:11–26, 2002.

[154] R. Wang, L. Liu, L. Lai, and Y. Tang. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.*, 4:379–394, 1998.

[155] R. Wang, Y. Lu, X. Fang, and S. Wang. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comp. Sci.*, 44(6):2114–2125, 2004.

[156] R. Wang, Y. Lu, and S. Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46:2287–2303, 2003.

[157] Z. Wang, B. J. Canagarajah, J. C. Boehm, S. Kassisa, M. H. Cobb, P. R. Young, S. Abdel-Meguid, J. L. Adams, and E. J. Goldsmith. Structural basis of inhibitor selectivity in MAP kinases. *Structure*, 6(9):1117–28, 1998.

[158] J. Weiser, P. S. Shenkin, and W. C. Still. Approximate solvent-accessible surface areas from tetrahedrally directed neighbor densities. *Biopolymers*, 50(4):373–380, 1999.

[159] J. Weiser, P. S. Shenkin, and W. C. Still. Optimization of gaussian surface calculations and extension to solvent-accessible surface areas. *J. Comput. Chem.*, 20(7):688–703, 1999.

[160] S. Xiang, S. A. Short, R. Wolfenden, and C. W. Carter. Transistion-state selectivity for a single OH group during catalysis by cytidine deaminase. *Biochem.*, 34:4516–4523, 1995.

[161] C. Y. Yang, R. Wang, and S. Wang. M-score: a knowledge-based potential scoring function accounting for protein atom mobility. *J. Med. Chem.*, 49(20):5903–11, 2006.

[162] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou. A knowledge-based energy function for protein-ligand, protein-protein, and protein-dna complexes. *J. Med. Chem.*, 48(7):2325–2335, 2005.

[163] J. Zhang, M. Aizawa, S. Amari, Y. Iwasawa, T. Nakano, and K. Nakata. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.*, 28(5-6):401–7, 2004.

A. Dokumentation

A.1. HYDE Einstellungen

HYDE ist in die Docking-Software FlexX integriert. Grundlegende Funktionen sind dem FlexX-Benutzer-Handbuch zu entnehmen. Im Folgenden werden die HYDE-spezifischen Einstellungen erläutert.

A.1.1. Konfiguration

In der Konfigurations-Datei `config.dat` werden die verwendeten Parameter-Dateien und Prgramm-Pfade definiert. Abweichend von den Standard-Einstellungen werden für HYDE folgende Dateien verwendet.

```
#application HYDE
SETTINGS      static_data/flexx_settings_hyde.dat
LOGP          static_data/logp_hyde.dat
AMINO         static_data/amino_hyde.dat
CONTACT       static_data/contact_hyde.dat
GEOMETRY      static_data/geometry_hyde_fscore.dat
GEOMETRY      static_data/geometry_hyde.dat
GRAPHIC       static_data/graphic_hyde.dat
#end_application
```

A.1.2. SETTINGS

In `flexx_settings_hyde.dat` werden zusätzlich zu den FlexX-Parametern HYDE-spezifische Einstellungen vorgenommen. Neben den Parametern für die rekursive Berechnung

der Zugänglichkeit lässt sich mit `ONLY_ACTIVES` einstellen, ob bei der Berechnung des *contact*-Termes (*lipo*, *ambig* und *clash*) nur die Atome der Bindetasche berücksichtigt werden sollen. Standardmäßig ist diese Option ausgeschaltet, damit entspricht der *contact*-Term dem Original. Wichtig: Bei der Berechnung der Zugänglichkeit für die HYDE-Funktion werden grundsätzlich nur die Atome der Bindetasche berücksichtigt.

```
#application HYDE
# ---------------------------------------------------------------------
# LOGP_CLASS is redefined via geometry settings (depends on
# accessibility algorithm)
# SAS 1, WSAS 2, WRAY 3, SWSA 4
ONLY_ACTIVES    0         # use only active site atoms for contact score
WSAS_MIN_LEVEL  0
WSAS_MAX_LEVEL  2
WSAS_EPS        0.001
  # approximation levels for SAS calculation (WARNING: sas computing time
  # increases with 4th-power of this params)
#end_application
```

A.1.3. LOGP

Abweichend von der Original-Version dienen die in `logp_hyde.dat` definierten Atom-Typen und logP-Inkremente vor allem der Berechnung des HYDE-Scores. Eine grobe Abschätzung des logP-Wertes über das `LIGAND` Menü ist weiterhin möglich. In der Datei sind die für unterschiedliche Zugänglichkeits-Algorithmen bestimmten logP-Inkremente definiert. Die Auswahl des zu verwendenden Wertes erfolgt über den Klassen-Eintrag (die erste Zahl in der Header-Zeile). Als Beispiel ist der Eintrag eines Methyl-Kohlenstoff-Atoms der Klasse 1 (entspricht dem SAS-Algorithmus) angeführt:

```
@subgraph 1 4 C_sp3_ssss_H3
smarts [#6X4H3](-[#1])(-[#1])(-[#1])-*
data
logp    0.75
ar_I    0.0
ar_III  0.0
access  0.54
sd      0.0
```

```
hb_cap  0
end
```

Neben einem Eintrag für den logP-Wert sind Angaben zu dem atomaren Brechungs-Index möglich, welche im Original Ghose/Crippen-Artikel zusätzlich angegeben sind. Weiterhin lassen sich (nur in der HYDE-Version) auch die mittlere Zugänglichkeit `access` und Werte für die Standardabweichung und die Wasserstoff-Brücken-Kapazität des Atoms definieren. Während `access` essentiell für die korrekte Berechnung von HYDE ist, sind die beiden anderen Parameter optional (die Werte finden derzeit keine Anwendung).

