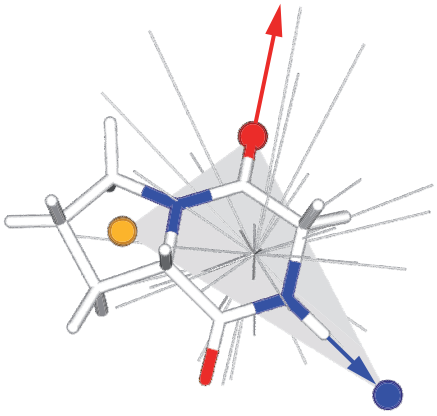
Modellierung molekularer und makromolekularer Zustände im geleiteten strukturbasierten virtuellen Screening



Dissertation zur Erlangung des Doktorgrades

an der Fakultät für Mathematik, Informatik und Naturwissenschaften

der Universität Hamburg

Fachbereich Informatik

vorgelegt von

Angela Maren Henzler

Hamburg, 2015

Tag der Disputation: 18.12.2015

Folgende Gutachter empfehlen die Annahme der Dissertation:
Prof. Dr. Matthias Rarey
Prof. Dr. Johannes Kirchmair

Kurzfassung

Strukturbasierte virtuelle Screening-Methoden kommen in den frühen Phasen des rationalen Wirkstoffentwurfs zum Einsatz. Sie setzen die dreidimensionale Struktur einer Proteinbindetasche voraus, für die passende kleine Moleküle aus einer Bibliothek identifiziert werden sollen. Bei der praktischen Anwendung sind neben der Zielstruktur häufig jedoch auch Liganden bekannt. Dieser Umstand stellt Erwartungen an das Screening-Resultat. Die erhaltenen Treffer sollen ein bereits bekanntes Bindungsmuster und/oder spezifische physikochemische Eigenschaften erfüllen.

Die vorliegende Arbeit beschreibt die Entwicklung von cRAISE, eine durch den Anwender kontrollierbare, strukturbasierte virtuelle Screening-Methode. Basierend auf dem Ansatz von RAISE (RApid Index-based Screening Engine) realisiert sie geleitetes Protein-Ligand-Docking, um den Inhalt einer Molekülbibliothek zu bewerten. Sie propagiert eine Pharmakophorhypothese und/oder ein Molekülprofil, um sich ausschließlich auf Moleküle zu konzentrieren, die den gegebenen Anforderungen gerecht werden. Ist die Hypothese adäquat formuliert, bietet der geleitete Ansatz neben einer drastischen Beschleunigung des Screening-Prozesses auch die Chance, Bindungsmodusvorhersagen und die Anreicherung aktiver Moleküle entscheidend zu verbessern.

Da Molekül- und Proteindateiformate nur einen von mehreren molekularen bzw. makromolekularen Zuständen darstellen können, sind insbesondere pharmakophorbasierte Docking-Strategien vom Eingabezustand abhängig. Sie repräsentieren beide Komponenten durch entscheidende, bindungsvermittelnde Merkmale, die durch Tautomerie und (De-)protonierung direkt beeinflusst sind. cRAISE integriert deshalb Protomerfreiheitsgrade von Protein und Ligand. Neben dem Bindungsmodus, sagt der neuartige Multizustandsansatz unabhängig von der Eingabe den angenommen Zustand im Protein-Ligand-Komplex voraus. Durch die indexbasierte Auswertung möglicher Protein-Ligand-Zustandskombinationen, bleibt die Methode trotz der erhöhten Anzahl von Freiheitsgraden effizient zum strukturbasierten Screening umfangreicher Molekülbibliotheken anwendbar.

Abstract

Methods for structure-based virtual screening are employed in early stages of rational drug design. They require a threedimensional structure of a protein binding pocket in order to extract matching small molecules out of a library. In practical applications however, beyond the structure of the protein target there is often knowledge about its bound ligands. This circumstance requires screening results to satisfy initial expectations. The obtained hits have to fulfill an already known binding pattern and/or lie within a specific physicochemical property range.

This thesis describes the development of cRAISE, a structure-based virtual screening method which is controllable by the user. Based on the RAISE (RApid Index-based Screening Engine) approach, it implements guided protein-ligand docking to assess the library content. It propagates a given pharmacophore hypothesis and/or a library profile to focus solely on molecules that meet the given conditions. If the hypothesis is chosen appropriately, the guided approach offers the chance to drastically accellerate the screening process, to improve binding mode predictions and to significantly enhance the enrichment of active compounds.

Since molecule and protein file formats depict only one of several possible molecular or macromolecular states, particularly, pharmacophore-based docking strategies depend on the input state. Such methods represent proteins and ligands by essential binding mediating features which are, however, under the influence of tautomerism and protonation. cRAISE integrates protomer degrees of freedom to resolve state dependencies. In addition to the binding mode, the novel multi-state docking approach independently predicts the adopted state in a protein-ligand complex. With the aid of an efficient index-based evaluation of possible receptor-ligand state combinations and despite an increased number of degrees of freedom, the approach remains practicable for the screening of large-scaled molecular libraries.

Danksagung

Diese Arbeit wäre ohne die Unterstützung vieler undenkbar gewesen. In erster Linie möchte ich mich bei meiner Familie bedanken, die mir die Freiheit gegeben hat, mich auf mein Projekt zu konzentrieren und es mir in keinster Weise übel nimmt, dass sie all die Jahre viel zu kurz gekommen ist. Außerdem kann ich mich glücklich schätzen, so viele wahre Freunde zu haben, auf die ich mich auch über Hunderte von Kilometern hinweg hundertprozentig verlassen kann. Sie haben mich in schwierigen Zeiten moralisch unterstützt, mich zum Auftanken gelegentlich aus der Arbeit gerissen und immer fest an mich geglaubt. Ihr seid der eigentliche Ursprung meiner Kraft.

Ich bin äußerst dankbar, dass meine Promotionsjahre von so großartigen Kollegen begleitet wurde. Explizit sollen die erwähnt werden, die direkt zu dieser Arbeit beigetragen haben: Zu aller erst Sascha Urbaczek, der so tiefen Einblick in diese Arbeit hatte, dass es mit ihm möglich war neue Ideen zu diskutieren und weiterzuentwickeln. Die Beiträge von Matthias Hilbig, Stefan Bietz, Christin Schärfer und Nadine Schneider halfen Teilprobleme zu lösen und ermöglichten es mir so, einen Schritt weiter zu gehen. Danke an Jochen Schlosser, der mich bei der Einarbeitung unterstützte, Andrea Volkamer, Mathias von Behren und Karen Schomburg, die meine Entwicklungen in andere Anwendungen integrierten und damit zu qualitativ besseren und generalisierten Methoden beitrugen. Hier einzelne Namen zu nennen, soll aber den Beitrag anderer in keinster Weise schmälern. Deshalb Dank an alle für die regen Diskussionen, die ehrliche Kritik und dafür, dass ihr keine Einzelkämpfer seid. Es ist nicht selbstverständlich eine Arbeitsatmosphäre zu finden, in der jeder ein offenes Ohr und Interesse für den anderen hat.

Danke Matthias dafür, dass Du diese tolle Arbeitsumgebung der offenen Türen geschaffen und mir die Möglichkeit zu dieser Arbeit gegeben hast. Sie ist mir so sehr ans Herz gewachsen. Danke für die Erfahrung, die ich in deiner Arbeitsgruppe, national und international sammeln durfte, für deinen Rat, deinen Rückhalt und dein Vertrauen in mich.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Symbol- und Abkürzungsverzeichnis

ΔG	Änderung der Gibbs-Energie	12
ΔH	Enthalpieänderung	12
ΔS	Entropieänderung	12
E	Potentielle Energie	39
F	Kraft	39
K_B	Bindungskonstante	12
K_D	Dissoziationskonstante	12
K_I	Inhibierungskonstante	12
R	Gaskonstante	12
T	Absolute Temperatur	12
1D	Eindimensional	30
2D	Zweidimensional	30
3D	Dreidimensional	8
Å	Angström	14
ACS	American Chemical Society	147
ADME	Absorption, Distribution, Metabolism, Excretion	9
ALDR	Aldosereduktase	156
AS	Aminosäure	10
AUC	Area Under the Curve	149
BLOB	Binary Large Objekt	82
cLogP	Berechneter LogP	17
CPU	Central Processing Unit	97
cRAISE	Complex-based RAISE	3
Da	Dalton	26
DNA	Desoxyribonukleinsäure	9
DUD	Directory of Useful Decoys	56

ECFP	Extended-Connectivity Fingerprint	30
FCFP	Function-Class Fingerprint	30
GB	Gigabyte	122
HID	Neutrales ND1-Tautomer des Histidins	23
HIE	Neutrales NE2-Tautomer des Histidins	23
HIP	Protonierter Zustand von Histidin	23
HPC	High Performance Computing	152
HTS	High throughput screening	8
InChI	IUPAC International Chemical Identifier	64
IUPAC	International Union of Pure and Applied Chemistry	14
J	Joule	12
K	Kelvin	12
LBVS	Ligandbasiertes virtuelles Screening	1
LogP	Logarithmus des Octanol/Wasser-Verteilungskoeffizienten	17
MCP	Max-Clique-Problem	28
MCS	Maximum Common Substructure	28
MCSS	Multiple Copy Simultaneous Search	51
MD	Moleküldynamik	36
MEP	Molekulares elektrostatisches Potential	47
MIF	Molekulares Interaktionsfeld	47
MM	Molekülmechanik	39
MOL2	Tripos Sybyl MOL2 Format	64
MW	Molekulargewicht	17
NMR	Nuclear Magnetic Ressonance	10
PBVS	Pharmakophorbasiertes virtuelles Screening	1
PDB	Protein Data Bank	10
PMF	Potential of Mean Force	41
PSA	Polar Surface Area	17
RAISE	RApid Indexbased Screening Engine	3
RMSD	Root Mean Square Deviation	81
RNA	Ribonukleinsäure	9
ROC	Receiver Operating Characteristics	148
SASA	Solvent Accessible Surface Area	31
SBVS	Strukturbasiertes virtuelles Screening	1
SDF	Structure Data File	64
SIFt	Struktureller Interaktionsfingerabdruck	52
SLN	Sybyl Line Notation	64

SMARTS	Smiles Arbitrary Target Specification	77
SMILES	Simplified Molecular Input Line Entry System	30
SQL	Structured Query Language	82
TFD	Torsion Fingerprint Deviation	81
TPSA	Topological Polar Surface Area	85
VS	Virtuelles Screening	1
VSC	Valence State Combination	69
VSEPR	Valence Shell Electron Pair Repulsion	65
WAH	Word Alligned Hybrid code	97
XML	Extensible Markup Language	84
ZINC	Zinc Is Not Commercial	55
ZNS	Zentrales Nervensystem	18

1 Einleitung

1.1 Motivation

In der pharmazeutischen Industrie haben sich computergestützte Methoden als kostengünstige Werkzeuge zum rationalen Entwurf von Wirkstoffen etabliert. In frühen Stadien des Entwicklungsprozess werden u. a. Verfahren zum virtuellen Screening (VS) eingesetzt. Ihr Ziel ist es, umfangreiche Molekülbibliotheken auf eine Submenge potentieller Kandidaten zu beschränken, sodass eine aufwändigere experimentelle Bewertung erfolgen kann. In den letzten Jahrzehnten haben sich drei wesentliche Entwicklungsstränge abgezeichnet: ligand-, struktur- und pharmakophorbasierte VS-Ansätze. Ligandbasierte Methoden (LBVS) werden bei fehlender Proteinstruktur genutzt, um zu bereits bekannten bioaktiven Molekülen ähnliche Verbindungen in einer Bibliothek zu identifizieren. Für ein strukturbasiertes Screening (SBVS) muss dagegen die dreidimensionale Struktur eines Proteins gegeben sein, zu der die Bindungsaffinität einzelner Moleküle bewertet wird. Pharmakophorbasierte Ansätze (PBVS) verlangen eine zuvor erstellte Pharmakophorhypothese, die von einem Protein und/oder Liganden abgeleitet wird. Durch verbesserte Verfahren zur experimentellen Bestimmung dreidimensionaler Proteinstrukturen ist es zukünftig wahrscheinlicher, dass Protein- und Ligandinformationen gegeben sind. Prinzipiell kann so jegliche VS-Strategie zum Einsatz kommen. Allerdings muss zuvor abgewägt werden, welche zur Anwendung geeignet ist, da die Ansätze bezüglich ihrer Bewertungsleistung, Selektivität und ihrer Effizienz individuelle Stärken und Schwächen aufweisen. Ein LBVS ermöglicht vielfältige Anfragen auf Basis simpler Moleküldeskriptoren. Im Vergleich zum SBVS ist es effizienter auf umfangreichen Bibliotheken anwendbar, jedoch weniger selektiv, da es keine Restriktion durch eine Proteinstruktur erfährt. SBVS-Methoden beschränken dagegen effektiv die Molekülbibliothek. Ihre aufwändigere Bewertung ist aufschlussreicher, aber nur in eingeschränktem Umfang einsetzbar. Ein PBVS realisiert effizient den Abgleich weniger Pharmakophormerkma-

le. Es nutzt nicht oder nur marginal die restriktive Proteininformation. Die Bewertung auf Basis weniger Merkmale ist rudimentär und erschwert die Priorisierung der Moleküle. PBVS-Methoden bieten jedoch die Möglichkeit, das Screening zu lenken und direkt auf das Resultat einzuwirken. Um die einzelnen Schwächen durch die Stärken anderer auszugleichen, werden die Ansätze zunehmend in Synergie eingesetzt und verschiedenartige Komponenten konkateniert. Da existierende VS-Methoden jedoch nicht mit Hinblick auf ein Zusammenspiel entwickelt wurden, verbleibt die Frage, wie die Protein- und Ligandinformation zugleich genutzt werden kann, sodass die Optimalität der Bewertungsleistung, Selektivität und Effizienz gewährleistet wird.

Durch Protonierung und Tautomerie nehmen Proteine und Moleküle einen von mehreren möglichen Zuständen an. Molekül- und Proteindateiformate können aber nur einen Zustand darstellen. Dieser prägt direkt charakteristische, bindungsvermittelnde Merkmale, die VS-Methoden häufig für eine reduzierte interne Repräsentation der betrachteten Komponenten verwenden. Sie verlassen sich somit auf eine adäquat präparierte Darstellung von Protein und/oder Molekül in Eingabedateien. Da der optimale Zustand vorab allerdings nicht absehbar ist, ist es gängige Praxis geworden, Bibliotheken initial zu präparieren und verschiedene Molekülzustände anzubieten. Das VS soll dann den passenden Zustand wählen. Für Proteine ist im SBVS eine analoge Verfahrensweise prinzipiell möglich. In der Praxis werden sie aber nur einmalig präpariert, indem ein mutmaßlich wahrscheinlicher Proteinzustand angenommen wird. In Eingabedateien bereitgestellte molekulare und/oder makromolekulare Zustände bringen jedoch entscheidende Nachteile mit sich, die die Genauigkeit, Konsistenz und die Effizienz des VS-Prozesses betreffen: Werden nur suboptimale Zustände bereitgestellt, können potentiell gute Kandidaten schlechter bewertet oder sogar völlig übersehen werden. Unterschiedliche Präparierungsprotokolle können zu unterschiedlichen Vorhersagen führen, auch wenn die Zustände eigentlich ein und dasselbe Molekül bzw. Protein beschreiben. Zusätzlich bereitgestellte Zustände expandieren substantiell Molekülbibliotheken. Da sie sich nur geringfügig unterscheiden, ist der Großteil der Information allerdings redundant. Im VS führt dies dazu, dass ähnliche interne Repräsentationen aufgebaut werden, die wiederholt ähnliche Berechnungen verursachen. Dies benötigt unnötigerweise erhöhte Rechenkapazitäten, die den Durchsatz in VS-Kampagnen mindern. Besonders im SBVS erhöht sich unter Hinzunahme von Proteinzuständen der Aufwand derart, dass das Verfahren nicht mehr praktikabel ist. Um den Nachteilen einer präparierenden Zustandsenumeration zu entgehen, müssen Freiheitsgrade, die molekulare und makromolekulare Zustände beschreiben, in die VS-Methode integriert und der optimale Zustand vorhergesagt werden. Bislang existieren keine Verfahren, die dies bewerkstelligen und die Anwendbarkeit zum Screening umfangreicher Molekülbibliotheken gewährleisten.

1.2 Zielsetzung

Das Ziel dieser Dissertation ist die Entwicklung eines SBVS-Verfahrens, das

- mittels einer Pharmakophorhypothese oder eines Molekülprofils ein durch den Anwender pharmakophor- bzw. molekülprofilgeleitetes SBVS unterstützt
- und die Freiheitsgrade integriert, die aus möglichen Protonierungszuständen und Tautomeren von Protein und Ligand resultieren.

Eine Lösung, die es erstmalig erlaubt diese Aspekte in einem umfangreichen VS zu betrachten, ist in cRAISE implementiert.

Die Methode soll insbesondere auf die Anforderungen eingehen, die bei gegebener Protein- und Ligandinformation entstehen. Die Basis bildet ein SBVS-Ansatz. Im Vergleich zu anderen Strategien schließen dessen Docking-Berechnungen unpassende Moleküle effektiv aus und ermöglichen eine aufschlussreiche Bewertung. Sind anhand zusätzlich gegebener Ligandinformation Bindungsmuster bekannt, soll ein integrierter PBVS-Ansatz verfolgt werden. Die zusätzlich bereitgestellte Information in Form einer Pharmakophorhypothese soll an den Docking-Prozess propagiert werden, sodass sich dieser ausschließlich auf Moleküle konzentrieren kann, die den gegebenen Anforderungen gerecht werden. Ist die Hypothese adäquat formuliert, soll der pharmakophorgeleitete Ansatz neben einer Beschleunigung des Prozesses auch die Chance bieten, Bindungsmodusvorhersagen und die Anreicherung bioaktiver Moleküle zu verbessern. Die Konzepte, der am Zentrum für Bioinformatik der Universität Hamburg entwickelten SBVS-Werkzeuge TrixX[1] und TrixX-BMI[2] dienten als Vorlage für die Neuentwicklung. Sie verfolgen die Idee eines zweigeteilten, indexbasierten SBVS. Es hat den Vorteil, dass umfangreiche Molekülbibliotheken effizient durchsucht werden können. Der gravierende Nachteil ist jedoch, dass die einmalig indexierte Molekülbibliothek, Information für jegliches erdenkliche Anfrageszenario bereitstellen und für diverse VS-Läufe statisch vorgehalten werden muss. Nur dann lohnen sich die Investitionen, die zu ihrem Aufbau getätigt werden müssen. In unterschiedlichen VS-Projekten ergeben sich allerdings unterschiedliche Anforderungen an die physikochemische Eigenschaften, die Konstitution und Topologie der extrahierten Moleküle. Die Neuentwicklung soll deshalb die Möglichkeit bieten, auch Molekülprofile an den Prozess zu propagieren, um lediglich Moleküle mit vorab festgelegten Eigenschaften zu identifizieren. Die simultane Filterung der Bibliothek soll sich nicht nachteilig auf die Effizienz des VS auswirken und vielfältige Anfragen auf einer umfangreichen, statischen Bibliothek unterstützen, ohne dass dies eine Neuetablierung der Indexstruktur erfordert.

TrixX und TrixX-BMI wurden unter Nutzung der Flex*-Software-Bibliothek entwickelt. Die Neuentwicklung sollte dagegen auf dem chemischen Modell von Naomi[3] und dessen assoziierter Software-Bibliothek beruhen. Naomi bildet den Rahmen, um Molekül- und Proteininformation auch während eines zweigeteilten, indexbasierten VS-Prozesses konsistent handhaben zu können und Störungen der Berechnungen durch einen widersprüchlichen Aufbau interner Protein- und Molekülrepräsentationen zu vermeiden. Inkonsistente Berechnungen können aber auch aufgrund unterschiedlicher Präparierungsprotokolle herrühren, die Protonierungszustände und Tautomere von Protein und Molekülen generieren. Besonders im deskriptorbasierten Docking und bei der Auswertung von Pharmakophorhypothesen sind die Auswirkungen der Abhängigkeit vom Eingabezustand zu spüren, da sich derartige Ansätze auf eine Bewertung entscheidender bindungsvermittelnder Merkmale stützen. Diese Arbeit setzt sich speziell mit dieser Problematik auseinander. Mit dem Ziel konsistente Bindungsmodusvorhersagen für jeglichen gegebenen Eingabezustand von Protein und Ligand zu gewährleisten, sind deshalb Protomerfreiheitsgrade in cRAISE integriert. Der Ansatz soll in erster Linie die Unabhängigkeit vom gegebenen Eingabezustand von Protein- und Ligand gewährleisten. Dies könnte prinzipiell auch durch eine Normalisierung der Eingaben durch die Annahme eines sehr wahrscheinlichen Grundzustands erreicht werden. Dies ist aber nicht ausreichend, da im VS sowohl Protein als auch Ligand in Abhängigkeit von ihrem komplexierten Gegenstück unterschiedliche Zustände annehmen können. Deshalb soll der Ansatz mögliche Protein- und Ligandzustände evaluieren und einen passenden Zustand im generierten Protein-Ligand-Komplex wählen. Das Resultat soll vergleichbar sein mit dem Ansatz einer präparierenden Zustandsenumeration, gefolgt von einer Bewertung jeder möglichen Protein-Ligand-Zustandskombination. Im Gegensatz zu diesem naiven Ensembleansatz, der oftmals zu aufwendig ist und deshalb nicht praktiziert wird, soll der neuartige Ansatz so effizient sein, dass er sich auch zum Screening umfangreicher Molekülbibliotheken eignet.

1.3 Struktur der Dissertation

Die vorliegende Dissertation beschreibt die theoretische Grundlage von cRAISE. Das folgende *Kapitel 2* widmet sich den grundlegenden Prinzipien des rationalen Wirkstoffentwurfs. Sowohl die Modelle des computergestützten Wirkstoffentwurfs im Allgemeinen, als auch die Modelle von cRAISE stützen sich fundamental auf diese Grundsätze. Das Kapitel konzentriert sich auf Aspekte, die für diese Arbeit als besonders relevant erachtet wurden. *Kapitel 3* ordnet den Gegenstand dieser Arbeit im Bereich des computergestützten Wirkstoffentwurfs ein. Es erklärt die verschiedenen VS-Strategien und

gibt einen Überblick über existierende Ansätze, die sich mit den zentralen oder ähnlich gearteten Fragestellungen befassen. cRAISE vereint eine Vielzahl anderer am Zentrum für Bioinformatik der Universität Hamburg entwickelter Methoden die in *Kapitel 4* vorgestellt sind. Sie übernehmen Randaufgaben, die zur Umsetzung eines SBVS notwendig sind. An vielen Stellen greift cRAISE auf deren zugrundeliegende Modelle zurück und erweitert sie, um seine Ziele zu verfolgen. Zudem werden das Konzept der Flex*-basierten TrixX-Versionen und ein am Berkeley Lab entwickeltes Indexierungssystem erläutert. Ersteres greift cRAISE erneut auf. Letzteres kommt in cRAISE, wie bereits in TrixX-BMI, zur Verwaltung von Molekülinformation zum Einsatz. Zusammen tragen die soweit beschriebenen Grundlagen unmittelbar zum Verständnis der in *Kapitel 5* skizzierten Strategie bei, die cRAISE zur Realisierung seiner Ziele verfolgt. Es bietet einen Überblick über den Ablauf eines gewöhnlichen SBVS, eines geleiteten Screenings bei gegebenen Randbedingungen und bei Berücksichtigung molekularer und makromolekularer Zustände. *Kapitel 6* beschreibt detailliert die Realisierung und erläutert die entwickelten Modelle. Neben der deskriptorbasierten Docking-Methode und ihrer Erweiterung zu einem SBVS-Ansatz, wird erklärt wie cRAISE Zusatzinformation für ein extern geleitetes Screening verarbeitet und Protonierungszustände und Tautomere von Protein und Molekülen integriert. *Kapitel 7* beschreibt die zur Validierung angewandten Bewertungsstrategien, die genutzten Maße, Daten und Experimente. Anhand dieser wird in *Kapitel 8* überprüft, ob die initial aufgestellten Ziele erfolgreich umgesetzt werden konnten. Es diskutiert Qualitäten und Fehlschläge der Vorhersagen und die dafür aufzuwendenden Ressourcen. *Kapitel 9* fasst die Erkenntnisse dieser Arbeit zusammen, beschreibt Limitierungen und schlägt Wege zur zukünftigen Bewältigung vor. Zudem werden Methoden vorgestellt, die mit der RAISE-Technologie bereits erfolgreich umgesetzt wurden. Die Arbeit endet mit einer Perspektive für weitere interessante Neuentwicklungen, die auf Grundlage der entstandenen Modelle und Methoden möglich sind.

2 Prinzipien des rationalen Wirkstoffentwurfs

Mit der Idee, dass die chemische Konstitution von Arzneimitteln in direktem Zusammenhang zu ihrem Wirkmechanismus stehen könnte, verfolgte Paul Ehrlich zu Beginn des 19. Jahrhunderts das Ziel, chemische Substanzen zu finden, die spezifische Affinitäten für „Rezeptoren“ zeigen.[4]

> „Wir müssen von einer Substanz ausgehend, Homologe und Derivate der verschiedensten Art darstellen und jede auf ihren Wirkungswert ausprobieren. Wir müssen also zielen lernen, und zielen lernen durch chemische Variation.“[5]

Nachdem der Erreger der Syphilis entdeckt wurde, beschloss Ehrlich ein Arzneimittel zu finden, das effektiv gegen diesen wirkt. Dafür modifizierte und testete er systematisch Hunderte von chemischen Substanzen und fand schließlich mit dem Präparat 606 eine wirksame Verbindung, die unter dem Namen Salvarsan als erstes Chemotherapeutikum in die Geschichte einging.[6] Im Einsatz zeigte die arsenhaltige Verbindung jedoch unerwünschte Nebeneffekte und erfüllte nicht die Erwartungen an eine „magische Kugel“, die ausschließlich die Krankheitsursache bekämpfte. Dennoch demonstrierten seine zahlreichen Experimente, dass es möglich war, systematisch Wirkstoffe zu entwickeln, die gezielt wirken. Die Prinzipien des von Ehrlich etablierten Prozesses, d. h. die Optimierung von Strukturen durch die Synthese und das *Screening* von Derivaten, haben bis heute Bestand und bilden die Basis für den modernen *rationalen Wirkstoffentwurf*. Dieses Kapitel widmet sich den Prinzipien, auf denen dieser Vorgang ruht.

2.1 Phasen des rationalen Wirstoffentwurfs

Generell kann die Entwicklung eines Wirkstoffes von der initialen Idee bis zum marktreifen Produkt in zwei mehrstufige Prozesse eingeteilt werden:

- In den ersten Stufen oder *frühen Phasen* sollen Wirkstoffkandidaten „gefunden“, „gesucht“, „entdeckt“ oder „entworfen“ werden (drug discovery).
- In den *späten Phasen* soll der Kandidat tatsächlich „entwickelt“ oder „erschlossen“ und zur Marktreife gebracht werden (drug development).

2.1.1 Frühe Phasen des rationalen Wirkstoffentwurfs

Die frühen Phasen realisieren die Zielstruktursuche, die Treffersuche, die Leitstruktursuche und die Leitstrukturoptimierung.[7, 8, 9]

Zielstruktursuche: Der Ausgangspunkt bei der Entwicklung eines Arzneimittels stellt eine Krankheit dar, zu deren Bekämpfung bislang keine oder nur unzureichende Mittel zur Verfügung stehen. Für sie muss innerhalb eines biochemischen Prozesses eine Zielstruktur gefunden werden, bei der durch Modulation der Aktivität (Blockierung oder Aktivierung) ein therapeutischer Effekt erzielt werden kann (target identification). Die Zielstruktur muss für einen mutmaßlichen Wirkstoff zugänglich sein und bei Modulation eine biologische Antwort auslösen, die *in vitro* und *in vivo* messbar ist und keine Seiteneffekte auslöst (target validation). In dieser Phase wird zumeist das assoziierte Gen des Ziels (i. d. R. ein Protein) und dessen dreidimensionale (3D) Struktur aufgeklärt.

Treffersuche: Die Intention der darauffolgenden Phase ist es, eine niedermolekulare Substanz zu finden, die zumindest im hohen nano- bis niedermikromolaren Bereich Aktivität gegen die Zielstruktur zeigt (hit identification). Dafür werden große Substanzbibliotheken experimentell *in vitro* getestet oder „gescreent“. Voraussetzung ist die Etablierung eines biochemischen Testsystems (Assay), mit dem die Affinität einer Substanz zur Zielstruktur gemessen werden kann. Die Testung erfolgt vollautomatisiert im Hochdurchsatzverfahren (High throughput screening, HTS). Substanzen, die eine geforderte Affinität aufweisen werden als *Treffer* oder *Hits* bezeichnet. Aus einem HTS ergeben sich keine oder viele Hits, die bestätigt, reduziert, klassifiziert und zu Hit-Serien zusammengefasst werden. Die Hits werden erneut mit dem ursprünglichen Assay und orthogonal mit Assays anderer Art getestet, um die initialen Ergebnisse zu bestätigen (hit validation). Zudem werden Dosis-Wirkungs-Kurven und Struktur-Aktivitäts-Beziehungen erstellt, um die Potenz der Hits zu vergleichen und eine Klassifizierung vorzunehmen.

Leitstruktursuche: Die Leitstruktursuche beginnt mit einer Reduzierung der molekularen Komplexität. Dabei sollen funktionelle Gruppen oder Bestandteile der Moleküle identifiziert werden, die nicht zur Affinität beitragen. Indirekt werden dadurch ein oder mehrere kleinste, aktive Pharmakophore (vgl. Abschnitt 2.6) und Leitstrukturen mit nachgewiesener Aktivität definiert. Zudem werden Leitstrukturen durch Variation

zu Leitstrukturserien erweitert. In einer Kaskade weiterer *in vitro* und *ex vivo* Tests werden diese Strukturen bezüglich einer Vielzahl von Aspekten bewertet. So werden u. a. Absorptions-, Verteilungs-, Metabolismus- und Ausscheidungsmerkmale (ADME), Toxizität und Synthetisierbarkeit beurteilt sowie Selektivitätsstudien durchgeführt. Anhand dieser Information werden Strukturen, die strenge Kriterien erfüllen und ausgearbeiteten Eigenschaftsprofile besitzen für die weiteren Phasen ausgewählt.

Leitstrukturoptimierung: Am Ende der Leitstruktursuche steht eine Auswahl potentieller Moleküle, für die eine Vielzahl guter Eigenschaften und Mängel dokumentiert sind. In der Phase der Leitstrukturoptimierung sollen die Mängel behoben und zugleich die guten Eigenschaften erhalten oder verbessert werden. Vor allem werden der Wirkstoffmetabolismus, pharmakokinetische Eigenschaften, Wirkung und Sicherheitsaspekte der Moleküle in komplexen Assays untersucht und optimiert. Dadurch soll der Einsatz *in vivo* in den klinischen Phasen vorbereitet werden. Am Ende dieser Phase stehen ein oder wenige *Kandidaten* mit gut dokumentierten Wirkstoffprofilen, die an die späten Phasen der Wirkstoffentwicklung übergeben werden.

2.1.2 Späte Phasen des rationalen Wirkstoffentwurfs

Die späten Phasen werden durch die präklinische Phase initiiert. Sie evaluiert die Wirkstoffkandidaten auf Zellebene und im Tiermodell und soll neben Sicherheitsaspekten, primär die Frage der Dosisfindung und Darreichungsform abklären. Zudem müssen Möglichkeiten zur groß angelegten Synthetisierung und Produktion bedacht werden. Es folgen mehrstufige klinische Studien am Menschen, um die Voraussetzung für die Zulassung des Kandidaten zu erbringen. Seine Wirkung muss bestätigt und Seiteneffekte minimiert werden. Die klinische Phase I wird an gesunden Probanden zur Bewertung der Pharmakokinetik, Sicherheit und Dosis durchgeführt. Phase II soll Hinweise zur Wirksamkeit erbringen und weitere Sicherheitsaspekte und die Tolerierbarkeit abklären. Sie wird an einer kleinen Gruppe kranker Probanden durchgeführt. Phase III, die an einer großen Gruppe von Probanden durchgeführt wird, muss die Wirksamkeit und die Tolerierbarkeit im Vergleich zum Placebo bestätigen. Es folgen die Zulassung, Produktion und Vertrieb. Bei Markteinführung beginnt die klinische Phase IV, die in einer Langzeitstudie Neben- und Wechselwirkungen aufklärt.[10]

2.2 Zielstrukturen

Neben der DNA und der RNA sind Proteine die häufigsten Zielstrukturen. Sie stehen im Fokus dieser Arbeit. Proteine sind Makromoleküle, die aus einer oder aus mehre-

ren Aminosäureketten bestehen. In Eukaryoten sind 20 unterschiedliche α-Aminosäuren (AS) über den genetischen Code durch 61 Codons auf direktem Wege kodiert[1] und proteinogen, d. h. sie bilden Bestandteile von Proteinen. Jede AS setzt sich aus einer Carboxy-Gruppe, einer Amino-Gruppe, einem Wasserstoffatom und einem AS-Rest oder *Seitenkette* zusammen. Während der Proteinbiosynthese bilden Carboxy- und Amino-Gruppe zweier aufeinanderfolgender AS eine Peptidbindung und kondensieren so zu Polymeren.[11] Die Folge der aneinandergereihten AS legt die *Primärstruktur* eines Proteins fest. Keto- und Amino-Gruppen des kondensierten Proteinrückgrats bilden intramolekulare Wasserstoffbrücken und formen dadurch lokale, dreidimensionale Strukturelemente wie Schleifen, Faltblätter oder Helices (*Sekundärstruktuen*). Durch Interaktionen zwischen Seitenketten verwinden sich die Sekundärstrukturelemente. Das Protein bildet so eine komplexere *Tertiärstruktur*. Mehrere Proteinuntereinheiten können sich zusammenlagern und Oligomere bilden (*Quartärstruktur*).[12] Die zuletzt geformte 3D-Struktur oder *Konformation* eines Proteins lässt sich experimentell durch Röntgenkristallographie oder NMR-Spektroskopie aufklären. In der Protein Data Bank (PDB) sind viele 3D-Strukturen der Öffentlichkeit zur Verfügung gestellt.[13]

Proteine übernehmen eine Vielzahl von Funktionen. Sie übermitteln inter- und intrazelluläre Signale, können als Transporter oder selbst als Botenstoffe fungieren, sie bilden weiterhin wichtige Bestandteile des Immunsystems und strukturgebende Baustoffe oder katalysieren als Enzyme wichtige Stoffwechselreaktionen. Das Hauptmerkmal vieler Proteine ist, dass sie *Liganden* binden. Liganden können entweder andere Proteine oder niedermolekulare Moleküle sein. Der Ort der Ligandbindung wird als *Bindestelle* oder, im Fall von Enzymen, als *aktives Zentrum* bezeichnet. Da sie oft nicht völlig zugänglich, sondern unter der Oberfläche vergraben liegen, spricht man auch von *Bindetaschen*. Sie entstehen bei der Faltung eines Proteins.[14] Ob ein Molekül bindet oder nicht hängt u. a. von der Konstitution der Bindetasche und den dort vorgefundenen AS ab, deren Seitenketten unterschiedliche physikochemische Eigenschaften aufweisen. AS lassen sich generell als unpolar, polar oder ionisierbar klassifizieren (vgl. Tabelle 2.1). Prinzipiell tendieren unpolare AS dazu, Kontakte untereinander zu bilden und ins Innere der Proteine zu ragen, um einer ungünstigen Konfrontation mit Wasser zu entgehen. Polare AS treten dagegen häufig an der Oberfläche eines Proteins auf und sind in der Lage mit Wasser zu interagieren.[15] Einige AS-Reste sind von dualer Natur (unpolares und polares Merkmal) oder können in Abhängigkeit des pH-Werts ihrer Umgebung geladen vorkommen.

[1] Unter bestimmten Bedingungen kann eine 21. AS, das Selenocystein, unter Einbeziehung des Stop-Codons (UGA) indirekt kodiert und über Umwege in Proteine eingebaut sein. Eine 22. AS, das Pyrrolysin, kann über das Stop-Codon UAG in Archaen und Bakterien kodiert sein.

Tabelle 2.1: Charakteristika der 20 proteinogenen Aminosäuren.

Aminosäure	Abkürzung	Kategorie
Isoleucin	Ile	unpolar
Valin	Val	unpolar
Leucin	Leu	unpolar
Phenylalanin	Phe	unpolar
Cystein	Cys	polar, nicht ionisierbar
Methionin	Met	polar, nicht ionisierbar
Alanin	Ala	unpolar
Glycin	Gly	unpolar
Threonin	Thr	polar, nicht ionisierbar
Serin	Ser	polar, nicht ionisierbar
Tryptophan	Trp	polar, nicht ionisierbar
Tyrosin	Tyr	polar, nicht ionisierbar
Prolin	Pro	unpolar
Histidin	His	polar, basisch ionisierbar
Glutamin	Gln	polar, nicht ionisierbar
Asparagin	Asn	polar, nicht ionisierbar
Glutamat	Glu	polar, sauer ionisierbar
Aspartat	Asp	polar, sauer ionisierbar
Lysin	Lys	polar, basisch ionisierbar
Arginin	Arg	polar, basisch ionisierbar

2.3 Bindungsaffinität

Beim Entwurf von Wirkstoffen hat man das Interesse, Moleküle zu identifizieren, die in die Proteinbindetasche binden und somit Affinität aufweisen. Manche Liganden bilden kovalente Bindungen zum Protein, sodass ein irreversibler Protein-Ligand-Komplex entsteht. Obwohl dies durchaus kontrovers diskutiert wird[16], stehen aufgrund von Sicherheitsaspekten primär nicht-kovalente, *reversible Bindungen* im Interesse der Medizinalchemie. Dabei findet in Lösung die Assoziation und Dissoziation von Protein und Ligand in stetigem Wechsel statt, sodass eine gewisse Konzentration gelöster Proteine $[P]$, gelöster Liganden $[L]$ und assoziierter Protein-Ligand-Komplexe $[PL]$ vorzufinden ist. Dieser Prozess wird durch folgende Reaktionsgleichung beschrieben:

$$P + L \underset{k_{\text{diss}}}{\overset{k_{\text{bind}}}{\rightleftharpoons}} PL \tag{2.1}$$

Nach einer gewissen Zeit stellt sich bei der Reaktion ein Gleichgewicht ein, sodass die Raten für die Bindung (k_{bind}) und die Dissoziation (k_{diss}) konstant sind und sich das Verhältnis der Konzentrationen von Protein mit gebundenem Ligand $[PL]_{\text{eq}}$ zu gelöstem

Protein $[P]_{\text{eq}}$ und Ligand $[L]_{\text{eq}}$ kaum ändert. Das Verhältnis der Raten spiegelt dann wider, ob das Gleichgewicht auf Seite der gelösten oder auf Seite der gebundenen Liganden zu finden ist. Die sogenannte *Bindungskonstante* K_B beschreibt das Gleichgewicht mit Fokus auf die Vorwärtsreaktion:

$$K_B = \frac{k_{\text{bind}}}{k_{\text{diss}}} = \frac{[PL]_{\text{eq}}}{[P]_{\text{eq}}\,[L]_{\text{eq}}} = \frac{1}{K_D} \tag{2.2}$$

Die Bindungskonstante ist ein Maß für die Affinität. Ist sie hoch, genügt bereits eine geringe Konzentration an Liganden, um im Gleichgewicht der Reaktion Protein-Ligand-Komplexe mit hoher Wahrscheinlichkeit vorzufinden. Mit der *Dissoziationskonstante* K_D betrachtet man das Gleichgewicht mit Fokus auf die Rückreaktion. Sie entspricht der Reziproken der Bindungskonstante. Im Falle enzymatischer Reaktionen wird K_D typischerweise als Inhibierungskonstante K_I bezeichnet. In beiden Fällen sind ihre Werte für hoch-affine Liganden besonders niedrig, sodass die Wahrscheinlichkeit ungebundene Liganden vorzufinden gering ist. Hohe K_B-Werte bzw. niedrige K_D/K_I-Werte sind somit wünschenswerte Eigenschaften für Wirkstoffkandidaten.