Achtung: Der in der `SETTINGS` Datei angegebene Wert für `LOGP_CLASS` wird bei Verwendung von HYDE automatisch bezüglich des verwendeten Zugänglichkeits-Algorithmus überschrieben.

A.1.4. GEOMETRY

In dieser Datei sind die wesentlichen Einstellungen der Scoring-Funktion zu finden. Die Datei ist aufgeteilt in einen Bereich für die globalen Parameter der Scoring-Funktion (Auswahl und Gewichtung der Scoring-Terme), spezifische Angaben zu den einzelnen Termen und die Definition der Wechselwirkungs-Geometrien. Die beiden in der Konfigurations-Datei angegebenen Dateien `geometry_hyde_fscore.dat` und `geometry_hyde.dat` unterscheiden sich lediglich in der Auswahl der Scoring-Terme. Die erste Datei verwendet die Original Scoring-Funktion mit angepassten Wechselwirkungs-Geometrien, `geometry_hyde.dat` enthält die Einstellungen für die HYDE-Funktion. Soll (z.B. für das Docking) die Original-Version für die Bewertung des Protein-Ligand Komplexes verwendet werden, so muss der zweite Eintrag in `config.dat` auskommentiert werden.

Nachfolgend die Einstellungen für die HYDE Funktion, unveränderte bzw. ungenutzte Scoring-Terme sind nicht angeführt:

```
@scoring_parameters
 G_match              1.0     0.0     1.0     #partial # -
 G_rotbonds           1.4     0.0     1.0     #partial # +
 G_constant           5.4     0.0     1.0     #partial # -
 G_lipo_contacts     -0.17    0.0     1.0     #partial # +
 G_ambig_contacts    -0.17    0.0     1.0     #partial # +
 G_close_contacts    -0.34    0.0     1.0     #partial # -
```

```
#application HYDE
 G_HYDE_saturation      1.2     1.0     0.0     #final  # -
 G_HYDE_desolvation     5.7     1.0     0.0     #final  # +/-
 G_HYDE_exposed         0.07    0.0     0.0     #never  # +
 G_HYDE_ligand          1.0     0.0     0.0     #never  # -
#end_application
```

Diese Einstellungen bewirken eine finale Bewertung eines Komplexes durch die HYDE-Funktion, während beim Komplex-Aufbau die Original-Funktion (mit angepassten Geometrien) verwendet wird. Da beim Docking alle Lösungen final bewertet werden, ist ein Docking auch mit diesen Einstellungen aus Zeit-Gründen nicht ratsam.

`G_HYDE_saturation` und `G_HYDE_desolvation` sind die beiden HYDE-Terme für die Berechnung der Wechselwirkungs- und der Desolvatations-Energie. Die Parameter entsprechen den physikalischen Konstanten $2.3RT \approx 5.7$ und $1/f_{sat} \approx 1.2$. `G_HYDE_exposed` ist ein Term, der eine Energie proportional zu der zugänglichen Oberfläche berechnet. Dieser Term wird üblicherweise separat von der Bindungs-Energie betrachtet und dient als Filter-Term für falsch gedockte Lösungen. `G_HYDE_ligand` ist die Wechselwirkungs-Energie analog `G_match` für intramolekulare Ligand-Wechselwirkungen. Da im Wechselwirkungs-Term `G_HYDE_saturation` intramolekulare Wechselwirkungen explizit berücksichtigt werden, ist dieser Term für die Berechnung von HYDE obsolet.

Zusätzliche lassen sich Einstellungen zur genauen Berechnung der HYDE-Funktion vornehmen. Angegeben ist die jeweilige Standard-Einstellung.

```
add_surface_score     0       # [0/1]
surface_modifier      1 0
```

Wahlweise Berücksichtigung von `G_HYDE_exposed` im Gesamt-Score. `surface_modifier` erlaubt die gesonderte Gewichtung von apolaren und polaren Atomen. Bei Standaard-Einstellungen werden nur apolare Atome berücksichtigt.

```
hyde_access_mode      wsas    # [sas/wsas/wray/swsa]
```

Verwendeter Zugänglichkeits-Algorithmus. Diese Einstellung wirkt sich auf den verwendeten logP Parameter-Satz aus.

```
  hyde_desolvation_mode 3       # [3-6]

# 3: dG = dG_0 * plogp(occ) * (acc(c) - acc(s)) / acc(m)
# 4: dG = dG_0 * plogp(occ) * (acc(c) - acc(s)) / acc(s)
# 5: dG = dG_0 * plogp(occ) * (sqrt(acc(c)) - sqrt(acc(s)))
#         / sqrt(access(m))
# 6: dG = dG_0 * plogp(occ) * (sqrt(acc(c)) - sqrt(acc(s)))
#         / sqrt(access(s))
```

Bestimmt die exakte Gleichung des `G_HYDE_desolvation` Scoring-Terms. Der in der Arbeit geschilderten Funktion entspricht die Einstellung `3`. Die alternativen Einstellungen sind experimentellen Ursprungs und sollten nur zu Test-Zwecken verwendet werden.

```
  hyde_ligand_access    pose    # [pose/fix/mean]
```

Auswahl der Ligand-Konformation, die für die Berechnung der Zugänglichkeit in Lösung vewendet wird. `pose` entspricht der Konformation des Liganden in der Bindetasche, `fix` steht für die eingelesenen Koordinaten und durch Angabe von `mean` wird immer die in der logP-Datei gespeicherte mittlere Zugänglichkeit eines Atom-Typs verwendet.

```
  hyde_site_access      fix     # [fix/mean]
```

In analoger Weise wird die freie Zugänglichkeit der Rezeptor-Atome angegeben. Da der Rezeptor als starr angenommen wird, entfällt die Auswahl von `pose`.

```
  hyde_sat_mode         dev_root # [acc/acc_pen/dev_mean/dev_root]
```

Berechnung der Wechselwirkungs-Energie im `G_HYDE_saturation` Term (bzw. der Änderung Δsat_i in der Wasserstoff-Brücken-Sättigung). `dev_root` entspricht der Differenz der Quadratwurzeln von $\sum f_{dev}(i,j)$ im freien und im gebundenen Zustand. Mit `acc` wird die Energie bezüglich der Änderung in der Zugänglichkeit berechnet. Bei Auswahl von `acc_pen` fließt zusätzlich zur Zugänglichkeit der Abweichungsfaktor f_{dev} in die Formel mit ein und `dev_mean` normiert die Änderung in der Sättigung mit einem Ausdruck für die mittlere Anzahl an Wechselwirkungen. Diese Kapazität wird dem jeweiligen Eintrag `hb_cap` der `LOGP` Datei entnommen.

```
  expand_radius             1.2   # [0.0-inf]
  bond_range                2     # [0.0-inf]
```

Mit diesen beiden Parametern lässt dich die Berechnung von `G_HYDE_saturation` bei Verwendung von `acc` oder `acc_pen` genauer spezifizieren. `expand_radius` gibt einen vergrößerten Radius für das betrachtete Wechselwirkungs-Atom an, `bond_range` die Anzahl der Bindungen, die für eine Berücksichtigung von zusätzlichen Atomen verwendet wird. Eine Kombination dieser Einstellungen ist möglich.

Im letzten Teil der Geometrie-Datei sind die spezifischen Wechselwirkungs-Geometrien definiert. Als Beispiel sei die Geometrie einer Carbonyl-Funktion angeführt. Statt einer radialsymmetrischen Fläche werden zwei sphärische Rechteck-Flächen definiert.

```
@geometry co   #   new version, needs additional reference atom c3
# c1=o, c2=c, c3=c
radius  1.9
delta 1.2 1.2
s_area e 0 0 1 1 0 0 110 160 70 110
s_area e 0 0 1 1 0 0 200 250 70 110
exclude_from_clash      1
energy                  -2.35
charge_scaling          1.766   -0.15
distance_scaling        0.2     0.5
angle_scaling           0       30      1.0
```

A.1.5. AMINO und CONTACT

In diesen Dateien werden den Protein- und Ligand-Atomen funktionelle Gruppen zugewiesen. Über Subgraphen erfolgt die Zuordnung von Referenz-Atomen und Wechselwirkungs-Geometrien. Der wichtigste Unterschied zu den Original-Dateien ist, dass bei Atomen, denen eine Carbonyl-Funktion `h_acc co` zugewiesen wird, ein zusätzliches Referenz-Atom angegeben werden muss. Dies ist erforderlich, da die neu definierte Geometrie nicht mehr radialsymmetrisch ist.

A.1.6. GRAPHIC

In dieser Datei erfolgen Standard-Einstellungen für die graphische Repräsentation der Atome mit FlexV. Als zusätzliche Option für die Einfärbung von Atomen wurde mit HYDE eine Energie-basierte Repräsentation eingeführt.

```
#application HYDE
# colormode ATOM_ENERGY
colormode LIGAND                      ATOM_ENERGY
scalar E_STABILIZING                   -2.0
scalar E_NEUTRAL                       0.0
scalar E_DESTABILIZING                 2.0
color  COL_STABILIZING                 dark green
color  COL_NEUTRAL                     white
color  COL_DESTABILIZING               red
list  ACTIVE_ENERGY_TYPES              1 2 3 4 5 6 7 8
#end_application
```

Neben Energie-Grenzwerten für stabilisierende, destabilisierende und neutral beitragende Atome werden an dieser Stelle die verwendeten Farben definiert. Weiterhin erfolgt hier die Auswahl der für die Einfärbung verwendeten Scoring-Terme. Zusätzlich zu den HYDE-Termen können auch die FlexX Scoring-Terme `G_match`, `G_lipo`, `G_ambig` und `G_clash` für eine Farbdarstellung verwendet werden.

A.2. Parametrisierung (PLOGP Menü)

Initialisierung (INIT)

Syntax: `INIT` <accessibility> <descriptor> <ID> <filter ia geometry> <filter contact> <filter sybyl> <filter logp> <filter unknown> <filter property>

Beschreibung: Initialisierung der Datenstruktur zur Parametrisierung von Atom-Solvatations-Parametern. Auswahl des Zugänglichkeits-Algorithmus und der verwendeten Atom-Deskriptoren. Die Auswahl verschiedener Filter erlaubt eine automatische Beschränkung der eingelesenen Moleküle anhand gewünschter Eigenschaften.