Betrachtet man die Bindungsreaktion aus thermodynamischer Sicht, so möchte man dass die Bindung von Protein und Ligand spontan verläuft. Dies ist der Fall wenn die Änderung der Gibbs-Energie ΔG negative Werte annimmt. Unter Standardbedingungen, bei konstantem Druck besteht zwischen Bindungs- bzw. Dissoziationskonstante und der Änderung der Gibbs-Energie folgender Zusammenhang:

$$\Delta G = -RT \ln K_B = RT \ln K_D \tag{2.3}$$

R entspricht der Gaskonstanten ($8{,}3144621\,\text{J}\,\text{mol}^{-1}\,\text{K}^{-1}$) und T der absoluten Temperatur ($298{,}15\,\text{K}$). Mit experimentell bestimmten Dissoziationskonstanten lassen sich mittels Gleichung 2.3 ΔG-Werte berechnen. Die Änderung der Gibbs-Energie lässt sich weiter in enthalpische bzw. energetische und entropische Beiträge separieren:

$$\Delta G = \Delta H - T\Delta S \tag{2.4}$$

ΔH bezeichnet die Enthalpieänderung und ΔS die Entropieänderung.

Enthalpische Beiträge: Bei der Protein-Ligand-Bindung liefern elektrostatische Wechselwirkungen Beiträge zur Enthalpieänderung. Dabei interagieren Ione, permanente und temporäre Dipole miteinander. Prinzipiell lassen sich diese *energetischen Wechselwirkungen* anhand der Art der beteiligten Interaktionspartner klassifizieren. Bei *ionischen Wechselwirkungen* interagieren zwei ionisierte Gruppen miteinander. *Ion-Dipol-Wechselwirkungen* werden dadurch etabliert, dass ein Ion mit einer polaren Gruppe

(permanenter Dipol) interagiert. *Dipol-Dipol-Wechselwirkungen* sind durch die Interaktion zweier permanenter Dipole charakterisiert. Darüber hinaus kann ein Ion oder Dipol in einer eigentlich unpolaren Gruppe einen temporären Dipol induzieren, indem es dessen Elektronenwolke stört d. h. diese verdrängt bzw. anzieht. Zuletzt können über kurze Distanzen fluktuierende Elektronen unpolarer Gruppen gegenseitig temporäre Dipole induzieren, sodass *Dispersionskräfte* wirken. *Wasserstoffbrücken* sind Dipol-Dipol- oder Ion-Dipol-Wechselwirkungen bei denen ein Wasserstoffatom beteiligt ist, das an ein elektronegativeres Atom gebunden ist. Unter dem Begriff *van-der-Waals-Wechselwirkung* werden die Wechselwirkungen zusammengefasst, die nicht die Beteiligung von Ionen fordern und keine Wasserstoffbrücken sind.

Entropische Beiträge: Entropische Beiträge resultieren aus der Immobilisierung bzw. aus der Mobilisierung der Reaktionspartner. Prinzipiell führt eine Protein-Ligand-Bindung immer zu einem Entropieverlust, da die Bewegungsfreiheit der Bindungspartner eingeschränkt wird. Entropische Beiträge können aber auch begünstigend sein. Ein Beispiel ist der *hydrophobe Effekt*: In Lösung etablieren Wassermoleküle ein dynamisches, jedoch regelmäßiges Wasserstoffbrückennetzwerk. Liegen unpolare Gruppen ungebunden in Lösung vor, so stören sie dieses Netzwerk. Da sie nicht direkt mit den Wassermolekülen interagieren können, müssen sich die Wassermoleküle umorientieren, sodass Brücken tangential zu den unpolaren Gruppen gebildet werden können. Dadurch entsteht ein Wasserkäfig, bei dem die Bewegungsfreiheit der Wassermoleküle eingeschränkt ist. Dies führt zu einem Entropieverlust. Um diesen Verlust möglichst gering zu halten und möglichst viele Wasserstoffbrücken zwischen Wassermolekülen zu etablieren, muss die Oberfläche störender Gruppen möglichst gering gehalten werden. Dies wird erreicht indem unpolare Gruppen aggregieren. Der hydrophobe Effekt ist die Triebkraft bei der Faltung von Proteinen[15], trägt jedoch auch oft entscheidend zur Ligandbindung bei.

Eine Bilanz enthalpischer und entropischer Beiträge kann Erkenntnisse darüber liefern, ob die Protein-Ligand-Bindung spontan verläuft und der Ligand affin ist.[17] Dafür muss die Änderung der Gibbs-Energie im System der Reaktion betrachtet werden. In diesem werden Interaktionen zwischen Wasser und Protein bzw. Ligand gebrochen (*Desolvatisierung*), Interaktionen zwischen Protein und Ligand gebildet und Protein, Molekül und Wasser mobilisiert bzw. immobilisiert. Können dabei ungünstige Beiträge kompensiert werden, sodass die günstigen Beiträge überwiegen, kann auf einen affinen Liganden geschlossen werden. Eine Bilanzierung durch die Summation einzelner Beiträge wird jedoch durch die Tatsache erschwert, dass Einzelbeiträge nicht notwendigerweise konstant sind, sondern durch kooperative Effekte innerhalb des Systems verstärkt oder gemindert werden können.[18]

2.4 Protein-Ligand-Interaktionen

Emil Fischer postulierte 1894 das Schlüssel-Schloss-Prinzip, welches bereits Ehrlich in seinen zahlreichen Vorträgen als Grund für den Wirkungseffekt nannte:

> „Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glucosid wie Schloss und Schlüssel zu einander passen müssen, um eine chemische Wirkung auf einander ausüben zu können."[19]

Damit beschrieb er die spezifische Komplementarität von Protein und Ligand und bezog sich auf deren geometrische Form, die für die Passung notwendig ist. Dieses Bild kann auf die räumliche Anordnung der physikochemischen Eigenschaften zwischen Bindetasche und Ligand übertragen werden. Ist deren Anordnung komplementär, so bilden sich intermolekulare Interaktionen:

Wasserstoffbrücken: 2011 definierte die International Union of Pure and Applied Chemistry (IUPAC) den Begriff der Wasserstoffbrücke nach 1997[20] erneut: Demnach[1] ist die Wasserstoffbrücke eine attraktive Interaktion zwischen einem Wasserstoffatom eines Moleküls oder eines molekularen Fragments X–H (wobei X elektronegativer als H ist) und einem Atom oder einer Gruppe von Atomen, desselben oder eines anderen Moleküls, für die es Belege für die Brückenbildung gibt. Eine typische Wasserstoffbrücke kann als X–H$\cdots$Y–Z dargestellt werden. X–H bezeichnet den Wasserstoffbrückendonor. Der Akzeptor Y kann ein Atom oder ein Anion sein, oder ein Fragment oder Molekül Y–Z, in welchem Y an Z gebunden ist.[21] In Protein-Ligand-Komplexen findet man zum einen die *klassischen Wasserstoffbrücken*. Sie bilden sich zwischen Stickstoff- bzw. Sauerstoffdonoren und Stickstoff- bzw. Sauerstoffakzeptoren und folgen im Allgemeinen sehr strikten Geometrien:[18]

- Die Anordnung von Donorschweratom, Wasserstoff und Akzeptorschweratom ist nahezu linear.
- Die Schweratomdistanz bewegt sich im Bereich von 2,6 und 3,2 Å.
- Die Brücke wird zum freien Elektronenpaar etabliert.
- Interaktionsatome aus π-Systemen etablieren bevorzugt eine Brücke in der dadurch aufgespannten Ebene.

Dagegen sind die sogenannten *schwachen Wasserstoffbrücken* sehr vielfältig. Hierbei können z. B. das π-System aromatischer Ringe oder Fluoratome als Akzeptoren und

[1] The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation.

C–H-Gruppen als Donoren fungieren.[22] Generell kann man Wasserstoffbrücken als eine spezielle Art von Dipol-Dipol- oder Ion-Dipol-Wechselwirkung betrachten, die die Beteiligung eines Wasserstoffatoms fordert.

Salzbrücken: *Salzbrücken* sind rein ionische Wechselwirkungen. Sie lassen sich prinzipiell auch als eine Art Wasserstoffbrücke betrachten bei denen Kationen als Donoren und Anionen als Akzeptoren fungieren. In Protein-Liganden-Komplexen setzt dies die Beteiligung einer ionisierten Seitenkette voraus. Somit können (abhängig vom umgebenen pH-Wert) ausschließlich Asp, Glu, Lys, Arg und His an Salzbrücken beteiligt sein. Die Schweratomdistanz zwischen den Interaktionspartnern ist allerdings im Vergleich zu klassischen Wasserstoffbrücken verkürzt.[23]

Metallinteraktionen: In Metalloenzymen koordinieren umgebene Seitenketten wie His, Asp, Glu und Cys ein Metallkation oftmals so, dass das Kation als Donor dienen kann. Zwar trägt dies kein Wasserstoff zur Bindung bei, jedoch interagiert es über eine freie Koordinierungsstelle mit einem klassischen Wasserstoffbrückenakzeptor oder Anion. In welche Richtung die Interaktion etabliert wird, hängt dabei sowohl vom Metalltyp, dessen möglichen Koordinationsgeometrien, als auch von der Anordnung und Zahl der koordinierenden Seitenketten ab.[24] Die Bindungen, die das Metall etabliert, sind dabei sehr kurz und liegen deutlich unter der Summe der Van-der-Waals-Radien der beteiligten Interaktionspartner (2.0–2.5 Å).[25]

Hydrophobe Interaktionen: Hydrophobe Interaktionen entstehen durch die Anlagerung von aliphatischen oder aromatischen Seitenketten an Alkyl- oder Aryl-Gruppen des Liganden. Ursache hierfür ist neben den schwach wirkenden Dispersionskräften hauptsächlich der hydrophobe Effekt (vgl. Abschnitt 2.3). Im Allgemeinen sind hydrophobe Interaktionen, im Gegensatz zu Wasserstoff-, Salzbrücken und Metallinteraktionen, ungerichtet und weitreichender. Die Schweratomdistanzen der beteiligten Atome bewegen sich im Bereich von 3,4 bis 4,4 Å.[18] Aromatische Gruppen bilden untereinander eine besondere Art hydrophober Interaktionen. Die π-Systeme aromatischer Ringsysteme können sogenannte *π-π-Interaktionen* bilden. Dabei ordnen sich die Ringe bevorzugt auf zwei Weisen an. Sie können leicht verschoben, parallel zueinander oder T-förmig angeordnet sein. Durch diese Anordnungen kann ein günstiger Quadrupol etabliert werden.[26, 27] Die aromatischen Aminosäuren Phe, Tyr, Trp und His können an π-π-Interaktionen beteiligt sein.

Weitere Interaktionen: Aromatische Ringe können sehr vielseitig interagieren.[27] Bei *Kation-π-Interaktionen* interagiert eine positiv ionisierte Gruppe (meist Lys, seltener Arg) mit dem π-System eines Aromaten so, dass die Elektronen die positive Ladung

ummanteln.[28] Wenn schwere Halogene X (X = Cl, Br, I) an elektronenziehende Gruppen gebunden sind, besitzen sie ein positives *σ-Loch* in Verlängerung zur C–X-Achse. Die drei freien Elektronenpaare bilden um die Achse einen negativen Gürtel. Dadurch können sie mit Elektrophilen und Nukleophilen interagieren. Interagieren sie über das σ-Loch so spricht man von *Halogenbindung*. Diese sind dadurch charakterisiert, dass sie spezifische, lineare Interaktionen mit klassischen H-Brücken Akzeptoren (z.B. C–Cl $\cdots$ O) etablieren.[29] Dabei steigt die Stärke der Bindung mit Größe des Halogens.[30] Fluor ist so klein, dass es kein σ-Loch besitzt. Seine freien Elektronen sind gleichmäßig verteilt. Dennoch interagiert es manchmal, wie andere Halogene auch, sehr schwach mit einem Carbonyl-Kohlenstoff oder einem Amid und bildet sogenannte *orthogonale multipolare Interaktionen*.[18] Divalenter Schwefel kann sowohl mit elektronenreichen Gruppen entlang der σ*-Richtung (einer C–S-Achse) als auch mit elektronenarmen Gruppen entlang seiner freien Elektronenpaare interagieren.[31] Dadurch kann divalenter Schwefel schwache Wasserstoffbrücken bilden.[32] Sowohl Methionin als auch Cystein treten aber auch mit Aryl-Gruppen in Kontakt.[27]

2.5 Charakteristika von Wirkstoffkandidaten

Die physikochemischen Eigenschaften niedermolekularer Verbindungen sind ausschlaggebend dafür, ob ein Wirkstoffkandidat letztendlich in Form eines Medikaments etabliert werden kann. Bei der Entwicklung gilt es verschiedene Faktoren zu berücksichtigen, die unter anderem über die Bioverfügbarkeit eines Wirkstoffs entscheiden, also ob dieser absorbiert und am Zielort freigesetzt werden kann. Konzepte wie die *Wirkstoffähnlichkeit* (Druglikeness) oder *Leitstrukturähnlichkeit* (Leadlikeness) versuchen auf Basis von beobachteten Eigenschaften bereits bekannter Wirkstoffe Kriterien zu definieren, um Kandidaten bezüglich dieser wichtigen Anforderungen einzuschätzen.

2.5.1 Wirkstoffähnlichkeit

Die Fähigkeit zur Absorption und Freisetzung eines Wirkstoffs hängt von dessen Membrangängigkeit und Löslichkeit ab. Beide Faktoren stellen gegensätzliche Forderungen an die physikochemischen Eigenschaften eines Moleküls. Die Permeabilität fordert einen kleinen, lipophilen Wirkstoff, sodass die Membran durchdrungen werden kann. Zur Freisetzung und Verteilung des Wirkstoffs am Zielort muss dieser im Serum löslich und somit hydrophil sein. Wirkstoffe werden beiden Anforderungen gerecht und weisen ein ausbalanciertes Eigenschaftsprofil auf. Lipinski formulierte eine Faustregel, die populäre *Rule of Five*, um Verbindungen entsprechend dieser Eigenschaften zu charakterisieren.[33] Die

Regel legt Grenzwerte für die Anzahl an Wasserstoffbrückendonoren und -akzeptoren, das Molekulargewicht und den Octanol/Wasser-Verteilungskoeffizienten[1] fest. Verbindungen sind *wirkstoffähnlich*, wenn sie diese Werte einhalten. Halten sie die Werte nicht ein, werden sie wahrscheinlich schlechter absorbiert und sind membranundurchlässiger. Ghose *et al.* verfeinerten die Regel und schränkten das Molekulargewicht und den Octanol/Wasser-Verteilungskoeffizienten weiter ein.[35] Beide *Druglike*-Kriterien sind in Tabelle 2.2 gegenübergestellt. Um eine orale Bioverfügbarkeit zu gewährleisten, erklärten Veber *et al.*, dass zudem die Anzahl rotierbarer Bindungen und der polare Oberflächenbereich (PSA) beschränkt sein sollten.[36]

Tabelle 2.2: Wirkstoff- und Leitstrukturkriterien.

Parameter	Lipinski-Druglike[33]	Ghose-Druglike[35]	Oprea-Leadlike[37]
MW[a]	≤ 500	$160 \leq 480$	≤ 450
cLogP[b]	≤ 5	$-0,4 \leq \text{aLogP} \leq 5,6$	$-3,5 \leq \text{cLogP} \leq 4,5$
N_{HBD}[c]	≤ 5		≤ 5
N_{HBA}[d]	≤ 10		≤ 8
N_{A}[e]		$20 \leq N_{\text{A}} \leq 70$	
MR[f]		$40 \leq \text{MR} \leq 130$	
$\text{LogD}_{7,4}$[g]			$-4 \leq \text{LogD}_{7,4} \leq 4$
N_{RNG}[h]			≤ 4
N_{RTB}[i]			≤ 10

[a] Molekulargewicht
[b] Berechneter Logarithmus des Octanol/Wasser-Verteilungskoeffizienten
[c] Anzahl der Wasserstoffbrückendonoren
[d] Anzahl der Wasserstoffbrückenakzeptoren
[e] Gesamtzahl der Atome
[f] Molrefraktion
[g] Berechneter Logarithmus des Verteilungskoeffizienten bei pH 7,4
[h] Anzahl der Ringe
[i] Anzahl nicht-terminaler rotierbarer Bindungen

2.5.2 Leitstrukturähnlichkeit

Während der Leitstrukturoptimierung tendiert ein Molekül dazu, größer zu werden. Eine Verbindung wird normalerweise mit zusätzlichen Gruppen ausgestattet, um phar-

[1] Der P-Wert bezeichnet den dimensionslosen Octanol/Wasser-Verteilungskoeffizienten. Er gibt das Verhältnis der Konzentrationen einer Substanz in einem Zweiphasensystem an, welches aus Oktanol und Wasser besteht. Das logarithmische Verhältnis skaliert den Wert, sodass positive Werte lipophile, negative Werte hydrophile Substanzen kennzeichnen. Der LogP wird experimentell bestimmt. Der cLogP bezeichnet den berechneten LogP. Andere LogP-Berechnungen sind der aLogP oder mLogP[33, 34].

makokinetische Eigenschaften zu verbessern. Auch nach der Optimierung sollte ein Kandidat noch immer wirkstoffähnlich sein. Um der Optimierung Raum zur Variation der Molekülstruktur zu bieten, sollten deshalb nur Verbindungen in diese Phase eingehen, bei denen ein Dekorieren keine Verletzung der Lipinski-Regeln mit sich bringt. Deshalb schlug Teague Eigenschaften vor, die Kandidaten vor der Optimierung erfüllen sollten.[38] Diese *Leadlike*-Kriterien wurden später, nach einer ausgedehnteren Analyse bekannter Leitstrukturen, weiter verfeinert.[37] Die Oprea-Leadlike-Kriterien sind in Tabelle 2.2 zusammengefasst. Im Vergleich zur Wirkstoffähnlichkeit betonen sie vor allem die geringere Komplexität von Leitstrukturen. Zudem existieren Regeln, die Fragmentähnlichkeit definieren und die Komplexität der Moleküle weiter beschränken.[39, 40]

2.5.3 Zielstrukturspezifische Eigenschaften

Die bislang vorgestellten Kriterien sind nicht immer angebracht.[41] Verbindungen, die im zentralen Nervensystems (ZNS) oder antibakteriell wirken, widersprechen den herkömmlichen Druglike- und Leadlike-Kriterien.[42] Um ins zentrale Nervensystem zu gelangen, müssen Wirkstoffe die selektive Blut-Hirn-Schranke überwinden. Sie besitzt passive und aktive Transportmechanismen, zu deren Durchquerung andere physikochemische Eigenschaften notwendig sind. Prinzipiell müssen Verbindungen kleiner und hydrophober sein. Aber auch die Flexibilität und die Ladung beeinflussen den Transport. Es existieren verschiedene Charakterisierungen von ZNS-aktiven Verbindungen. Sie vereinen Leitstruktur- oder Wirkstoffkriterien mit transportrelevanten Kriterien.[43, 42] Die Eigenschaften antibakterieller Wirkstoffe widersprechen praktisch jedem klassischen Druglike-Kriterium. Sie sind im Vergleich zu herkömmlichen Wirkstoffen größer und komplexer. Besonders natürliche Wirkstoffe, zeigen ein erhöhtes Molekulargewicht, sind polarer und besitzen mehr Ringe.[44] Aufgrund des unterschiedlichen Aufbaus eukaryotischer und prokaryotischer Zellen wirken sie dennoch. Um Bakterien anzugehen, muss die anders strukturierte Zellwand überwunden werden. Dafür sind relaxierte Wirkstoffkritierien notwendig. Bei der Entwicklung antibakterieller Wirkstoffe muss außerdem berücksichtigt werden, ob Gram-positive oder Gram-negative Bakterien anvisiert werden sollen. Differenzierte cLogP-Werte favorisieren in beiden Fällen den Wirkstofftransport durch die unterschiedlich aufgebauten Zellwände.[45]

2.5.4 Unerwünschte Verbindungen

Nicht alle im HTS detektierten Treffer sollten bei der Entwicklung weiterverfolgt werden. Manche Verbindungen zeigen fälschlicherweise Aktivität. Sie werden unter dem Begriff *Screening-Artefakte* zusammengefasst. Andere Verbindungen interagieren auch

mit nicht anvisierten Proteinen. Sie werden als *pharmakologisch wahllose Liganden* bezeichnet.

Screening-Artefakte: Screening-Artefakte zeigen reproduzierbare Aktivität in einem initialen, versagen aber in versierteren Assays.[42] Grund für die falsche Aktivität ist nicht die direkte Interaktion mit dem Zielprotein, sondern eine störende Interaktion mit dem Testsystem, sodass ein positives Resultat detektiert wird. Ein Beispiel hierfür sind Farbstoffe, die in einem kolorimetrischen Assay ausgelesene Wellenlängen ändern und so als falsch-positiv oder falsch-negativ detektiert werden können. Auch Verbindungen, die in Lösung aggregieren, sind problematisch, da sie ein falsches Ergebnis induzieren können. In zellulären Assays können Verbindungen auch nicht mit der Zielstruktur direkt, sondern über den Umweg eines anderen zellulären Proteins, eine Kaskade induzieren, die zuletzt zum selben, positiv gewerteten Effekt führt. Außerdem können toxische Verbindungen in zellulären Testsystemen zu einer Reduzierung der Zellzahl und des Signals führen, was dann fälschlicherweise als Aktivität interpretiert wird.

Pharmakologisch wahllose Liganden: Polypharmakologie kann ein gewünschter Effekt sein, um mehrere Zielstrukturen simultan anzugehen.[46] Sie kann aber auch zu unvorhersehbaren Seiteneffekten führen. In den meisten Fällen sollte ein Wirkstoff deshalb selektiv für ein spezifisches Zielprotein sein. Es gibt Schätzungen, dass Wirkstoffe dennoch durchschnittlich mit 6,3 Zielstrukturen im niedermikromolaren Bereich interagieren.[47] Ein wahlloser Ligand kann an verwandte oder an nicht verwandte Zielstrukturen binden. Ein Beispiel für den ersten Fall sind die humanen Proteinkinasen, die in etwa 500 verschiedenen Varianten existieren. In einem zellulären Assay, kann ein Kinase-Inhibitor leicht das normale Zellverhalten ändern und somit Screening-Artefakte auslösen. Sollen nicht explizit Kinasen anvisiert werden, dann sollten Kinase-Inhibitoren deshalb vom Screening ausgeschlossen werden. Einige Zielstrukturen binden *per se* wahllos Liganden. Beispielsweise binden metabolische Cytochrome (CYPs) und auch der hERG-Kanal bevorzugt lipophile und geladene Liganden.[42] Eine Interaktion mit diesen Proteinen möchte man aber gewöhnlicherweise vermeiden, da deren Modulation Auswirkungen auf den Wirkstoffmetabolismus haben und zu dramatischen Randeffekten führen kann.

2.6 Pharmakophore

Nach dem Schlüssel-Schloss-Prinzip rufen Moleküle mit ähnlicher Form und ähnlich angeordneten physikochemischen Eigenschaften eine ähnliche Wirkung hervor. Dies impliziert, dass bezüglich einer Zielstruktur bioaktive Moleküle im Komplex auch ähnliche

Interaktionsmuster zeigen müssen. Ein Beispiel hierfür sind die in Abbildung 2.1 dargestellten Interaktionsmuster zwischen dem α-Adrenorezeptor und dem Neurotransmitter Epinephrin bzw. dessen kompetitiven Wirkstoffs Clonidin (nach Wermuth[48]). Beide etablieren über unterschiedliche funktionelle Gruppen dieselben Interaktionen zur Rezeptorseite: eine ionische Bindung, eine Wasserstoffbrücke und eine π-π-Interaktion.

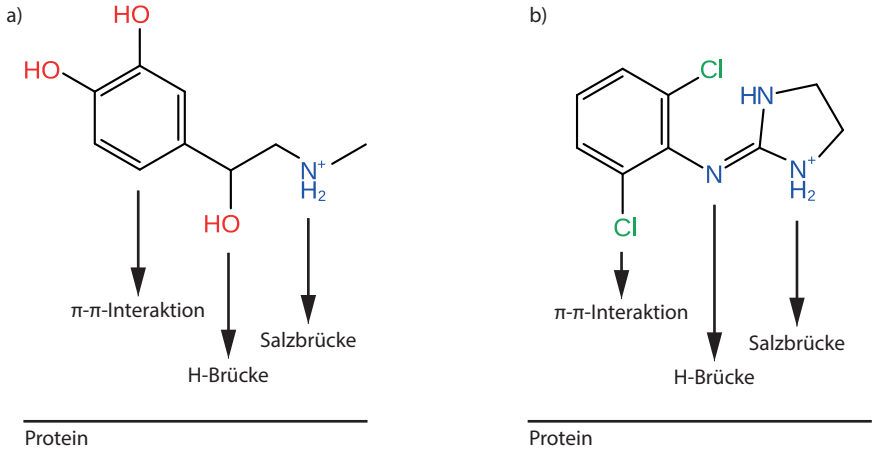


Abbildung 2.1: Sowohl a) Epinephrin als auch b) Clonidin etablieren die selbe Art Interaktionen zum α-Adrenorezeptor (nach Wermuth[48]).

Pharmakophore sind ein sehr populäres Mittel, um diesen Sachverhalt zu beschreiben. Das zugrundeliegende Konzept lässt sich auf die Arbeit von Lemont B. Kier aus den 1960/70er Jahren zurückführen.[49, 50] Van Drie bietet einen kurzen und unterhaltsamen Abriss über die historische Entwicklung und Etablierung des Pharmakophorbegriffs.[51] Um dessen einheitliche und korrekte Verwendung zu propagieren, definierte 1998 auch die IUPAC den Begriff:[52]

> Ein Pharmakophor (pharmakophores Muster) ist die Zusammenstellung sterischer und elektronischer Merkmale, die notwendig ist, um optimale supramolekulare Interaktionen mit einer spezifischen biologischen Zielstruktur zu gewährleisten und ihre biologische Antwort auszulösen (oder zu blockieren). Ein Pharmakophor repräsentiert nicht ein reales Molekül oder eine reale Zusammenstellung von funktionellen Gruppen, sondern ist ein absolut abstraktes Konzept, das für eine Gruppe von Verbindungen gemeinsame molekulare Interaktionsfähigkeiten zur Zielstruktur darstellt. Das Pharmakophor

> kann als der größte gemeinsame Nenner, der von einer Menge von aktiven Molekülen geteilt wird, betrachtet werden. (aus dem Englischen übersetzt[1])

Ein Pharmakophor setzt sich demnach aus einer Beschreibungen der *sterischen* und *elektronischen* Merkmale bioaktiver Moleküle zusammen. Die sterischen Merkmale beschreiben den Raum, den Liganden bei der Bindung einnehmen dürfen. Um ungünstige Überlappungen von Protein- und Ligandatomen zu vermeiden, ist dies ein Raum, der möglichst nicht durch die Zielstruktur beansprucht wird. Zudem sollen Molekülatome Räume einnehmen, die günstige Interaktionen zur Zielstruktur favorisieren. Im besten Fall beschreiben die sterischen Merkmale somit den Raum der Überlappungsfreiheit zwischen den Interaktionspartnern garantiert und zugleich die Anzahl günstiger Kontakte maximiert. Elektronische Merkmale beschreiben die physikochemischen Eigenschaften bioaktiver Moleküle, die zur Bildung von Interaktionen zur Zielstruktur notwendig sind. Im besten Fall sind sie komplementär zu den physikochemischen Eigenschaften der Zielstruktur, um eine Interaktion zu ermöglichen. Zudem ist ihre räumliche Anordnung derart, dass Interaktionen mit optimalen Geometrien gebildet werden können.

Anhand eines einzelnen Liganden kann nicht ohne Weiteres festgestellt werden, welche Merkmale für die Auslösung der biologischen Antwort verantwortlich sind. Unterschiedliche Liganden können weniger, mehr oder andere Interaktionen zur Zielstruktur etablieren und unterschiedliche Bereiche innerhalb der Proteinbindetasche einnehmen. Die definitionsgemäße Spezifikation eines Pharmakophors erfordert deshalb gemeinsame Merkmale, anhand einer Menge bioaktiver Moleküle zu identifizieren. Dadurch kann auf die Merkmale geschlossen werden, die zur Auslösung der biologischen Antwort mutmaßlich notwendig sind. Das Pharmakophor ist allerdings nur dann vollständig beschrieben, wenn zur betrachteten Zielstruktur alle Liganden bekannt sind. Nur dann können alle okkupierten Bereiche und Interaktionsmöglichkeiten beobachtet und die Gesamtheit an sterischen und elektronischen Merkmalen objektiv abgeleitet werden. Aufgrund des nahezu unendlich großen Molekülraums und der daraus resultierenden Ungewissheit, ob für eine Zielstruktur weitere bioaktive Moleküle existieren, ist die Beschreibung eines Pharmakophors in der Praxis nie ideal. Sie stellt nur ein mögliches Modell eines Pharmakophors bezüglich einer beschränkten Anzahl bereits bekannter Liganden dar.

[1] Pharmacophore (pharmacophoric pattern): A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response. A pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their target structure. The pharmacophore can be considered as the largest common denominator shared by a set of active molecules.

Aus diesem Grund spricht man von einem *Pharmakophormodell* oder einer *Pharmakophorhypothese*. Es beschreibt ähnlich angeordnete Molekülmerkmale, jedoch nicht deren zugrundeliegende Muster von Atomen und Bindungen. Aufgrund dieser Abstraktion, ermöglicht eine Pharmakophorhypothese neuartige Liganden mit anderem Molekülgerüsten aber bioisosterisch vergleichbaren Molekülgruppen zu identifizieren.

2.7 Molekulare und makromolekulare Zustände

In dieser Arbeit werden als molekulare und makromolekulare Zustände, Ausprägungsformen eines kleinen Moleküls bzw. eines Proteins bezeichnet, die unterschiedliche chemische und physikalische Eigenschaften aufweisen. Hierzu gehören insbesondere Protonierungszustände und Tautomere, die in dieser Arbeit unter dem Begriff *Protomere*[1] zusammengefasst sind, und im weiteren Sinne auch Konformere.

2.7.1 Protonierungszustände

Protonierungen und Deprotonierungen treten in den meisten Säure-Base-Reaktionen auf. Eine Protonierung ist die Addition eines Protons (H+) zu einem Atom, Ion, Molekül oder Protein unter der Formierung einer konjugierten Säure. Umgekehrt ist eine Deprotonierung die Subtraktion eines Protons unter Formierung einer konjugierten Base. Durch eine Protonierung bzw. Deprotonierung ändert sich die Masse und Ladung um eine Einheit. Manche Ionen und Moleküle können auch mehre Protonierungen erfahren und so einen von mehreren mögliche Zuständen ausbilden. Der Vorgang verläuft meist sehr schnell und ist im Normalfall reversibel. Die Rate der (De-)protonierungen kann jedoch auch gering sein, vor allem wenn der Zustandswechsel eine signifikante strukturelle Änderung beinhaltet. Da sie unterschiedliche physikochemischen Eigenschaften aufweisen, spielen Protonierungszustände eine wichtige Rolle bei der Protein-Ligand-Bindung.[53] Insbesondere kann ein Molekül bzw. Protein je nach Zustand als potentieller Wasserstoffbrückendonor oder -akzeptor fungieren. Damit (De-)protonierungen stattfinden, muss ein Mangel bzw. Überschuss freier Protonen in der Umgebung einer ionisierbaren Gruppe vorzufinden sein. Aus diesem Grund sind (De-)protonierungen pH-abhängig. Betrachtet man Proteine, so können die ionisierbaren Aminosäuren Gln, Asn, Asp, Glu, Lys, Arg und His prinzipiell ihren Protonierungszustand ändern (vgl. Tabelle 2.1). Bei einem physiologischem pH liegen Gln und Asn aber bevorzugt neutral, Asp und Glu als Säure und Lys und Arg als Base vor. Histidin nimmt eine Sonderrolle ein.

[1] Gewöhnlicherweise bezeichnet der Begriff *Protomer* eine strukturelle Einheit eines oligomeren Proteins.

Aufgrund seines pKa-Werts, der nahe am physiologischen pH liegt, kann diese Aminosäure besonders leicht protoniert werden und nahezu beliebig einen neutralen (HID und HIE) oder protonierten (HIP) Zustand einnehmen (vgl. Abbildung 2.2).

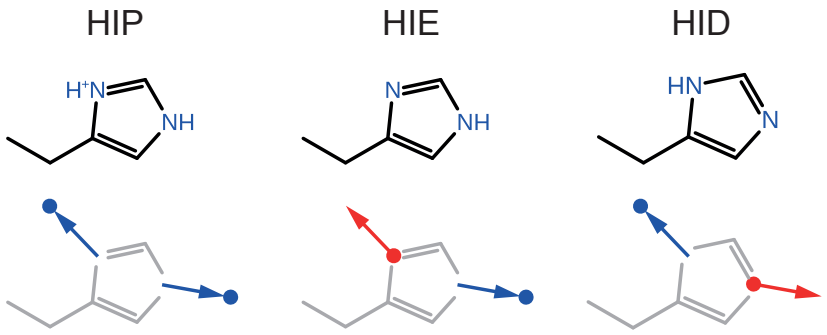


Abbildung 2.2: Wahrscheinliche Zustände von Histidin: HIP bezeichnet den protonierten Zustand von Histidin. HID und HIE sind neutrale, tautomere Zustände, bei denen das Wasserstoffatom am ND1- bzw. NE2-Atom zu finden ist. Die unterschiedlichen Zustände führen zu einer unterschiedlichen Anzahl und Anordnung von Wasserbrückendonoren (blau) und -akzeptoren (rot).

2.7.2 Tautomere

In manchen Fällen, wie bei Histidin, wird die Protonierung von Isomerie begleitet (vgl. Abbildung 2.2). Als *Isomere* werden Moleküle mit gleicher Summenformel aber unterschiedlicher Strukturformel bezeichnet. Eine Art von Isomerie ist die *Konstitutionsisomerie*. Konstitutionsisomere unterscheiden sich in der Reihenfolge von Atomen und Bindungen. Zu dieser Klasse gehören auch *Tautomere*. Sie unterscheiden sich nur durch die Stellung eines Atoms oder Atomgruppe. Prototropie ist die häufigste Form der Tautomerie. Protorope Tautomere sind Tautomere, bei denen lediglich ein Wasserstoffatom eine andere Position einnimmt und benachbarte Einfach- und Doppelbindungen ihren Platz tauschen. Bei anderen tautomeren Formen können auch Anionen ihre Stellung ändern. Prototrope Tautomere können als spezielle Form von Protonierungszuständen unter Erhaltung der Gesamtladung betrachtet werden. Sie stehen miteinander in einem chemischen Gleichgewicht und können innerhalb einer reversiblen chemischen Reaktion ineinander übergehen. Aufgrund ihrer raschen Umwandlung, betrachtet man sie als ein und dieselbe chemische Verbindung. Durch den Zustandswechsel ändert sich allerdings die Anordnung potentieller Wasserstoffbrückendonoren und -akzeptoren. Die Mengenverhältnisse der Tautomere sind untereinander konstant. Das genaue Verhältnis hängt jedoch von verschiedenen Faktoren wie Temperatur, Lösung und pH-Wert ab.[54]

Typische tautomere Paare sind Amid und Imid, Amin und Imin, Enamin und Imin und Lactam und Lactim. Manche Tautomerien ändern die Gestalt eines Moleküls. Dies ist insbesondere bei der Keto-Enol-Tautomerie und bei Ring-Ketten-Tautomerien der Fall.

2.7.3 Konformere

Eine andere Art der Isomerie ist die *Stereoisomerie*. Bei Stereoisomeren ist die Topologie der Atome gleich, aber deren räumliche Anordnung verschieden. Stereoisomere lassen sich weiter in *Konfigurationsisomere* und *Konformationsisomere* unterscheiden. Konfigurationsisomere unterscheiden sich in der räumlichen Anordnung der Atome, die nicht durch eine simple Drehung von Atomen oder Atomgruppen an Einfachbindungen entstehen. Zur Konfigurationsisomerie zählen beispielsweise die cis-trans-Isomerie, die Enantiomerie und die Diastomerie. Die Stereoisomere, die durch Drehungen um Einfachbindungen (rotierbare Bindungen) entstehen, werden als *Konformationen* bezeichnet. Auch hier stehen die unterschiedlichen Formen in einem chemischen Gleichgewicht. Konformationen, die besonders häufig vorkommen, weil sie energetische günstiger als andere sind, werden als *Konformere* bezeichnet. Sowohl kleine Moleküle als auch Proteine mit rotierbaren Einfachbindungen können Konformere ausbilden. Die Änderung von terminalen funktionellen Gruppen an Einfachbindungen ist energetisch betrachtet die einfachste Konformationsänderung. Auf die äußere Gestalt wirkt sich diese Änderung kaum aus, allerdings resultiert sie häufig in einer anderen räumlichen Ausrichtung potentieller Wasserstoffbrückendonoren und -akzeptoren. Zentralere rotierbare Einfachbindungen bewirken zudem, dass sich die äußere Form von Molekülen substantiell ändert. Proteine können ebenso Konformationsänderungen erfahren. Zum einen können einzelne Seitenketten der Aminosäuren, insbesondere langkettige Aminosäuren wie Lys, zum anderen auch Schleifen des Proteinrückgrats oder sogar ganze Proteindomänen ihre Konformation ändern. Nach der Induced-Fit-Theorie[55] kann die Bindung eines Liganden eine solche Konformationsänderung im Protein hervorrufen.

3 Methoden des computergestützten Wirkstoffentwurfs

Die Entwicklung eines neuen Wirkstoffes ist ein langwieriger und kostenintensiver Prozess. In der Regel werden 10 bis 15 Jahre und etwa 100 Millionen US-Dollar investiert, bevor ein Medikament zugelassen wird.[56] Der Kostenaufwand steigt von den frühen zu den späten Entwicklungsphasen stetig an. Jederzeit besteht die Gefahr, dass ein Kandidat den Anforderungen an einen Wirkstoff nicht genügt und verworfen werden muss. Da klinische Phasen in der Öffentlichkeit Beachtung finden, kann ein spätes Scheitern neben finanziellen Verlusten auch zu Imageschäden führen. Deshalb gilt bei der Wirkstoffentwicklung die Prämisse: „Versage früh, versage schnell, versage günstig."[57] Die frühen Phasen des rationalen Wirkstoffentwurfs werden daher von einer Reihe computergestützter Methoden begleitet, um unzureichende Kandidaten früh zu erkennen und einen unnötigen Einsatz aufwändigerer experimenteller Methoden zu vermeiden. Computergestützte Methoden sind mittlerweile integrale Bestandteile des Entwicklungsprozesses[58] und haben nachweislich zum erfolgreichen Entwurf von Wirkstoffen beigetragen.[59] Dieses Kapitel stellt *in silico* Methoden zum virtuellen Screening (VS) vor, ein Verfahren, das hauptsächlich während der Treffer- und Leitstruktursuche zur Anwendung kommt.