<accessibility> erlaubt die Auswahl des Zugänglichkeits-Algorithmus (0: occurrence, 1: SAS, 2: WSAS, 3: WRAY, 4: SWSA). Durch Auswahl von 'occurrence' erfolgt ein einfaches Abzählen der Häufigkeit, mit der ein Deskriptor in einem Molekül zu finden ist. Die übrigen Algorithmen berechnen die zugängliche Oberfläche des Deskriptors relativ zur Kugel-Oberfläche.

Anhand von <descriptor> wird ein Molekül in atomare Einheiten unterteilt. Zur

Auswahl stehen 0: interaction geometry, 1: contact subgraph, 2: sybyl atom type und 3: atomic solvation parameter. 'interaction geometry' bezeichnet die FlexX-interne Wechselwirkungs-Geometrie (z.B. o/nh, coc oder ch3_phe), 'contact subgraph' den Kontakt-Typen (z.B. acceptor_C3OH), 'sybyl atom type' unterscheidet die Atome nach ihrem Sybyl Typ und 'atomic solvation parameter' steht für die in der `logp.dat` Datei definierten Atom-Typen. Eine Kombination der unterschiedlichen Klassifizierungs-Typen ist möglich.

<ID> ist eine einfache Text-Zeile, in der genauere Angaben zum betrachteten Datensatz spezifiziert werden können. Anhand der Filter-Terme können bereits beim Einlesen des Datensatzes Moleküle aufgrund unerwünschter Eigenschaften aussortiert werden. <filter ia geometry>, <filter contact> und <filter sybyl> erlauben die Auswahl der bereits genannten Deskriptoren. <filter logp> gibt einen zulässigen Werte-Bereich für den logP-Wert vor und <filter property> ermöglicht die Auswahl physiko-chemischer Eigenschaften durch eine Filter-Datei.

Hinzufügen eines Moleküls zum Datensatz (ADD)

Syntax: `ADD` <smiles> <name> <logP>

Beschreibung: Das Hinzufügen eines Moleküls zum Parametrisierungs-Datensatz erfolgt über die Eingabe der entsprechenden SMILES Repräsentation <smiles>. Die Angabe eines Namens ist optional. Wird <name> nicht angegeben, so wird das Molekül nach seinem SMILES-String benannt. <logP> ist der logP-Wert des Moleküls, der bei der Parametrisierung als freie Variable verwendet wird.

Voraussetzung: Mit `INIT` wurde zuvor eine Parametrisierungs-Matrix initialisiert.

Anzeigen einer Statistik (INFO)

Syntax: `INFO` <descriptor> <matrix> <logp> <accessibility>

Beschreibung: Ausgabe der wichtigsten Informationen zu dem eingelesenen Datensatz. Grundsätzlich ausgegeben wird eine Übersicht über die gewählten Einstellungen und die Anzahl der eingelesenen Moleküle inklusive einer Auflistung der im Datensatz vorhandenen Deskriptoren. Optional wird deren Häufigkeit und mittlere Zugänglichkeit und, wenn `SOLVE` bereits aufgerufen wurde, auch der berechnete parti-

elle logP-Wert angezeigt. Weiterhin möglich ist die Ausgabe der Parametrisierungs-Matrix in Tabellen-Form sowie die Ausgabe von Häufigkeits-Verteilungen mit Informationen über die verwendeten logP-Werte und die Zugänglichkeiten einzelner Deskriptoren.

Voraussetzung: `INIT` wurde zuvor aufgerufen.

Löschen des Datensatzes (DELETE)

Syntax: `DELETE`

Beschreibung: Löscht die Parametrisierungs-Matrix inklusive der eingelesenen Moleküle aus dem Speicher. Beim Verlassen des Programms wird dieser Punkt automatisch ausgeführt.

Anzeigen des Datensatzes (DATASET)

Syntax: `DATASET`

Beschreibung: Ausgabe des Datensatzes im Format <smiles> <name> <logP> entsprechend den bei `ADD` eingegebenen Werten. Moleküle, die durch eines der Filter-Kriterien aussortiert worden sind, erscheinen nicht in der Auflistung. Die Ausgabe kann durch das Kommando `SELOUTP` in eine Datei umgeleitet werden.

Anzeigen der Zugänglichkeits-Matrix (MATRIX)

Syntax: `MATRIX` <normalize>

Beschreibung: Ausgabe der Parametrisierungs-Matrix. Die Zeilen entsprechen den eingelesenen Molekülen, die Spalten stehen für die Deskriptoren. In der letzten Spalte stehen die experimentellen logP-Werte. Die einzelnen Felder sind durch Leerzeichen getrennt. Durch das Kommando `SELOUTP` kann die Ausgabe in eine Datei umgeleitet werden. Wird <normalize> mit 'yes' beantwortet, werden die Elemente der Matrix bezüglich der Zugänglichkeiten der Deskriptoren normalisiert.

Automatische Parametrisierung (SOLVE)

Syntax: `SOLVE` <algorithm> <data reduction> <data analysis>

Beschreibung: Automatische Bestimmung der Koeffizienten (plogP-Werte) entweder über eine multiple lineare Regression (unter Verwendung der GNU Scientific Library) oder anhand einer einfachen Simplex-Optimierung. Vor der Regressions-Analyse ist es möglich, den Datensatz anhand spezifischer Eigenschaften der Deskriptoren zu reduzieren. In einem ersten Schritt können Moleküle aus dem Datensatz entfernt werden. Zur Auswahl stehen: a = all, s = selection, e = exclusion, p = properties.