3.1 Virtuelles Screening

2013 definierte die IUPAC den Begriff des *virtuellen* oder *in silico Screenings* als „die Bewertung von Verbindungen unter Nutzung von Computermethoden" und merkte an, dass „der Ursprung des Bewertungsmodells eine makromolekulare Struktur sein kann oder auf physikochemischen Parametern oder Struktur-Aktivitäts-Beziehungen basieren kann." (aus dem Englischen übersetzt[1]).[60]

[1] virtual screening/in silico screening: Evaluation of compounds using computational methods. Note: The source of the model could be a macromolecular structure or based on physicochemical parameters

Im Gegensatz zu einem experimentellen HTS bietet ein VS den Vorteil, dass virtuelle Molekülbibliotheken getestet werden können und Materialkosten zur Synthese von Substanzen und Erstellung von Testsystemen entfallen. Berücksichtigt man nur organische Verbindungen mit einem Molekulargewicht von weniger als 500 Da, umfasst der chemische Raum allerdings bereits $\sim 10^{120}$ Moleküle. Für Computermethoden ist eine vollständige Suche durch solch einen Raum nicht praktikabel. Die praktische Anwendung eines VS beschränkt sich daher darauf, neben der Bewertung von Molekülen, die Bibliothek auf eine handhabbare Größe zu reduzieren, sodass Substanzen schließlich synthetisiert oder erworben und experimentell getestet werden können.[61] Ein virtuelles Abbild einer kombinatorischen Bibliothek, des Sortiments eines Lieferanten oder einer betriebseigenen Substanzsammlung wird nach potentiellen Kandidaten durchsucht und die Bibliothek bezüglich der betrachteten Fragestellung beschränkt. Dabei sind VS-Methoden mit Bibliotheken im Umfang von 10^6–10^{12} Molekülen konfrontiert.[62]

Die zentrale Fragestellung in einem VS ist stets die, neuartige chemische Strukturen zu identifizieren, die die Funktion des anvisierten makromolekularen Ziels modulieren. In Abhängigkeit der Datenlage, d. h. welches Wissen über das Zielprotein und Liganden bereits vorhanden ist, realisieren VS-Methoden verschiedene Strategien:

- Ein *ligandbasiertes virtuelles Screening* (LBVS) setzt voraus, dass Information über bioaktive Moleküle zur betrachteten Zielstruktur gegeben ist.
- Ein *strukturbasiertes virtuelles Screening* (SBVS) fordert eine aufgeklärte dreidimensionale Struktur des Zielproteins.
- Ein *pharmakophorbasiertes virtuellen Screening* (PBVS) bedarf einer zuvor postulierten Hypothese über typischerweise etablierte Protein-Ligand-Interaktionen.
- Sind Zielstruktur aufgeklärt, bioaktive Moleküle bekannt und/oder Pharmakophorhypothesen postuliert, können *integrierte VS-Ansätze* LBVS-, SBVS- und PBVS-Strategien kombinieren, um die Bibliothek sukzessiv zu beschränken.

Welche Strategie zur Anwendung kommt, hängt neben der Datenlage auch davon ab, welches konkrete Ergebnis von einem VS erwartet wird. Welche Art von Molekülen soll extrahiert werden? Sollen analoge oder völlig neuartige Molekülgerüste identifiziert werden? cRAISE, das im Zuge dieser Arbeit entstand, vereint Konzepte verschiedener Screening-Strategien. Deshalb werden grundlegende Prinzipien und ausgewählte Ansätze, die zum Verständnis dieser Arbeit beitragen, im Folgenden vorgestellt. Für einen ausführlichen Überblick sei auf [63] und [64] verwiesen.

or ligand structure–activity relationships.

3.2 Ligandbasiertes virtuelles Screening

Das Prinzip eines LBVS beruht auf der Annahme, dass ähnliche Moleküle ähnliche Eigenschaften und somit eine ähnliche biologische Aktivität zum Zielprotein zeigen. Um ein LBVS durchzuführen, ist zumindest ein bereits bekannter Ligand (*Referenzligand*) notwendig. Zu ihm werden dann ähnliche Moleküle einer Bibliothek bestimmt. Das Ergebnis eines LBVS oder einer *Ähnlichkeitssuche* ist eine Liste von Verbindungen aus der Bibliothek, die entsprechend der berechneten Ähnlichkeit sortiert ist (*Hitliste*). Ein detaillierter Überblick über LBVS-Methoden ist u. a. durch [65], [66] und [67] gegeben.

3.2.1 Das Problem der molekularen Ähnlichkeit

Molekulare Ähnlichkeit lässt sich bezüglich der Konstitution, der Topologie oder der 3D-Struktur ermitteln. Ein LBVS ist dabei mit unterschiedlichen Problemen konfrontiert:

Konstitutionelle Ähnlichkeit: Eine konstitutionelle Ähnlichkeit ist vorhanden, sobald Moleküle eine ähnliche atomare Zusammensetzung besitzen. Wird eine ähnliche Anzahl korrespondierender Elemente in Molekülen vorgefunden, kann dies auf ähnliche physikochemische Eigenschaften hinweisen. Im Vergleich zur Bewertung der topologischen und strukturellen Ähnlichkeit ist ein konstitutioneller Vergleich relativ einfach zu bewerkstelligen. Dafür muss lediglich die Anzahl enthaltener Atome verglichen werden, die bereits aus den Summenformeln der Moleküle entnommen werden kann.

Topologische Ähnlichkeit: Die topologische Bewertung von Molekülen beruht auf der Idee, die größte gemeinsame Substruktur (Maximum common substructure, MCS) in den zugehörigen Molekülgraphen zu identifizieren. Ein Molekülgraph repräsentiert jedes Atom durch einen Knoten und die Bindungen durch entsprechende Kanten. Die Kardinalität der größten gemeinsamen Substruktur kann als Maß für die Ähnlichkeit von Molekülen dienen. Die Bestimmung der MCS ist ein Graphproblem mit der Aufgabe, eine topologie- und typerhaltende Zuordnung von Atomen zu identifizieren. Es kann durch Finden einer maximalen Clique im Assoziationsgraph[1] G_A gelöst werden:

> **Max-Clique-Problem:** Sei ein ungerichteter Graph $G = (V, E)$ gegeben. Eine Clique C ist eine Teilmenge von V, so dass für jedes Paar von Knoten $u, v \in C$ gilt $(u, v) \in E$. Das Auffinden der größten Clique in G ist ein NP-vollständiges Problem, das als Max-Clique-Problem (MCP) bezeichnet wird.

[1] Assoziationsgraph G_A: G_A enthält einen Knoten für jedes Knotenpaar $\{v_1, v_2\}$ aus den Molekülgraphen $G_1 = (V_1, E_1, f_1)$ und $G_2 = (V_2, E_2, f_2)$ mit identischen Knotenbezeichnungen $f_1(v_1) = f_2(v_2)$. Zwei Knoten in G_A sind adjazent, wenn sie entweder in G_1 und G_2 oder weder in G_1 noch in G_2 über eine Kante verbunden sind.

Zur Lösung des MCP werden in der Chemieinformatik gewöhnlicherweise Varianten des Bron-Kerbosch-Algorithmus[68] eingesetzt.

Strukturelle Ähnlichkeit: Moleküle sind strukturell ähnlich, wenn sie ähnliche Konformationen aufweisen. Um einen Vergleich zu ermöglichen, muss dafür das Problem der molekularen Überlagerung gelöst werden. Dies bedeutet, dass eine Transformation $\mathbf{T}_{\min}$ gefunden werden muss, die die Summe der Abstandsquadrate zwischen definierten Atompaaren (x_i, y_i) des auszurichtenden Moleküls x und der Referenz y minimiert:

$$\mathbf{T}_{\min} = \arg\min_{\mathbf{T}} \sum_{i}^{N} (\mathbf{T}x_i - y_i)^2 \tag{3.1}$$

Für identische, rigide Moleküle ist eine Zuordnung zwischen den Atomen x_i und y_i und die Transformation $\mathbf{T}_{\min}$ relativ einfach zu bestimmen, z. B. mit Hilfe des Kabsch-Algorithmus[69, 70]. In einem VS ist man jedoch mit Molekülen unterschiedlicher Art konfrontiert, für die eine Atomzuordnung nicht immer offensichtlich ist. Deshalb muss auch im dreidimensionalen Fall zunächst das MCS-Problem gelöst werden.[71] Bei Bestimmung der Transformation muss zudem die Flexibilität der Moleküle berücksichtigt werden. Es existieren verschiedene Strategien zum Auffinden einer optimalen molekularen Überlagerung. Für einen Überblick algorithmischer Konzepte sei auf [72] verwiesen. Man kann diese Methoden, analog zu den in Abschnitt 3.3.4 vorgestellten Docking-Strategien, bezüglich der Handhabung der Ligandflexibilität klassifizieren.

3.2.2 Deskriptorbasiertes virtuelles Screening

Die Lösung des MCS-Problems bzw. des Problems der molekularen Überlagerung ist oftmals zu aufwändig, um effizient in einem VS eingesetzt werden zu können. LBVS-Methoden verfolgen daher häufig einen vereinfachten Ansatz, der die explizite Berechnung der MCS bzw. der Transformation umgeht: Ein Molekülvergleich auf Basis der Topologie kann dadurch realisiert werden, dass kleine vordefinierte Substrukturen oder *Fragmente* in Molekülen identifiziert werden. Je mehr gemeinsame Fragmente gefunden werden, desto höher ist die Wahrscheinlichkeit, dass Moleküle eine große MCS teilen. Analog dazu kann ein Molekülvergleich auf Basis der 3D-Struktur dadurch realisiert werden, dass kleine vordefinierte räumliche Anordnungen charakteristischer Strukturmerkmale in Molekülen identifiziert werden. Ähnliche Konstellationen der Merkmale können ein Indiz dafür sein, dass sich Moleküle gut überlagern lassen. Somit kann die Ähnlichkeit zweier Moleküle als Grad der Überlappung ihrer charakteristischen Merkmale gesehen werden. Charakteristische Molekülmerkmale können in Form von *Deskriptoren* kodiert werden. Die Ähnlichkeit zum Referenzmolekül wird dann über ein *Ähnlichkeitsmaß* ausgedrückt, das im Deskriptorvergleich angewendet wird. Abbildung 3.1

skizziert die wesentlichen Bestandteile eines solchen deskriptorbasierten LBVS. Welche Merkmale als charakteristisch ausgewiesen werden und somit die Ähnlichkeit bestimmen, hängt von der betrachteten Problemstellung ab. Manche Merkmale eignen sich besser um chemische Substanzklassen zu detektieren, andere beschreiben elektrostatische Eigenschaften, Löslichkeit, Säure-/Basestärke oder die Form von Molekülen besser.

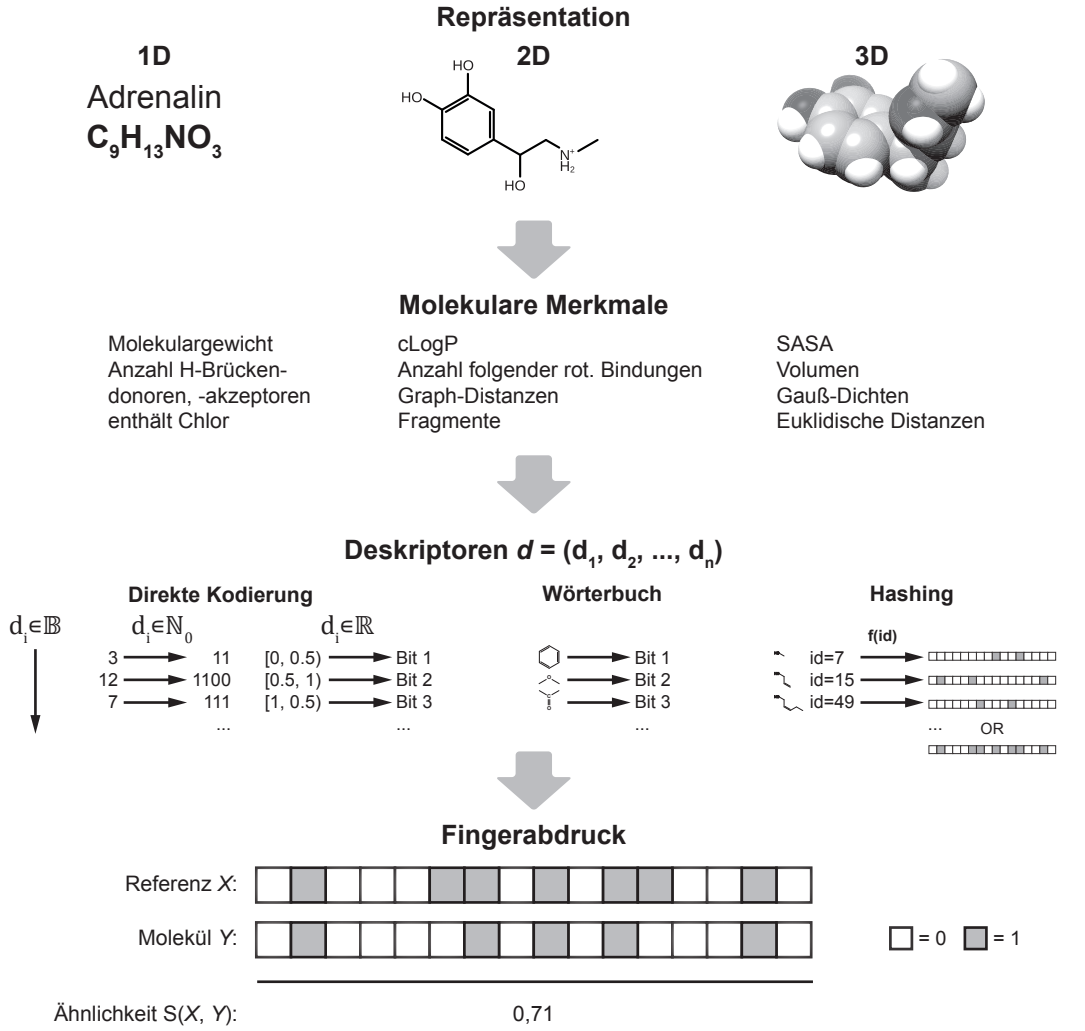


Abbildung 3.1: Molekülmerkmale werden von der 1D-, 2D- oder 3D-Repräsentation eines Moleküls berechnet. Molekulare Deskriptoren sind eine Zusammenstellung dieser Merkmale, die direkt, wörterbuchbasiert oder mit einer Hash-Funktion binär in einem Fingerabdruck kodiert werden. Ein Ähnlichkeitsmaß, vergleicht die Fingerabdrücke.

3.2.3 Molekulare Deskriptoren

Es existieren über 3000 molekulare Deskriptoren, die aus einer Zusammenstellung berechneter Moleküleigenschaften oder -merkmale bestehen.[64] Sie werden anhand der Dimension der Molekülrepräsentation, auf der die Berechnung erfolgt, klassifiziert:[73]

- Ein *1D-Deskriptor* beschreibt eine globale Eigenschaft eines Moleküls, die anhand seiner Konstitution direkt abgeleitet werden kann.
- Ein *2D-Deskriptor* beschreibt eine Eigenschaft, die anhand der chemischen Struktur ohne Berücksichtigung von 3D-Koordinaten bestimmt werden kann.
- Ein *3D-Deskriptor* beschreibt eine Eigenschaft, die sich ausschließlich mit Hilfe von 3D-Atomkoordinaten berechnen lässt.

Ein 1D-Deskriptor kann beispielsweise das Molekulargewicht sein, das direkt aus der Summenformel eines Moleküls berechnet werden kann. 1D-Deskriptoren und Deskriptoren, die numerische Werte kodieren, werden oft als Filter genutzt, um Moleküle bezüglich der Wirkstoff- oder Leitstrukturkriterien zu bewerten (vgl. Abschnitt 2.5.1 und 2.5.2). Die Anzahl aromatischer Bindungen oder der Oktanol-Wasser-Koeffizient cLogP können anhand der Topologie eines Moleküls abgeleitet und durch einen 2D-Deskriptor repräsentiert sein. 2D-Deskriptoren, die vollständig die Topologie eines Moleküls beschreiben, z. B. kanonische SMILES (Simplified Molecular Input Line Entry System), können Filter etablieren, um unerwünschte Verbindungen in Molekülbibliotheken zu identifizieren und auszuschließen (vgl. Abschnitt 2.5.3 und 2.5.4). Viele 2D-Deskriptoren registrieren charakteristische Molekülfragmente: MACCS- und BCI-Deskriptoren[74] beschreiben ausschließlich die Existenz vordefinierter Fragmente. Dagegen enumeriert der Daylight-Fingerabdruck[75] unabhängig alle eindeutigen Subgraphen im Molekülgraph bis zu einer gegebenen Größe. ECFPs (Extended-Connectivity Fingerprint) kodieren Atomtypen, Ladungen und Bindungseigenschaften der zirkulären Umgebung einzelner Atome. Sie erzielen gute Hit-Raten, allerdings sind die Screening-Resultate wenig divers.[76, 77] Um die Diversität zu erhöhen, muss der Molekülvergleich von der chemischen Struktur abstrahieren: FCFPs (Function-Class Fingerprint) sind eine Variante der ECFPs, die Atome auf funktionelle Merkmale wie Wasserstoffbrückendonoren oder -akzeptoren, negativ oder positiv ionisierbar sowie aromatisch oder halogenoid abstrahieren.[78] Der CATS-Deskriptor[79] weist jedem Atom eines von fünf funktionellen Merkmalen zu, misst paarweise Merkmalsdistanzen und vergleicht Histogramme der Distanzen, die in einem topologischen Fingerabdruck kodiert sind. Die Abstraktion von der chemischen

Struktur ermöglicht einen Grundgerüstwechsel (*Scaffold hopping*) während des Screenings und erleichtert so das Auffinden neuer Chemotypen. Die Molekülform und die spezifische räumliche Anordnung elektrostatischer Oberflächenmerkmale berücksichtigen 2D-Deskriptoren jedoch nicht. 3D-Deskriptoren wie der CATS3D[80], SURFCATS[81] oder ROCS[82, 83, 84] vergleichen Moleküle bezüglich dreidimensionaler Merkmale. ROCS führt einen reinen Formvergleich von Molekülen durch. Die Form einer Konformation wird durch atomzentrierte Gauß-Funktionen kontinuierlich beschrieben. Sie können während des Screenings effizient verglichen werden, um das Überlappungsvolumen zur Referenz zu bewerten. Ein formbasiertes VS kann bereits gute Hit-Raten liefern.[85] Die gängigsten 3D-Deskriptoren sind allerdings pharmakophorartiger Natur. Sie erweitern die Repräsentation der Form durch Beschreibungen der räumlichen Anordnung von Pharmakophormerkmalen und werden oft entsprechend der in Abschnitt 3.4.1 vorgestellten Pharmakophorfingerabdrücke kodiert. Weitere Beispiele für 3D-Deskriptoren sind das van-der-Waals-Volumen oder der wasserzugängliche Oberflächenbereich (SASA), zu deren Berechnung die Kenntnis der Atompositionen im Raum notwendig ist. Der Vorteil von 1D- und 2D-Deskriptoren ist ihre Unabhängigkeit von der Molekülkonformation. 3D-Deskriptoren erfordern die Erzeugung von Konformationen und sind daher aufwändiger zu berechnen. Allerdings bilden sie realistischer räumliche, molekulare Eigenschaften ab, die bei der Protein-Ligand-Bindung ausschlaggebend sind.

3.2.4 Fingerabdrücke

Molekulare Deskriptoren werden oft binär, in Form von *Fingerabdrücken* kodiert. Dies hat den Vorteil, dass molekulare Merkmale kompakt repräsentiert, vorberechnet und ein darauffolgender Deskriptorabgleich rasch durchgeführt werden kann. Fingerabdrücke sind Bit-Vektoren, wobei ein Bit das Vorhandensein oder Fehlen eines charakteristischen Merkmals anzeigt. Auch numerische Werte können in einem Fingerabdruck dargestellt werden. Die einzelnen Einträge der Fingerabdrücke können bei der Kodierung direkt, wörterbuchbasiert oder über eine Hash-Funktion zugewiesen sein:

Direkte Kodierung: Werte diskreter Variablen mit beschränktem Wertebereich können auf direktem Wege binär kodiert und in reservierten Position(en) im Fingerabdruck gespeichert werden. Werden mehrere Variablen kodiert, muss ausschließlich ein Versatz berücksichtigt werden, der einer Variablen einen bestimmten Bereich des Fingerabdrucks zuordnet. Bei kontinuierlichen Variablen müssen initial Wertebereiche vordefiniert und jedem Bit zugeordnet werden (*Binning*). Fällt ein Wert in einen vordefinierten Bereich, wird das Bit an der entsprechenden Position gesetzt.