Bei Auswahl von 'all' wird der gesamte Datensatz in der Parametrisierung verwendet. 'selection' ermöglicht eine Auswahl der Deskriptoren, die im finalen Datensatz enthalten sein dürfen. Moleküle, in denen von der Auswahl abweichende Deskriptoren enthalten sind, werden bei der Parametrisierung nicht beachtet. Umgekehrt ist über 'exclusion' eine Auswahl von Deskriptoren möglich, die nicht im Datensatz enthalten sein dürfen. Mit 'properties' kann der Datensatz bezüglich der Mindest-Häufigkeit oder der minimalen mittleren Zugänglichkeit von Deskriptoren eingeschränkt werden.

Anschließend an den Reduktions-Schritt können einzelne Deskriptoren explizit von der Analyse ausgenommen werden, sie erhalten einen plogP-Wert von 0. Die Ergbnisse der Parametrisierung werden gespeichert und können anschließend über das `INFO` Kommando angezeigt werden.

Voraussetzung: Mit `INIT` wurde eine Matrix initialisiert und über `ADD` wurden dem Datensatz Moleküle hinzugefügt.

Reduktion des Datensatzes (SHRINK)

Syntax: `SHRINK` <# molecules> <delete descriptors> <backup>

Beschreibung: Reduktion des Datensatzes um <# molecules> Moleküle. Es werden bevorzugt Moleküle mit häufig repräsentierten Deskriptoren gelöscht. Der Original-Datensatz kann für eine spätere Verwendung temporär gespeichert werden. Wird <delete descriptors> mit 'no' beantwortet, bricht der Reduktions-Algorithmus ab, sobald durch das Löschen eines weiteren Moleküls ein Deskriptor verloren gehen würde.

Voraussetzung: Mit `INIT` wurde eine Matrix initialisiert und über `ADD` wurden dem Datensatz Moleküle hinzugefügt.

Clustern des Datensatzes (CLUSTER)

Syntax: `CLUSTER` <# molecules> <backup>

Beschreibung: Reduktion des Datensatzes auf eine Größe von <# molecules> Molekülen. Die Auswahl der Moleküle erfolgt unter Anwendung des Bawden-Lajiness Algorithmus. Ausgehend von einer leeren Menge werden aus dem Original-Datensatz schrittweise Moleküle ausgewählt, die zu den bereits ausgewählten eine maximale Unähnlichkeit besitzen. Die (Un)Ähnlichkeit wird anhand des Tanimoto-Koeffizienten bestimmt und bezieht sich auf die An- bzw. Abwesenheit von Deskriptoren in den Molekülen. Der Original-Datensatz kann für eine spätere Verwendung temporär gespeichert werden.

Voraussetzung: Mit `INIT` wurde eine Matrix initialisiert und über `ADD` wurden dem Datensatz Moleküle hinzugefügt.

Austausch mit Backup-Datensatz (SWAP)

Syntax: `SWAP`

Beschreibung: Der temporär gespeicherte Datensatz wird verwendet. Stattdessen wird der aktuelle Datensatz in den temporären Speicher geschoben.

Voraussetzung: Eine mit `INIT` initialisierte Matrix ist vorhanden und ein Datensatz liegt im temporären Speicher. Dies ist der Fall nach Verwendung von `SHRINK` oder `CLUSTER`.

Wiederherstellen des Datensatzes (RESTORE)

Syntax: `RESTORE`

Beschreibung: Der aktuelle Datensatz wird durch den temporär gespeicherten Datensatz ersetzt. Die aktuellen Daten werden gelöscht.

Voraussetzung: Eine mit `INIT` initialisierte Matrix ist vorhanden und ein Datensatz liegt im temporären Speicher. Dies ist der Fall nach Verwendung von `SHRINK` oder `CLUSTER`.

B. plogP Parameter

Auflistung der in HYDE verwendeten Atom-Typen mit den in der Regressionsanalyse ermittelten Parametern. Die mittleren Zugänglichkeiten *acc* und Parameter *plogP* wurden für drei unterschiedliche Zugänglichkeits-Algorithmen bestimmt.

Atom Typ ID	**WSAS**		**WRAY**		**SWSA**	
	acc	*plogP*	*acc*	*plogP*	*acc*	*plogP*
C_sp3_ssss_H3	0.54	0.74	0.67	0.74	0.67	0.67
C_sp3_ssss_H2	0.24	0.48	0.39	0.48	0.46	0.50
C_sp3_ssss_H1	0.08	0.12	0.16	0.21	0.26	0.31
C_sp3_ssss_H0	0.00	0.00	0.00	0.00	0.09	0.56
C_sp2_ssd_H2	0.56	0.77	0.66	0.70	0.70	0.58
C_sp2_ssd_H1	0.23	0.34	0.37	0.38	0.46	0.39
C_sp2_ssd_H0	0.02	0.00	0.05	0.00	0.25	0.49
C_sp2_saa_H1	0.26	0.43	0.39	0.42	0.47	0.36
C_sp2_saa_H0	0.02	0.00	0.09	0.00	0.21	0.23
C_sp2_aaa_H0	0.03	0.10	0.13	0.15	0.21	0.27
C_sp2_ggg_H0	0.10	0.10	0.23	0.15	0.45	0.27
C_sp_dd_H0	0.27	0.00	0.27	0.00	0.37	0.16
C_sp_st_H1	0.63	0.18	0.69	0.20	0.78	0.14
C_sp_st_H0	0.18	0.00	0.35	0.00	0.53	0.16
N_sp3_sss_H2	0.22	-1.31	0.28	-1.30	0.27	-1.49
N_sp3_sss_H1	0.10	-1.72	0.18	-1.78	0.17	-1.82
N_sp3_sss_H0	0.03	-1.43	0.08	-1.77	0.08	-1.68
N_sp3_ssss_H3	0.21	-1.31	0.25	-1.34	0.25	-1.49
N_sp3_ssss_H2	0.10	-1.72	0.16	-1.82	0.15	-1.82
N_sp3_ssss_H1	0.01	-1.43	0.07	-1.81	0.06	-1.68
N_sp3_ssss_H0	0.00	0.00	0.00	0.00	0.01	0.00
N_sp2_sd_H1	0.24	-1.05	0.18	-1.04	0.10	-0.53
N_sp2_sd_H0	0.06	-1.05	0.11	-1.04	0.16	0.00
N_sp2_ssd_H2	0.15	-1.05	0.18	-1.04	0.30	0.48
N_sp2_ssd_H1	0.06	-1.05	0.09	-1.04	0.11	-1.04
N_sp2_ssd_H0	0.00	0.00	0.00	0.00	0.26	-1.04
N_sp2_saa_H1	0.09	-0.55	0.10	-0.45	0.09	-1.04
N_sp2_saa_H0	0.04	0.00	0.05	0.00	0.17	-1.04
N_sp2_snn_H0	0.11	0.00	0.11	0.00	0.08	-1.04

Fortsetzung auf der nächsten Seite

Atom Typ ID	**WSAS**		**WRAY**		**SWSA**	
	acc	*plogP*	*acc*	*plogP*	*acc*	*plogP*
N_sp2_aa_H0	0.08	-1.05	0.11	-1.04	0.00	0.00
N_sp2_ss_H1	0.24	-0.95	0.26	-0.96	0.09	-1.21
N_sp2_ss_H0	0.04	-1.15	0.10	-1.15	0.30	-1.10
N_sp2_ssm_H2	0.16	-0.94	0.20	-0.96	0.19	-0.67
N_sp2_ssm_H1	0.04	-1.15	0.09	-1.15	0.09	-1.11
N_sp2_ssm_H0	0.00	0.00	0.03	0.00	0.00	0.00
N_sp2_sss_H2	0.15	-0.94	0.19	-0.96	0.17	-1.21
N_sp2_sss_H1	0.06	-1.15	0.10	-1.15	0.08	-1.10
N_sp2_sss_H0	0.00	0.00	0.03	0.00	0.00	0.00
N_sp2_ssg_H2	0.13	-0.94	0.17	-0.96	0.16	-1.21
N_sp2_ssg_H1	0.05	-1.15	0.09	-1.15	0.08	-1.10
N_sp2_ssg_H0	0.02	0.00	0.02	0.00	0.00	0.00
N_sp_t_H0	0.55	-0.84	0.37	-0.86	0.15	-1.21
N_sp_st_H1	0.10	-0.84	0.10	-0.86	0.10	-1.21
N_sp_st_H0	0.17	0.00	0.17	0.00	0.00	0.00
O_sp3_ss_H2	0.36	-1.38	0.36	-1.38	0.36	-1.38
O_sp3_ss_H1	0.21	-1.13	0.28	-1.25	0.26	-1.43
O_sp3_ss_H0_AA	0.11	-1.38	0.16	-1.38	0.15	-1.45
O_sp3_ss_H0_Aa	0.11	-0.48	0.16	-0.52	0.15	-0.64
O_sp3_ss_H0_aa	0.30	0.00	0.30	0.00	0.33	0.19
O_sp3_s_H0	0.37	-1.13	0.43	-1.13	0.40	-1.43
O_sp2_d_H0	0.17	-0.80	0.24	-0.98	0.20	-1.78
O_sp2_a_H0	0.27	-0.97	0.35	-1.12	0.33	-1.60
O_sp2_n_H0	0.38	0.00	0.21	0.00	0.31	-0.34
O_sp2_aa_H0	0.14	-0.63	0.16	-0.34	0.16	-0.42
O_sp2_saa_H1	0.11	-0.54	0.11	-0.33	0.11	-0.42
O_sp2_saa_H0	0.00	0.00	0.00	0.00	0.00	0.00
Fluorine	0.36	0.30	0.56	0.29	0.49	0.00
Chlorine	0.40	0.56	0.52	0.52	0.59	0.41
Bromine	0.58	0.68	0.65	0.68	0.69	0.55
Iodine	0.65	0.90	0.75	0.96	0.75	0.81
Hydrogen	0.00	0.00	0.00	0.00	0.00	0.00
Silicon	0.00	0.00	0.00	0.00	0.00	0.00
Phosphorus	0.00	0.00	0.00	0.00	0.00	0.00
S_sp3_ss_H1	0.64	0.00	0.64	0.00	0.00	0.00
S_sp3_ss_H0	0.42	0.00	0.42	0.00	0.00	0.00
Sulfur	0.00	0.00	0.00	0.00	0.00	0.00
S_sp2_d_H0	0.60	0.00	0.60	0.00	0.00	0.00
Metal	1.00	-2.00	1.00	-2.00	1.00	-2.00
O/N_bridge	0.00	0.00	0.00	0.00	0.00	0.00