Wörterbuchbasierte Kodierung: Wörterbücher ordnen vordefinierten Fragmenten eindeutige Bit-Positionen im Fingerabdruck zu. Solche Fingerabdrücke, wie der MACCS- oder der BCI-Deskriptor, setzen das entsprechende Bit sobald ein Fragment aus dem Wörterbuch im Molekül enthalten ist.[74] Die Größe des Fingerabdrucks hängt somit vom genutzten Wörterbuch ab und das Screening-Resultat ist sehr vom zugrundeliegenden Wörterbuch geprägt. Fragmente, die darin nicht enthalten sind, werden im Deskriptor nicht registriert und daher nicht zum Molekülvergleich herangezogen.

Hash-Funktion: Mit einer Hash-Funktion lässt sich die Verwendung eines Wörterbuchs vermeiden. Die Hash-Funktion der Daylight-Methode erzeugt anhand aller vorgefundenen Fragmente eindeutige Schlüssel, die im Bit-Vektor repräsentiert werden. Dadurch garantiert sie die Darstellung einer variablen Anzahl von Fragmenten in einem Bit-Vektor fixer Länge. Der Daylight-Fingerabdruck kann zudem „gefaltet“ werden, um dichter besetzte und kompaktere Bit-Vektoren zu erhalten.[75]

Fingerabdrücke werden zumeist zur Kodierung von 2D-Deskriptoren genutzt. Das Konzept der Fingerabdrücke kann allerdings auch 3D umgesetzt werden, um die räumliche Anordnung von Pharmakophormerkmalen zu kodieren (siehe Abschnitt 3.4.1).

3.2.5 Bewertung der Molekülähnlichkeit

Während eines deskriptorbasierten LBVS reduziert sich der Vergleich zweier Moleküle stets auf einen Vergleich zweier Vektoren $X = (x_1, x_2, \cdots, x_N)$ und $Y = (y_1, y_2, \cdots, y_N)$ mit jeweils N Attributen. Die Attribute können dabei reellwertige, physikochemische Eigenschaften oder die Bits binärer Fingerabdrücke sein. Auf direktem Wege kann die Ähnlichkeit $S(X, Y)$ zweier Vektoren mit Hilfe von Ähnlichkeitskoeffizienten bestimmt werden. Für identische Moleküle nimmt $S(X, Y)$ den Maximalwert an. Indirekt lässt sich die Ähnlichkeit über Distanzkoeffizienten $D(X, Y)$ bestimmen. Für identische Moleküle ist $D(X, Y) = 0$, sodass die Ähnlichkeit $S(X, Y) = 1 - D(X, Y)$ ist. Ist der Maximalwert der Distanz nicht 1, so muss gegebenenfalls normalisiert werden, sodass sich die Ähnlichkeit im Bereich von $0 \leq S(X, Y) \leq 1$ bewegt und den Überlappungsgrad gemeinsamer Merkmale anzeigt. Ein Überblick und eine Diskussion über molekulare Ähnlichkeits- und Distanzkoeffizienten ist in [86] gegeben. Für Fingerabdrücke wird zumeist der Tanimoto-Koeffizient genutzt. Dieser und andere direkte Ähnlichkeitsmaße hängen von der Anzahl gesetzter Bits ab und bewerten die Existenz gemeinsamer Merkmale. Distanzen wie die Euklidische Distanz betonen dagegen Unterschiede. Daher werden kleine Moleküle, bei denen in der Regel weniger Bits gesetzt sind, durch direkte Ähnlichkeitsmaße bezüglich der Referenz als unähnlicher bewertet wohingegen indirekte

Maße kleine Moleküle ähnlicher bewerten.[65] Welches Maß letztendlich genutzt wird, hängt vom gewünschten Resultat im konkreten Anwendungsszenario ab.

3.3 Strukturbasiertes virtuelles Screening

Nach dem Schlüssel-Schloss-Prinzip (vgl. Abschnitt 2.4) kann ein Molekül Ligand eines Proteins sein, wenn es überlappungsfrei in die Bindetasche passt und seine elektrostatischen Oberflächenmerkmale die des Proteins ergänzen. Voraussetzung zur Bewertung dieser Komplementarität ist die aufgeklärte 3D-Struktur des Zielproteins (auch *Rezeptor* genannt). Ist sie gegeben, so wird in einem SBVS für jedes Molekül der Bibliothek ein *Protein-Ligand-Docking* durchgeführt. Es bewertet ein Molekül bezüglich seiner Komplementarität zur Zielstruktur und fügt das Resultat zur Hitliste des VS-Laufes hinzu. Im Idealfall reflektiert die Bewertung die Bindungsaffinität und findet sich im vorderen Bereich der Hitliste wieder, wenn das Molekül bioaktiv ist. Ein SBVS beschränkt effektiv die Molekülbibliothek und verwirft Moleküle, die nicht in die Bindetasche passen. Im Folgenden werden das Docking-Problem und Strategien etablierter Methoden zu dessen Lösung vorgestellt. Für einen ausführlichen Überblick zu diesem Thema sei auf aktuelle Rezensionen [87, 88, 89, 90, 91, 92] verwiesen.

3.3.1 Das Docking-Problem

> **Docking-Problem:** Sei ein Molekül L und die 3D-Struktur eines Proteins P gegeben. Entscheide ob L an P binden kann. Ist dies der Fall, bestimme die Geometrie von L (den *aktiven Bindemodus*) und die Bindungsaffinität ΔG.

Ist L ein Protein so spricht man von einem *Protein-Protein-Docking*. Andernfalls handelt es sich um ein Protein-Ligand-Docking, das auch als *molekulares Docking* bezeichnet wird. Methoden, die versuchen das Docking-Problem zu lösen, bestehen im Wesentlichen aus zwei Komponenten: Einer *Suchstrategie*, die den Raum möglicher Ligandplatzierungen (*Posen*) innerhalb der Proteinbindetasche durchsucht und einer *Bewertungsfunktion*, die Posen bewertet und ΔG abschätzt. Für ein Molekül generiert ein Docking zumeist mehrere Platzierungsvorschläge, die entsprechend der Bewertung sortiert werden. In einem SBVS wird ausschließlich die Bewertung der besten Pose als Teillösung zur Hitliste hinzugefügt. Folgend wird der Begriff „Docking“ als Synonym für Protein-Ligand-Docking genutzt, das im Fokus dieser Arbeit stand.

3.3.2 Suchraum

Der Suchraum eines Protein-Ligand-Dockings wird durch die in Abbildung 3.2 dargestellten Freiheitsgrade aufgespannt. Da viele Docking-Methoden nicht all diese Freiheitsgrade berücksichtigen, lassen sich die Verfahren in *starre*, *flexible* und *voll-flexible* Methoden unterteilen:

Starres Docking: Starre Ansätze berücksichtigen ausschließlich die Freiheitsgrade, die zur Translation und Rotation des Liganden um das Protein notwendig sind. Der Suchraum eines starren Dockings besitzt somit $d = 6$ Dimensionen.

Flexibles Docking: Flexible Docking-Methoden erlauben neben der Translation und Rotation eine Variation der Ligandkonformation. Besitzt ein Molekül n rotierbare Bindungen, so erstreckt sich der Suchraum über $d = 6 + n$ Dimensionen.

Voll-flexibles Docking: Voll-flexible Ansätze berücksichtigen die Translation und Rotation des Moleküls, Ligandflexibilität und zudem die Flexibilität der Zielstruktur. Besitzt ein Protein m rotierbare Bindungen, so ergibt sich in einem voll-flexiblen Docking ein Suchraum mit $d = 6 + n + m$ Dimensionen.

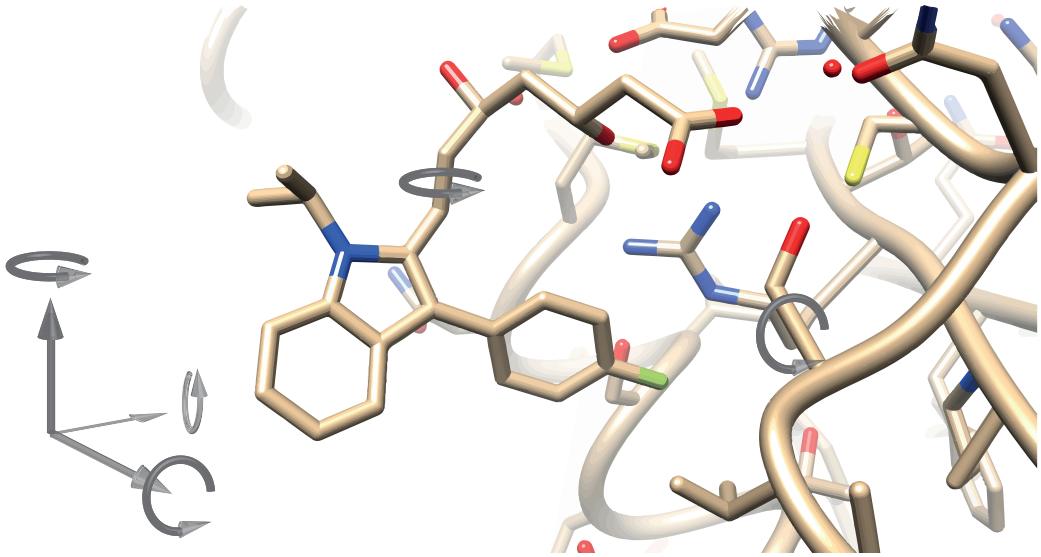


Abbildung 3.2: Der Suchraum eines molekularen Dockings wird durch Freiheitsgrade aufgespannt, die aus der Translation und Rotation des Liganden und aus der Flexibilität von Ligand und Protein resultieren.

Seit der Veröffentlichung von Dock[93], der ersten starren Docking-Methode, im Jahr 1982, wurden Methoden publiziert, die immer mehr Freiheitsgrade integrierten.

Die Berücksichtigung von Ligandflexibilität erschwert das Docking-Problem zwar, ist heutzutage aber *de facto* Standard. Mittlerweile existieren auch Docking-Methoden, die Proteinflexibilität integrieren.[94, 95] Sie modellieren zumindest einzelne flexible Bestandteile des Proteins. Aktuell werden voll-flexible Docking-Methoden im VS aber nur selten angewandt.[96] Die enorme Anzahl zusätzlicher Freiheitsgrade erfordern Rechenkapazitäten, die einen Durchsatz großer Bibliotheken erschweren und zugleich die Selektivität des Zielproteins verringern.

3.3.3 Systematische Beschränkung des Suchraums

Durch die Kontinuität der Freiheitsgrade ist bereits für starre Docking-Methoden ein unendlich großer Suchraum gegeben und das Docking-Problem somit nicht exakt lösbar. Um zumindest eine abschätzende Vorhersage und den Einsatz im VS zu ermöglichen, reduzieren Docking-Methoden deshalb ihren Suchraum mithilfe systematischer Einschränkungen:

Definition der Proteinbindetasche: Die Bindetasche prägt mutmaßlich die Funktion eines Proteins. Daher beschränken sich Docking-Methoden darauf, Posen nur in der Proteinbindetasche zu generieren. Ist eine Proteinstruktur inklusive eines komplexierten Liganden gegeben (*Holoprotein*), dann ist die Bindetasche durch die Proteinatome definiert, die den Referenzligand direkt umgeben. Zumeist wird eine Umgebung gewählt, die Atome in der Entfernung von 5 bis 10 Å vom Referenzliganden einschließt. Ist zur Proteinstruktur keine entsprechende Ligandstruktur gegeben, kann für das *Apoprotein* eine Tasche manuell definiert oder die Bindetasche vorhergesagt werden. Methoden zur Bindetaschenvorhersage bewerten geometrisch oder energetisch die Proteinoberfläche oder analysieren Proteinbereiche bezüglich ihrer evolutionären Konservierung.[97, 98]

Diskretisierung der Freiheitsgrade: Unter der Annahme, dass sich die Güte einer Pose bei leichter Variation kaum ändert, können Variablen zur Translation, Rotation und Torsionsänderung diskretisiert und von Docking-Algorithmen schrittweise variiert werden. Da Bewertungsfunktionen von Atomkoordinaten abhängen, sollte die Granularität der Diskretisierung so fein gewählt werden, dass die Bewertungsfunktion ausreichend genau arbeiten kann. Eine Diskretisierung äußerst sich oft auch in einer reduzierten Molekül- und Proteinrepräsentation: Oberflächenbereiche, Interaktionssphären oder Volumen können durch gleichmäßig verteilte Oberflächen-, Interaktions- oder Gitterpunkte diskretisiert sein, um kontinuierliche Bereiche annähernd zu beschreiben.

Rationale Beschränkung des Suchraums: Die Diskretisierung der Variablen reicht oft nicht aus, sodass der Raum vollständig durchsucht werden kann. Eine Strategie ist

es, die chemische Information des Zielproteins zu nutzen, um den Suchraum systematisch einzuschränken. Es reicht aus, Posen zu erzeugen, die Protein-Ligand-Interaktionen etablieren können. Nur sie haben das Potential den aktiven Bindemodus besonders affiner Liganden widerzuspiegeln. Gerichtete Docking-Methoden[99] typisieren dafür den diskretisierten Raum und bilden Interaktionsstellen in der Proteinbindetasche ab, an die nur Posenatome komplementären Typs platziert werden dürfen. Eine weitere Einschränkung des Suchraums kann dadurch erreicht werden, dass zugleich geometrische Beschränkungen für Interaktionsstellen definiert werden, die während der Platzierung eingehalten werden müssen. Dadurch werden die teils strengen Geometrien typischer Protein-Ligand-Interaktionen modelliert (vgl. Abschnitt 2.4).

Beschränkung des Ligandkonformationsraums: Zur Beschränkung des hochdimensionalen Konformationsraums von Liganden variieren flexible Docking-Methoden Torsionswinkel häufig gemäß der Winkel, die in niederenergetischen Konformeren vorgefunden werden. Diese Winkel können u. a. in sogenannten Torsionswinkelbibliotheken (z. B. [100, 101]) registriert sein, die auf einer statistischen Analyse aufgeklärter Protein-Ligand-Komplexe beruhen. Auf diese Weise wird eine freie Rotation einzelner Bindungen vermieden und der Konformationsraum der Liganden auf typische Torsionen beschränkt.

Beschränkung des Proteinkonformationsraums: Auf ähnliche Weise können vollflexible Docking-Methoden den Konformationsraum des Proteins beschränken. Der Raum flexibler Seitenketten kann ausschließlich auf statistisch häufig vorgefundene, niederenergetische Konformere reduziert werden, wie sie in Rotamerbibliotheken (z. B. [102, 103]) registriert sind.[104] Um die Flexibilität des Proteinrückgrats abzubilden, kann ein Ensemble von Proteinkonformationen zusammengestellt werden. Momentaufnahmen des Proteins aus molekulardynamischen (MD) Simulationen (vgl. Abschnitt 3.3.4) oder NMR-Experimenten können dieses Ensemble bilden und einen statistischen Durchschnitt eingenommener Proteinkonformationen repräsentieren.[105, 106] Manche vollflexible Docking-Ansätze können solch ein Ensemble verarbeiten, um während der Platzierung verschiedene Proteinkonformationen zu evaluieren.[107, 108, 109, 110, 111, 112]

Vermeidung kombinatorischer Explosionen: Auch wenn jede Variable nur eine endliche Anzahl an Zuständen einnehmen kann, wird der Suchraum für jede weitere Dimension exponentiell erweitert. Ein simultanes Traversieren des Transformations-, Ligand- und Proteinkonformationsraums kann somit zu einer kombinatorischen Explosion führen. Eine Strategie zur Vermeidung kombinatorischer Explosionen ist das aufeinanderfolgende Traversieren individueller Subräume um einen Großteil des kombinierten Suchraums auszuschließen. Voll-flexible Docking-Methoden, die ein Induced-Fit (vgl. Abschnitt 2.7.3) bei Protein-Ligand-Bindung simulieren, wenden diese Strategie

u. a. an.[113, 114, 115] Dabei werden zunächst Ligandfreiheitsgrade variiert während das Protein starr gehalten wird. Die Freiheitsgrade vielversprechender Posen werden daraufhin fixiert und schließlich die Variablen des Proteins optimiert.

Vereinte Freiheitsgrade: Zuletzt sei noch die Möglichkeit der Vereinigung von Freiheitsgraden genannt. Beispielsweise können durch eine Normalmodenanalyse relevante Moden extrahiert werden, die – als Freiheitsgrade im Docking integriert – die wesentlichen Bewegungen eines Proteins beschreiben.[116, 117]

3.3.4 Suchstrategien

Docking-Methoden lassen sich entsprechend der Art und Weise wie der Suchraum traversiert wird klassifizieren. Viele Methoden sind *systematische Suchen*. Zu diesen gehören *Multikonformermethoden*, die Ligandkonformationen initial vor der eigentlichen Platzierung erzeugen, mit welchen sie dann ein starres Docking durchführen. *Fragmentbasierte Methoden* zerlegen Liganden in Fragmente, um diese dann in die Proteinbindetasche einzupassen. *Stochastische Methoden* variieren Freiheitsgrade zufällig, um den Suchraum effizient abzutasten. *Simulationsmethoden* sind deterministisch und ahmen den dynamischen Vorgang der Protein-Ligand-Bindung nach. *Hierarchische Methoden* kombinieren mehrere Suchstrategien.

Multikonformermethoden: Multikonformermethoden sind eine Kombination zweier Ansätze. Sie verbinden einen Konformergenerator mit einer starren Docking-Methode. Der Vorteil ist, dass Konformere vorberechnet und in einer Datenbank gespeichert werden können. Während des eigentlichen Docking-Prozesses können sie dann direkt genutzt werden, ohne dass der Konformationsraum während der Platzierung erneut traversiert werden muss. Zur Platzierung bestimmen Multikonformermethoden wie Dock 3.0[118], Flog[119], Fred[120] oder TrixX-BMI[2] Transformationen, die einzelne Konformationen in die Proteinbindetasche einpassen. Dafür werden Moleküle und Rezeptor durch charakteristische Merkmale repräsentiert, z. B. durch Graphen mit knotenzentrierten Sphären, atomzentrierte Gauß-Funktionen, Pharmakophormerkmale, Oberflächenpunkte und andere Formbeschreibungen. Abhängig von der Repräsentation wird dann ein Algorithmus angewandt, um die Formen zu vergleichen und eine Transformation zu bestimmen, z. B. mit einer systematische Suche vordefinierter Orientierungen, einem Graph-Matching, einem geometrisches Hashing oder einer Cliquen-Suche. Multikonformermethoden werden aufgrund ihrer Effizienz gerne im VS eingesetzt. Um große Bibliotheken zu verarbeiten, ist es jedoch notwendig, dass eine möglichst geringe Anzahl repräsentativer Konformationen erzeugt und gespeichert wird. Da die Platzierung primär auf einer Auswertung der komplementären Form von Protein und Ligand basiert,

ist der Docking-Erfolg allerdings davon abhängig, ob die bioaktive Konformation im initial erzeugten Konformerensemble enthalten ist. Ist sie es nicht, führt dies mitunter zu einem Docking-Fehlschlag.

Fragmentbasierte Methoden: Fragmentbasierte Methoden schränken den Suchraum mit Hilfe der gegebenen Proteinstruktur ein. Die Idee ist es, nur solche Konformationen zu erzeugen, die auch wirklich eine Chance haben in die Proteinbindetasche zu passen. Dafür werden Moleküle initial fragmentiert, d. h. an rotierbaren Bindungen geschnitten. Die erhaltenen Fragmente werden dann starr in die Bindetasche eingepasst. Entsprechend der Vorgehensweise wie vollständige Moleküle rekonstruiert werden, unterscheidet man zwischen Methoden, die einen *inkrementellen Aufbau* durchführen und Methoden, die eine *Placement-and-Linking*-Strategie verfolgen. Beim inkrementellen Aufbau werden zunächst sogenannte Basisfragmente oder Anker in die Bindetasche gelegt. Ausgehend von diesen werden die verbleibenden Fragmente des Liganden sukzessiv angefügt, bis der Ligand vollständig rekonstruiert wurde. Beim Anfügen werden verschiedene Torsionen, u. a. aus Torsionswinkelbibliotheken, exploriert. Wird während des Aufbaus festgestellt, dass die Teillösung keine valide oder nur eine ungünstige Pose ergeben kann, so wird der Aufbau an dieser Stelle abgebrochen. Docking-Methoden, die einen inkrementellen Aufbau durchführen sind z. B. FlexX[121], Hammerhead[122] oder Dock 4.0[123]. Placement-and-Linking-Algorithmen (z. B. Surflex[124], eHits[125]) versuchen die initial platzierten Fragmente innerhalb der Tasche so zu verbinden, dass eine vollständige Konformation erzeugt wird. Aufgrund der Diskretisierung der Freiheitsgrade bei der Platzierung müssen beim Verbinden der Fragmente jedoch Distanztoleranzen in Kauf genommen werden. Die erhaltenen Posen spiegeln daher nicht immer valide Konformationen wider. Sie können verzogen sein und ungewöhnliche Bindungslängen und -winkel besitzen. Kann ein Bestandteil überhaupt nicht platziert werden, so ist es nicht mehr möglich einen vollständigen Liganden zu rekonstruieren.

Stochastische Methoden: Unter der Annahme, dass ΔG durch eine Bewertungsfunktion $f(\vec{x})$ ausreichend gut angenähert werden kann, beschreibt die Funktion die Energiehyperfläche eines d-dimensionalen Suchraums. Das globale Minimum dieser Fläche repräsentiert die Energie der optimal in die Proteinbindetasche positionierten, bioaktiven Konformation. Das Docking-Problem kann somit als ein kontinuierliches Optimierungsproblem betrachtet werden, wobei die einzelnen Komponenten des Vektors $\vec{x}$ jeweils einen Freiheitsgrad darstellen. Unter Verwendung eines Optimierungsalgorithmus, der das globale Minimum der Ziel- oder Fitnessfunktion $f(\vec{x})$ identifiziert, kann die zugehörige Pose bestimmt werden:

$$\min_{\vec{x}\in\mathbb{R}^d} f(\vec{x}) : \mathbb{R}^d \to \mathbb{R} \tag{3.2}$$

Abhängig davon welche Freiheitsgrade in die Zielfunktion integriert werden, kann ein Optimierungsalgorithmus ein starres, flexibles oder voll-flexibles Protein-Ligand-Docking realisieren. Da die kontinuierliche Energiehyperfläche jedoch nicht vollständig durchsucht werden kann und zudem sehr rau ist, versuchen stochastische Methoden das Minimum zu identifizieren indem sie ihre Parameter zufällig variieren. Energiebarrieren können so leichter überwunden werden und die Gefahr, nur ein lokales Minimum aufzufinden, kann reduziert werden. Algorithmen, die dazu von Docking-Methoden genutzt werden sind Monte-Carlo-Methoden (ICM[126] oder QXP[127]), Genetische Algorithmen (AutoDock[128, 129] oder Gold[130]), eine Kombination aus Simulated Annealing, Evolutionäre Programmierung und Tabu-Suche (PRO_LEADS[131]), Ameisenalgorithmen (PLANTS[132, 133, 134]) und Partikelschwarmoptimierungen (SODOCK[135], ParaDockS[136] oder FIPSDock[137]).

Simulationsmethoden: Molekulardynamische Simulationen ermöglichen, die dynamische Entwicklung biologischer Systeme innerhalb eines gewissen Zeitraums zu untersuchen. Weiß man welche Kraft zu einem gewissen Zeitpunkt auf ein Atom ausgeübt wird, so ist es möglich, die Beschleunigung eines Atoms und dessen zukünftige Position innerhalb des Systems zu bestimmen. Molekülmechanische (MM) Methoden basieren auf dem zweiten Newtonschen Gesetz oder der sogenannten Bewegungsgleichung:

$$F_i = m_i a_i = -\nabla_i E \tag{3.3}$$

F_i ist die Kraft ist, die auf ein Atom i ausgeübt wird. Sie hängt von der Masse m_i und der Beschleunigung a_i des Atoms ab. Die Kraft kann ebenso als Gradient der potentiellen Energie E ausgedrückt werden. MM-Methoden schätzen E durch ein differenzierbares Kraftfeld ab (vgl. Abschnitt 3.3.5). Newtons Bewegungsgleichung kann so die Ableitung der potentiellen Energie mit der zeitabhängigen Positionsänderung in Bezug setzen:

$$-\frac{dE}{dr_i} = -m_i \frac{d^2 r_i}{dt^2} \tag{3.4}$$

Durch Integration erhält man eine Trajektorie, die die Positionen, Geschwindigkeiten und Beschleunigungen der Atome im Verlauf der Zeit beschreibt. Sind Positionen und Geschwindigkeiten aller Atome bekannt, so kann der Zustand des System zu jedem Zeitpunkt vorhergesagt werden. MD-Simulationen sind zeit- und rechenintensiv und die numerische Integration lässt sich nur approximativ lösen. Für eine Anwendung im VS sind sie nicht geeignet. Dennoch können sie vereinzelt, eingeschränkt den Docking-Prozess ergänzen, um Posen zu verfeinern und erneut zu bewerten.[138]

Hierarchische Methoden: Hierarchische Docking-Methoden kombinieren iterativ verschiedene Suchstrategien, um eine möglichst schnelle Posenselektion zu forcieren. Initial

wird der Suchraum nur grob abgetastet, um rasch vielversprechende Lösungskandidaten zu selektieren. Eine genauere, lokale Suche des Raums wird dann nur für die besten Kandidaten durchgeführt. Klassische Vertreter dieser Methoden sind HierVLS[139] oder Glide[140, 141]. Sie kombinieren systematische mit stochastischen Suchen. Die jeweiligen Docking-Phasen werden dabei von einem hierarchischen Bewertungsmodell begleitet (vgl. Abschnitt 3.3.5).

3.3.5 Bewertungsfunktionen

Bewertungsfunktionen leiten die Suchen und entscheiden, ob (Teil-)lösungen verworfen, weiterverfolgt oder zuletzt als Resultat des Docking-Prozesses präsentiert werden sollen.

Empirische Bewertungsfunktionen: Empirische Bewertungsfunktionen beruhen auf der Annahme, das ΔG durch Summation der wichtigsten energetischen Beiträge intermolekularer Interaktionen angenähert werden kann. Ein klassisches Beispiel ist die Bewertungsfunktion von FlexX[121], die auf der Böhm-Funktion[142] basiert:

$$\begin{aligned}\Delta G = \Delta G_0 & \\ &+ \Delta G_{\text{rot}} N_{\text{rot}} \\ &+ \Delta G_{\text{hb}} \sum_{\text{neutral H-bonds}} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{\text{io}} \sum_{\text{ionic int.}} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{\text{aro}} \sum_{\text{aro. int.}} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{\text{lipo}} \sum_{\text{lipo. cont.}} f^*(\Delta R)\end{aligned} \tag{3.5}$$

ΔG wird durch Beiträge für Wasserstoffbrücken ΔG_{hb}, ionische Wechselwirkungen ΔG_{io}, aromatische Wechselwirkungen ΔG_{aro} und lipophile Kontakte ΔG_{lipo} zwischen Protein und Ligand angenähert. Sie sind durch entropische Beiträge ergänzt, um den Verlust an Freiheitsgraden rotierbarer Ligandbindungen ΔG_{rot} und der restlichen Freiheitsgrade ΔG_0 bei Komplexbildung zu beschreiben. Die stückweise linearen Bestrafungsfunktionen $f(\Delta R, \Delta\alpha)$ und $f^*(\Delta R)$ skalieren die Einzelbeiträge bezüglich der Einhaltung der optimalen Interaktionsgeometrie und der Kontaktfläche. Bei leichten Abweichungen von der Idealgeometrie erhält eine Interaktion ihren vollen Beitrag, der bei größeren Abweichungen stetig reduziert wird. Die Interaktionsbeiträge sind darauf angepasst, experimentell bestimmte ΔG-Werte aufgeklärter Protein-Ligand-Komplexe zu beschreiben. Da die zur Kalibrierung verwendeten Kristallstrukturen allerdings einen

niederenergetischen Zustand des gebundenen Liganden widerspiegeln, bei dem primär stabilisierende Kräfte wirken, sind destabilisierende Beiträge in empirischen Bewertungsfunktionen unterrepräsentiert. Weitere Beispiele für empirische Bewertungsfunktionen sind PLP[143, 144, 145, 146], ChemScore[147, 148] oder die Funktionen von PLANTS[134].

Wissensbasierte Bewertungsfunktionen: Wissensbasierte Bewertungsfunktionen beruhen auf der Annahme, dass in Kristallstrukturen häufig beobachtete Kontakte zwischen Paaren von Protein- und Ligandatomen energetisch günstig sind und einen entscheidenden Beitrag zur Bindungsenergie liefern müssen. Von einer Menge von Kristallstrukturen können so, basierend auf einer Boltzmann-Statistik der Häufigkeitsverteilungen beobachteter Atompaardistanzen, distanzabhängige Atompaarpotentiale $A_{ij}(r)$ bestimmt werden. Für einen unbekannten Komplex liefert die Summation dieser Potentiale dann eine Einschätzung der Bindungsenergie. Ein Beispiel für eine wissensbasierte Bewertungsfunktion ist der DrugScore[149, 150] oder das von Muegge entwickelte PMF (potential of mean force)[151, 152, 153]:

$$\mathrm{PMF} = \sum_{\substack{kl \\ r<r^{ij}_{\text{cut-off}}}} A_{ij}(r), \quad A_{ij}(r) = -k_B T \ln \left[f^{j}_{\mathrm{Vol\_corr}}(r) \frac{\rho^{ij}_{\text{seg}}(r)}{\rho^{ij}_{\text{bulk}}} \right] \tag{3.6}$$

k_B bezeichnet die Boltzmann-Konstante, T die absolute Temperatur und r die Distanz eines Atompaares. $f^{j}_{\mathrm{Vol\_corr}}(r)$ ist ein Korrekturfaktor zur Berücksichtigung des vom Liganden beanspruchten Volumens. $\rho^{ij}_{\text{seg}}(r)$ ist die Teilchendichte des Paares vom Typ ij in der zugrundeliegenden Strukturdatenbank, die in einem gewissen radialen Bereich *seg* beobachtet wird. ρ^{ij}_{bulk} stellt die Teilchendichte von i und j in einem Referenzzustand dar, d. h. wenn i und j nicht interagieren. Der Quotient $\frac{\rho^{ij}_{\text{seg}}(r)}{\rho^{ij}_{\text{bulk}}}$ bezeichnet die radiale Verteilung eines Proteinatoms vom Typ i mit einem Ligandatom vom Typ j in einer Strukturdatenbank von Protein-Ligand-Komplexen. Verschiedene wissensbasierte Bewertungsfunktionen unterscheiden sich primär in den Paarungen ij und den verwendeten Strukturen auf deren Basis die Potentiale bestimmt wurden. Die Bestimmung von Paarpotentialen erfordert keine Affinitätsdaten, daher können wissensbasierte Bewertungsfunktionen im Vergleich zu empirischen in größerem Umfang kalibriert werden.

Kraftfeldbasierte Bewertungsfunktionen: Kraftfelder beschreiben die Zusammenstellung von Wechselwirkungstermen:

$$E = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{3.7}$$

Bindungsvermittelte Wechselwirkungsterme beschreiben intramolekulare Energien, die zwischen kovalent gebundenen Atomen wirken, z. B. Beiträge, die bei Abweichungen von

der optimalen Bindungslänge r_0, Bindungswinkel ϕ_0^{ijk} oder Torsionswinkel τ_0 beobachtet werden. Nicht-bindungsvermittelte Wechselwirkungsterme beschreiben die energetischen Beiträge, die zwischen nicht kovalent gebundenen Atomen wirken, z. B. Wasserstoffbrücken, ionische und van-der-Waals-Wechselwirkungen. Die Terme sind auf Prinzipien der Molekülmechanik formuliert, die das molekulare System basierend auf der klassischen Mechanik modellieren. Bindungen und Bindungswinkel werden typischerweise als Oszillatoren durch ein harmonisches Potential beschrieben:

$$E_{\text{stretch}} = c_{ij}(r_{ij} - r_0)^2, \quad E_{\text{bend}} = c_{ijk}(\phi - \phi_0^{ijk})^2 \tag{3.8}$$

Da Torsionswinkelpotentiale in der Regel mehrere Minima besitzen, lassen sie sich nicht als Oszillatoren modellieren. Eine Kosinus-Funktion beschreibt in diesem Fall die Periodizität des Potentials bei Rotation um die Bindung:

$$E_{\text{tors}} = c(1 + \cos(n\tau - \tau_0)) \tag{3.9}$$

Elektrostatik lässt sich mit Hilfe eines Monopol-Ansatzes über das Coulombsche Gesetz formulieren, wobei Atome als Punktladungen q_i, q_j im Abstand r_{ij} betrachtet werden:

$$E_{\text{es}} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{3.10}$$

Van-der-Waals-Wechselwirkungen werden mit dem Lennard-Jones-Potential beschrieben. Bei Atomen mit einem Abstand r_{ij} bestraft es ungünstige Überlappungen und forciert die optimale Lage der Atome, sodass Dispersionkräfte attraktiv wirken können:

$$E_{\text{vdw}} = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{3.11}$$

Wasserstoffbrücken werden teils auch mit Varianten des Lennard-Jones-Potentials modelliert. Die Konstanten c, c_{ij}, c_{ijk}, ϵ_0, ϵ und σ entsprechen kraftfeldspezifischen Parametern, die in Abhängigkeit betrachteter Atomtypen die Terme kalibrieren. Kraftfelder wie AMBER[154] und CHARMM[155] werden klassischerweise in molekulardynamischen Simulationen eingesetzt, um die Kräfte zu modellieren, die zu einem gewissen Zeitpunkt auf Atome des Systems wirken. Ihre Terme finden sich jedoch auch in kraftfeldbasierten Bewertungsfunktionen wieder (z. B. in Gold[156, 130], Glide[140, 141] oder AutoDock[157]). Dabei werden sie auch mit empirischen Termen kombiniert, um Solvatisierungs- und entropische Effekte zu berücksichtigen.

Hierarchische Bewertungsmodelle: Hierarchische Docking-Methoden, wie z. B. HierVLS[139] oder Glide[140, 141], sind durch sukzessiv durchgeführte Bewertungs- und

Filterphasen gekennzeichnet. In jeder Phase wird eine andere Bewertungsfunktion eingesetzt und die Anzahl der Posen kontinuierlich reduziert. Bei der initialen, groben Abtastung des Suchraums sind die Methoden mit mehreren Tausend bis Zehntausend Platzierungen konfrontiert. Daher müssen früh eingesetzte Bewertungsfunktionen rasch evaluiert werden. Eine Strategie ist es, initial ausschließlich die Komplementarität der Posen zu bewerten und eine genaue Bewertung von Interaktionsgeometrien zu vernachlässigen. Einmalig vorberechnete und an Gitterpunkten annotierte Bewertungen können dazu genutzt werden, eine frühe Einschätzung zu erhalten, indem die vorberechneten Werte abgefragt werden. Frühe Bewertungsfunktionen sind zumeist Varianten der finalen Bewertungsfunktion. Sie berücksichtigen zuvor gemachte Abschätzungen (z. B. Grad der Diskretisierung eines Freiheitsgrades) und besitzen andere Parametrisierungen, um Posen „weicher“ zu bewerten. Spät eingesetzte Bewertungsfunktionen sind kostenintensiver, liefern aber eine genauere Abschätzung der Bindungsenergie. Sie werden nur für wenige Posen durchgeführt. Die frühen Phasen sind von einfach und schnell durchzuführenden Tests durchzogen. Sie sollen prospektiv ungünstige Posen verwerfen und nur solche selektieren, die eine Chance besitzen, eine gute Bewertung in den späteren Phasen zu erzielen. Die Strategie möglichst früh schlechte Posen auszusortieren, bietet die Möglichkeit, den Docking-Prozess massiv zu beschleunigen.[158]

Konsensusbewertung: Eine Konsensusbewertung[159] oder *Datenfusion* ist eine Bewertungsstrategie, die vor allem in VS-Kampagnen genutzt wird, um die Ergebnisse verschiedener Bewertungsfunktionen zu vereinen. Oft ist ihr Ziel die Schwächen individueller Bewertungsfunktionen durch Stärken anderer auszugleichen, um eine verbesserte Anreicherung zu erhalten. Eine Konsensusbewertung bietet auch die Möglichkeit Screening-Resultate bei Nutzung unterschiedlicher Deskriptoren oder unterschiedlicher Referenzdaten zu kombinieren.[160, 161, 162, 163] Immer häufiger werden mit diesem Mittel auch Ergebnisse ligand-, struktur- und pharmakophorbasierter VS-Läufe vereint.[164, 165, 166] Fusionierungsmethoden verarbeiten die Bewertungen oder Ränge der ursprünglichen Hitlisten weiter: Die *Summe der Ränge* addiert die Ränge der ursprünglichen Hitlisten. Entsprechend der Summe werden die Verbindungen erneut sortiert. Die *Summe der Scores* berechnet in jeder Hitliste, für jede Verbindung zunächst einen relativen Score, indem die ursprüngliche Bewertung durch die Beste der Hitliste geteilt wird. Für eine Verbindung werden dann die relativen Scores über die Hitlisten summiert. Bei der *parallelen Selektion* werden Verbindungen der Reihe nach von den oberen Rängen der Hitlisten selektiert bis die gewünschte Anzahl an Verbindungen erreicht ist. Wurde eine Verbindung bereits gewählt, so wird ihr direkter Nachfolger aus der entsprechenden Hitliste gewählt. Die *Bewertung durch Votum* stimmt für ein Molekül, sobald es innerhalb der oberen $x\,\%$ einer Hitliste vorgefunden wird. Abhängig

von der Anzahl n der Hitlisten kann ein Molekül somit 0 bis maximal n Stimmen erhalten, anhand derer eine vereinte Hitliste zusammengestellt wird. Bei Gleichstand kann als zweites Kriterium die Summe der Scores herangezogen werden. Bei einer *Pareto-Bewertung* wird für jede Verbindung gezählt, wie häufig andere Verbindungen in allen Hitlisten besser bewertet werden. Bei einer identischen Bewertung wird die Summe der Ränge als zweites Bewertungskriterium herangezogen. Der *Z-Score* standardisiert alle ursprünglichen Bewertungen S_{hi} einer Verbindung i in Hitliste h:

$$Z_{hi} = \frac{S_{hi} - \mu_h}{\sigma_h} \tag{3.12}$$

μ_h ist die durchschnittliche Bewertung und σ_h die Standardabweichung.

3.4 Pharmakophorbasiertes virtuelles Screening

Ein PBVS macht sich das in Abschnitt 2.6 vorgestellte Pharmakophorkonzept zunutze. Demnach etablieren Liganden einer Zielstruktur ein identisches Interaktionsmuster, das durch die größte gemeinsame räumliche Anordnung spezifischer Pharmakophormerkmale reflektiert wird. Ist ein Pharmakophormodell gegeben, das diese Merkmale umfasst, lassen sich mit ihm im PBVS passende Liganden aus einer Molekülbibliothek extrahieren. Es realisiert den Abgleich des Pharmakophormodells mit den Merkmalen der Moleküle. Das Resultat ist eine Liste nicht priorisierter Verbindungen, also eine Submenge der Bibliothek. Ein PBVS ist dabei stets Teil eines mehrstufigen Prozesses:

- Erstellung der Pharmakophorhypothese/n (*Pharmakophormodellierung*)
- Validierung der Hypothese (*Pharmakophorvalidierung*)
- Anwendung des bestätigten Pharmakophormodells im PBVS

Unter dem Begriff *Pharmakophormethoden* sind Methoden dieser Phasen zusammengefasst. Kommerziell verfügbare Software wie das Discovery Studio von Accelrys[167], die Small-Molecule Drug Discovery Suite von Schrödinger[168], die Molecular Operating Environment (MOE) der Chemical Computing Group[169] oder LigandScout von Inte:Ligand[170] bieten Module zur Erstellung, Validierung und Anwendung von Pharmkophorhypothesen an. Zudem sind Methoden Gegenstand akademischer Entwicklungen. Für einen Überblick sei auf [171, 172, 173, 174, 175, 176, 177, 48, 178] verwiesen.