C. DUD Datensatz

Übersicht über die Zahl der aktiven Verbindungen und Decoys im DUD Datensatz [68]. Die in Klammern genannten PDB-Dateien sind in der PDB durch besser aufgelöste Strukturen ersetzt worden. Anschließend ist eine Übersicht über initiale Enrichment-Ergebnisse gegeben. Dargestellt sind die Ergebnisse der Bewertung mit FlexX in schwarz und HYDE in rot.

Protein	PDB Code	# Aktive	# Decoys
	Nukleäre Hormonrezeptoren		
AR	2ao6 (1xq2)	79	2854
$ER_{agonist}$	1l2i	67	2570
$ER_{antagonist}$	3ert	39	1448
GR	1m2z	78	2947
MR	2aa2	15	636
PPARγ	1fm9	85	3127
PR	1sr7	27	1041
RXRα	1mvc	20	750
	Kinasen		
CDK2	1ckp	72	2074
EGFr	1m17	475	15996
FGFr1	1agw	120	4550
HSP90	1uy6	37	979
P38 MAP	1kv2	454	9141
PDGFrb	model	170	5980
SRC	2src	159	6319
TK	1kim	22	891
VEGFr2	1vr2	88	2906
	Serin Proteasen		
FXa	1f0r	146	5745
Thrombin	1ba8	72	2456
Trypsin	1bju	49	1664

Fortsetzung auf der nächsten Seite

Protein	PDB Code	# Aktive	# Decoys
	Metalloenzyme		
ACE	1o86	49	1797
ADA	1ndw (1stw)	39	927
COMT	1hld	11	468
PDE5	1xp0	88	1978
	Folatenzyme		
DHFR	3dfr	410	8367
GART	1c2t	40	879
	sonstige Enzyme		
AChE	1eve	107	3892
ALR2	1ah3	26	995
AmpC	1xgj	21	786
COX-1	1q4g (1p4q)	25	911
COX-2	1cx2	426	13289
GPB	1a8i	52	2140
HIVPR	1hpx	62	2038
HIVRT	1rt1	43	1519
HMGR	1hw8	35	1480
InhA	1p44	86	3266
NA	1a4g	49	1874
PARP	1efy	35	1351
PNP	1b8o	50	1036
SAGG	1a7a	33	1346

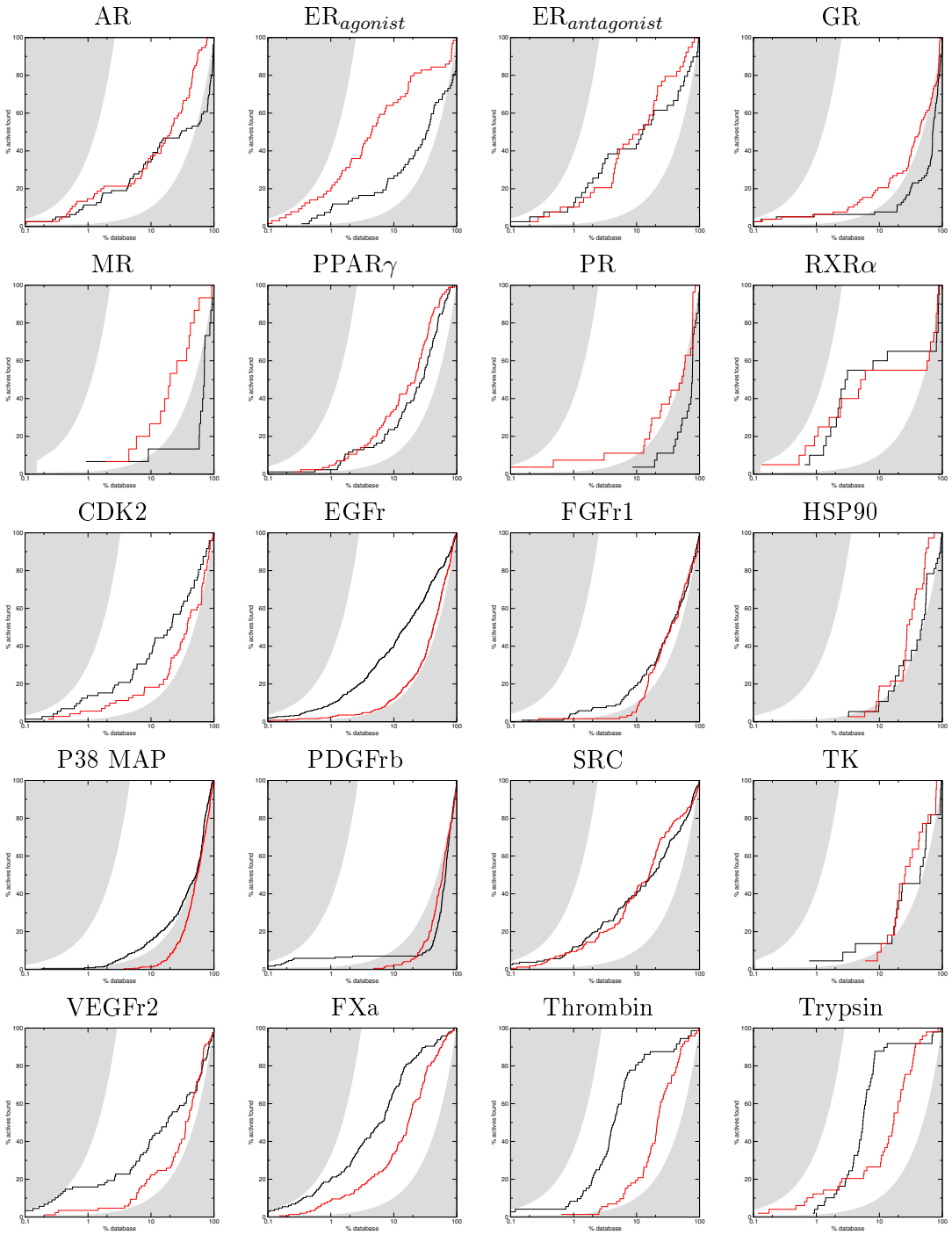


Abbildung C.1.: Enrichment Ergebnisse DUD-Datensatz. Bewertung durch HYDE (rot) und FlexX (schwarz).

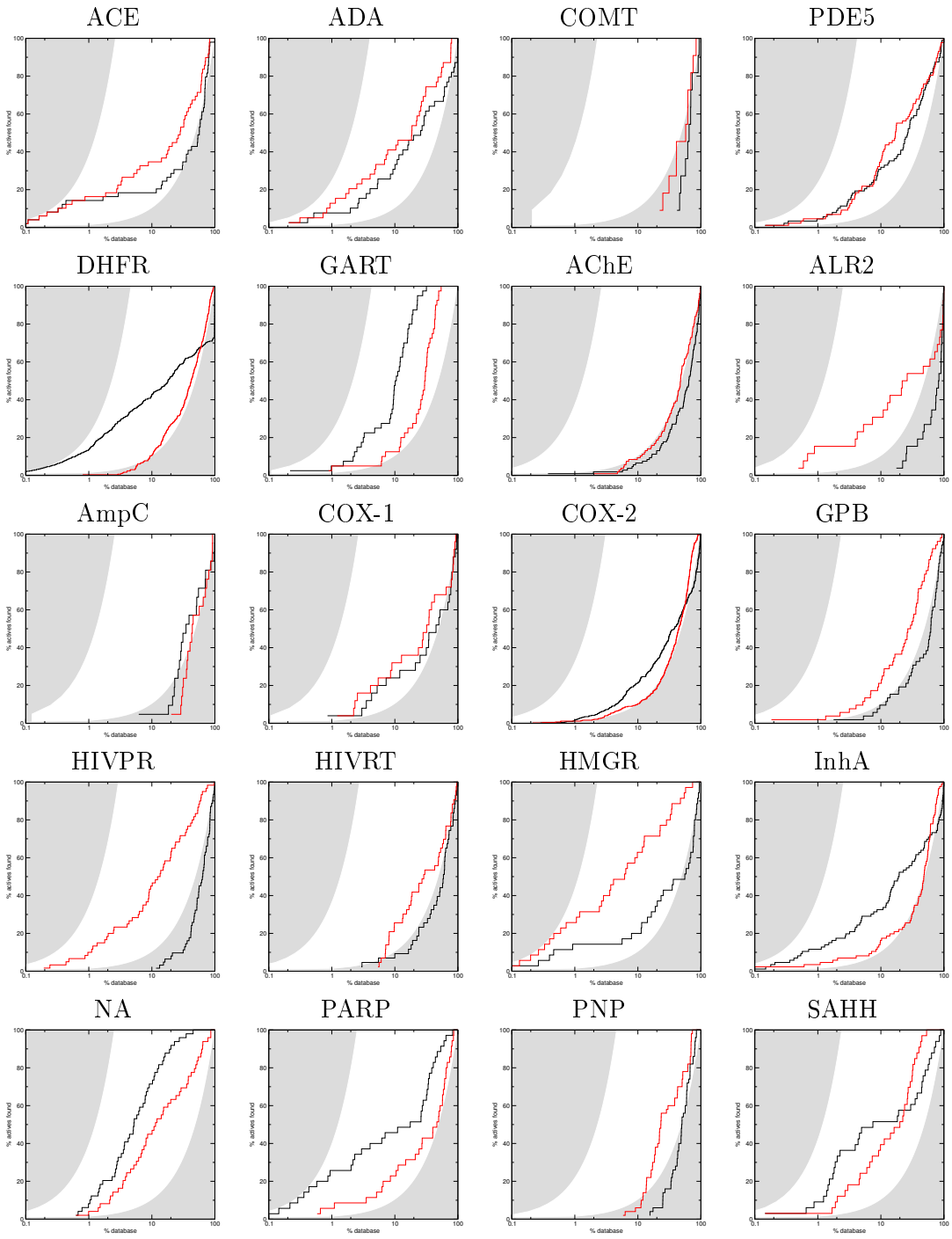


Abbildung C.2.: Enrichment Ergebnisse DUD-Datensatz (Forts.)

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Hamburg, 16. Februar 2009

(Ingo Reulecke)