3.4.1 Repräsentation von Pharmakophorhypothesen

Pharmakophorhypothesen werden durch gängige Methoden sehr ähnlich repräsentiert. Sie bestehen aus einer Zusammenstellung weniger Pharmakophormerkmale (3 bis 7, selten mehr) und enthalten Information über deren Anordnung im Raum (*3D-Pharmakophormodell*). Merkmale können entweder *atom-* oder *feldbasiert* dargestellt werden. Die atombasierte Repräsentation ist jedoch weiter verbreitet. *Pharmakophorfingerabdrücke* sind ein Mittel zur binären Kodierung und Klassifizierung von Merkmalsmustern, die in Molekülen beobachtet werden können. Sie dienen hauptsächlich zur internen Repräsentation pharmakophorer Moleküleigenschaften und zum effizienten Merkmalsabgleich.

Atombasierte Repräsentation: Die atombasierte Repräsentation ordnet ein Phamakophormerkmal eindeutig einem Atom oder einer funktionellen Gruppe zu. Die Merkmale können unterschiedlich spezifiziert sein und entsprechend des Grads der Abstraktion von der Molekülstruktur klassifiziert werden. Ein Abstraktionslevel niedriger Ordnung entspricht dabei Merkmalen mit hoher Spezifität aber geringer Universalität[179, 180]:

- *Level-4-Merkmale* sind durch Punkt(e) im Raum spezifiziert und entsprechend der Interaktionsmöglichkeit mit der Zielstruktur typisiert, z. B. sind sie ein Wasserstoffbrückendonor, -akzeptor, hydrophobes, positiv bzw. negativ ionisierbares Merkmal. Zentrierte Toleranzsphären beschreiben die variable Position des Merkmals.
- *Level-3-Merkmale* besitzen zudem eine assoziierte Richtung oder eine weitere Sphäre, die an die Stelle des potentiellen Interaktionspartners projiziert wird, z. B. ein Wasserstoffbrückendonormerkmal mit einer projizierten Akzeptorsphäre.
- *Level-2-Merkmale* sind mit einem Teil des Molekülgraphs assoziiert und wie Level-4-Merkmale geometrisch unbeschränkt, z. B. ein Phenolfragment.
- *Level-1-Merkmale* erweitern Level-2-Merkmale um eine geometrische Beschränkung, z. B. ein Phenol, das in 2 bis 4 Å parallel auf ein benzenoides System trifft.

Merkmalsschemata verschiedener Pharmakophormethoden unterscheiden sich auf den ersten Blick nur minimal. Die Unterschiede sind jedoch ein Grund weshalb unter derselben Voraussetzung verschiedene Hypothesen und Screening-Ergebnisse produziert werden.[181, 182] Schemata unterscheiden sich in Art und Umfang der Typisierung, durch welchen Abstraktionslevel ein Merkmalstyp beschrieben wird, welche Atome Punkte und wie viele abgeleitet werden, wo sie platziert und wie groß ihre Toleranzsphären skaliert werden. Abbildung 3.3 vergleicht schematisch die Pharmakophormerkmale von Phase[183] (Schrödinger), Catalyst (Accelrys) und MOE (nach Spitzer[181] und Wolber[180]). Typischerweise formulieren Hypothesen zumin-

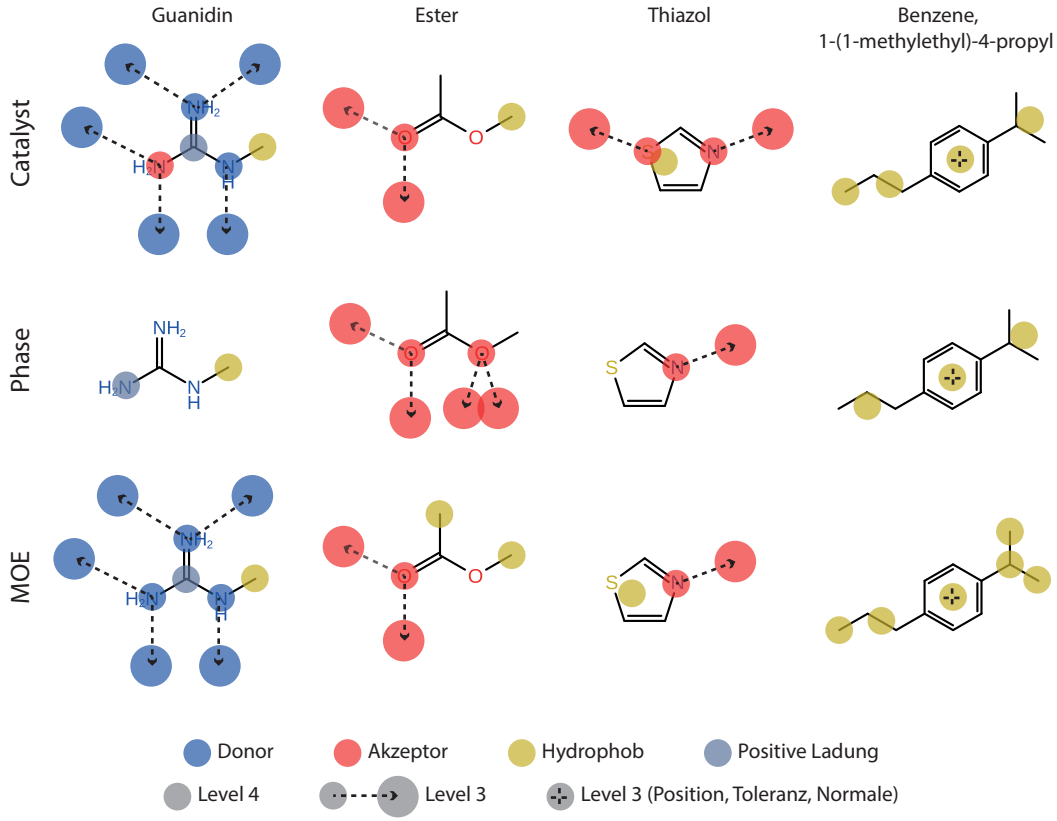


Abbildung 3.3: Catalyst und MOE modellieren Wasserstoffbrückenmerkmale als projizierte Punkte, Phase standardmäßig mit assoziierten Richtungen (optional mit projizierten Punkten). Im Gegensatz zu Catalyst und MOE, deren Merkmale auf Schweratomkoordinaten positioniert sind, zentriert Phase Donormerkmale (nicht dargestellt) auf Wasserstoffe (nach Spitzer[181] und Wolber[180]).

dest Wasserstoffbrückendonor-, Wasserstoffbrückenakzeptor- und hydrophobe Merkmale. Donormerkmale werden mit wasserstoffgebundenen Sauerstoffen und Stickstoffen, Akzeptormerkmale mit Sauerstoffen und Stickstoffen mit freien Elektronenpaaren assoziiert. Nicht klassische Donoren, z. B. SH– oder CH–Gruppen, bzw. schwache Akzeptoren wie Sauerstoffe von Ethergruppen werden aufgrund ihrer uneindeutigen Interaktionen nicht einheitlich gehandhabt.[174] Wasserstoffbrückenmerkmale sind zumeist als Level-3-Merkmale modelliert, um die Direktionalität der Brücken abzubilden. Eine Darstellung mittels projizierter Sphären ermöglicht die Vereinigung der Sphären bei Überlappungen, auch wenn sie von unterschiedlichen Atomen ausgehen. Um die Geometrie von Wasserstoffbrücken weniger stark zu gewichten, positionieren manche Methoden die Merkmale näher zu ihrem Ursprung auf Wasserstoffe bzw. freie Elektronenpaare

oder direkt auf Schweratomzentren. Hydrophobe Interaktionen sind geometrisch nicht klar definiert. Daher unterscheiden sich verschiedene Modellierungsmethoden, obwohl sie teilweise auf demselben Algorithmus[184] beruhen, bei der Bestimmung hydrophober Merkmale.[180] Meistens sind sie jedoch auf Ringzentren, entlang aliphatischer Ketten und auf Verzweigungen zu finden und als Level-4-Merkmale modelliert. In aromatischen Ringen können sie mit der Normalen der Ringebene versehen und als *aromatische* Level-3-Merkmale beschrieben sein, um potentielle π-π-Interaktionen abzubilden. Fein granulierte Level-4-Merkmale werden mitunter dazu genutzt, die Molekülform zu beschreiben und werden dann als *sterische* oder *Van-der-Waals-Merkmale* bezeichnet.[185, 186] Zusätzlich zu den drei beschriebenen typischen Merkmalstypen werden auch positiv und negativ geladene, geometrisch unbeschränkte Level-4-Merkmale mit ionisierbaren Gruppen assoziiert. Allerdings ist umstritten, ob dies tatsächlich notwendig ist, da geladene Typen bei einem Merkmalsabgleich implizit fordern, dass entsprechende Gruppen nur rein ionische Wechselwirkungen bilden können.[174] Manche Programme erlauben benutzerspezifische Merkmale zu definieren. So können funktionelle Gruppen als geometrisch unbeschränkte Level-2- oder geometrisch beschränkte Level-1-Merkmale definiert werden. MOE bietet zudem die Möglichkeit aus mehreren Merkmalsschemata zu wählen und überlässt so dem Anwender, welche Merkmalstypen in welchem Abstraktionslevel dargestellt werden sollen.

Feldbasierte Repräsentation: Bei der feldbasierten Repräsentation werden Pharmakophormerkmale nicht anhand der Molekülstruktur, sondern anhand eines molekularen Feldes abgeleitet.[187] Ein molekulares Feld beschreibt das elektrostatische Potential (MEP) in der Umgebung eines Moleküls. Das MEP wird berechnet, indem an Punkten eines dreidimensionales Gitters oder einer diskretisierten Oberfläche eine Sonde (z. B. ein Proton) platziert wird. Mit Hilfe des Coulombschen Gesetz wird dann das elektrostatische Potential zwischen der Sonde und den Partialladungen der Molekülatome bewertet und an den Punkten annotiert. Die molekularen Interaktionsfelder (MIFs) von Grid[188], welche eigentlich zur Identifizierung von potentiellen Interaktionsstellen in Proteinen gedacht sind, können auch dazu genutzt werden, kleine Moleküle mit funktionellen Gruppen zu sondieren und deren Potential zur Interaktion zu bewerten. Das XED-Kraftfeld versucht Orbitale der Atome zu modellieren, um die Ladungsverteilung um das Molekül zu beschreiben.[189] Prinzipiell leiten feldbasierte Methoden sehr viele potentielle Pharmakophorpunkte ab, die sich nicht ohne Weiteres effizient in Pharmakophormethoden nutzen lassen. Es existieren verschiedene Strategien, um Extrema in den Feldern zu identifizieren oder die Felder anders zu beschreiben. Grind[190] identifiziert Extrema in MIFs, die mit Grid erstellt wurden, um die wahrscheinlichsten Regionen für einen Akzeptor, Donor oder anderer Merkmale zu beschreiben. Das MEP kann aber

auch z. B. durch atomzentrierte Gauß-Funktionen annähernd gut beschrieben werden, die sich dann effizient vergleichen lassen.[191, 192]

Pharmakophorfingerabdruck: Ein *Pharmakophorfingerabdruck* (oder *-schlüssel*) repräsentiert die Zusammenstellung pharmakophorer Merkmale eines Moleküls, indem er binär kodiert, welche *Merkmalskonfiguration(en)* im Molekül vorzufinden sind:

> ***k*-Konfiguration:** Eine k-Konfiguration ist eine Teilmenge von k Pharmakophormerkmalen.[193] Der Typ der Konfiguration ist durch die Merkmalstypen gegeben. Eine 3-Konfiguration, die aus einem Donor (D), einem Akzeptor (A) und einem hydrophoben Merkmal (H) besteht, impliziert den Konfigurationstyp DAH.
>
> **Partition:** Eine Partition umfasst alle Konfigurationen eines bestimmten Konfigurationstyps. Die DDH-Partition beschreibt beispielsweise alle 3-Konfigurationen, die aus zwei Donoren und einem hydrophoben Merkmal bestehen.
>
> **Gemeinsame Konfiguration:** Besitzen zwei Moleküle Konfigurationen, die derselben Partition angehören und lassen sich ihre Merkmale innerhalb einer gewissen Toleranz zur Deckung bringen, so besitzen sie eine gemeinsame Konfiguration.

Pharmakophormethoden (und teilweise LBVS- und SBVS-Methoden) nutzen Pharmakophorfingerabdrücke, um effizient Moleküle zu vergleichen. Dafür identifizieren sie gemeinsame Konfigurationen. Für den Merkmalsabgleich kodieren Pharmakophorschlüssel zusätzlich die relative Anordnung der Merkmale. Kontinuierliche Distanzwerte zwischen Merkmalspaaren werden diskreten Intervallen (Bins) zugeordnet und durch ein Bit repräsentiert. Je feiner die Intervalle gewählt werden, desto genauer sind die Distanzen kodiert. Pharmakophorschlüssel der Pharmakophormethoden unterscheiden sich darin, welche Arten von Konfigurationen beschrieben werden. Oft werden 3-Konfigurationen verwendet, um die relative räumliche Anordnung von Merkmalen zu registrieren. Manche Methoden nutzen 4-Konfigurationen, da sie zudem auch zwischen den Merkmalen chiraler Moleküle unterscheiden können.[194, 195] Es werden aber auch größere oder unterschiedliche Konfigurationen genutzt. Die Methoden sind dann jedoch speicherintensiver. Gröbere Distanzintervalle ermöglichen eine ressourcenschonende Verarbeitung, allerdings sind die Berechnungen ungenauer.[174] Deskriptoren, die Kombinationen an Konfigurationen und zusätzliche Information über relative Häufigkeiten der k-Tupel kodieren, können darüber hinaus eingesetzt werden, um Pharmakophorprofile von molekularen Datenbanken zu erstellen.

3.4.2 Pharmakophormodellierung

Eine Pharmakophorhypothese kann prinzipiell ligand- oder strukturbasiert erstellt werden: Die *ligandbasierte Pharmakophormodellierung* verfolgt das Ziel, eine Menge von

Liganden so zu überlagern, dass offensichtlich wird, welche Merkmale ähnlich angeordnet und deshalb mutmaßlich an der Bindung zur unbekannten Struktur des Zielproteins beteiligt sind. Ligandbasierte Ansätze sind die gängigsten Methoden, da sie lediglich die Kenntnis über bereits bekannte, bioaktive Moleküle benötigen. Bei der *strukturbasierten Pharmakophormodellierung* wird die dreidimensionale Zielstruktur herangezogen. Man kann hierbei zwischen *apoprotein-* und *komplexbasierten* Methoden unterscheiden. Erstere nutzen zur Erstellung von Pharmakophorhypothesen ausschließlich Informationen des Proteins und versuchen potentielle Interaktionsstellen innerhalb der Proteinbindetasche zu identifizieren, um dadurch auf essentielle Pharmakophormerkmale der Liganden schließen zu können. *Komplexbasierte* Methoden nutzen die dreidimensionale Struktur vollständiger Protein-Ligand-Komplexe. Sie versuchen typischerweise etablierte Interaktionen zu identifizieren und daraus die gemeinsamen Merkmale zu extrahieren.

Ligandbasierte Pharmakophormodellierung: Grundsätzlich müssen ligandbasierte Methoden das Problem der molekularen Überlagerung lösen (vgl. Abschnitt 3.2.1). Sie durchsuchen den Konformationsraum der teils sehr verschiedenen Moleküle und versuchen zugleich, gemeinsame Merkmale räumlich zur Deckung zu bringen. Dabei identifizieren sie direkt oder indirekt eine größte gemeinsame Konfiguration in Molekülen (analog zum MCS-Problem). Um Ausrichtungen zu erzeugen, werden die Abstandsquadrate der Konfigurationsmerkmale minimiert. Existierende Ansätze unterscheiden sich primär im Zeitpunkt der Konformergenerierung. Konformere können vorberechnet in einer Datenbank gespeichert sein oder während des Merkmalsabgleichs erzeugt werden. Manche Ansätze nutzen auch eine Kombination und justieren vorberechnete Konformere während des Merkmalsabgleichs.[196] Vorberechnete Konformere haben den Vorteil, dass Überlagerungen schnell durchgeführt werden können. Einer der ersten Ansätze war DISCO[197, 198], der Merkmale mit Hilfe des Bron-Kerbosch-Algorithmus zur Cliquen-Suche abgleicht. LIGANDSCOUT[199] identifiziert ein maximales bipartites Matching zum Abgleich korrespondierender Merkmalspaare. PHASE[183] nutzt eine baumbasierte Partitionierungsmethode, um k-Konfigurationen mit ähnlichen Zwischenmerkmalsdistanzen zu gruppieren und seinen Suchraum zu beschränken. Integriert in Discovery Studio (CATALYST) identifiziert der HIPHOP-Algorithmus[193] mit vorberechneten Konformeren[200] die größte gemeinsame Konfiguration zu einem Referenzmolekül. Konfigurationen werden stetig vergrößert und überprüft, ob sie mit der Referenz zur Deckung gebracht werden können. Die zuletzt gefundene Konfiguration definiert die Pharmakophorhypothese. Algorithmen, die Konformere während des Merkmalsabgleichs erzeugen, sind weniger speicher-, aber dafür rechenintensiver. Klassische Beispiele sind

die genetischen Algorithmen Gasp[201] und sein Nachfolger Galahad[202]. Gasp optimiert eine initiale Population zufälliger Überlagerungen. Jedes Individuum ist durch ein Chromosom repräsentiert, dessen Gene Torsionswinkel rotierbarer Bindungen und eine Zuordnung jedes Merkmals auf ein Referenzmerkmal kodieren. In jedem Evolutionsschritt wird ein Nachkomme erzeugt, indem ein Elternchromosom mutiert oder ein Crossover zweier Individuen durchgeführt wird. Dadurch werden Torsionswinkel oder Merkmalszuordnungen variiert. Die Fitnessfunktion bewertet die Population. Sie berücksichtigt die Abstandsquadrate der ausgerichteten Merkmale, die Ausdehnung des Überlappvolumens und die Torsionsspannung der Liganden. Galahad variiert zunächst Torsionen, um Konformationen mit großem Überlappungsvolumen, niederer Torsionsspannung und ähnlichen Konfigurationen zu erzeugen. Konfigurationen sind über Fingerabdrücke kodiert, die Tripletts, Quadrupel, etc. und Distanzen zwischen Merkmalen enthalten. Im zweiten Schritt werden die Konformere starr gehalten und ein Algorithmus aus der Bilderkennung genutzt, um die finale Ausrichtung der Liganden zu erzeugen.

Apoproteinbasierte Pharmakophormodellierung: Steht die 3D-Struktur des Proteins zur Verfügung, identifizieren apoproteinbasierte Modellierungsmethoden an der Proteinoberfläche Merkmale, die an der Bindung von Liganden beteiligt sein könnten. Initial beschreiben sie ein negatives Abbild eines ligandbasierten Pharmakophors, das invertiert wird, um eine Hypothese zu erhalten. Dafür muss zunächst die Bindetasche definiert werden (vgl. Abschnitt 3.3.3). Da Bindetaschen größer als Liganden sind, erzeugen die Methoden mehr Merkmale. Die Herausforderung besteht vor allem darin, eine möglichst geringe Anzahl entscheidender Merkmale zu selektieren. Proteinflexibilität führt zwar zu variierenden Merkmalspositionen, allerdings vernachlässigen die meisten strukturbasierten Ansätze[203, 204] die Auswertung von Konformationen. Dafür werden Toleranzsphären größer skaliert, um die Modelle „weicher“ zu gestalten und zumindest indirekt etwas Flexibilität zu gewähren. Durch das Protein lassen sich räumliche Ausschlussmerkmale definieren, die von Liganden nicht eingenommene Räume beschreiben. Diese Level-4-Merkmale stellen implizit Forderungen an die Form potentieller Liganden und führen im VS zu selektiveren Hypothesen. Um Hypothesen anhand der Proteinstruktur abzuleiten, kann die Bindetasche mit funktionellen Gruppen sondiert, deren Interaktionspotential zu Proteinatomen bewertet und die besten Interaktionsstellen selektiert werden. Ein Vertreter dieses Ansatzes ist das in LigBuilder integrierte Pocket-Modul.[205] Es sondiert die Tasche mit positiv geladenen, sp^3-hybridisierten Stickstoffen, mit negativ geladenen, sp^2-hybridisierten Sauerstoffen und Kohlenstoffatomen. Ihre Interaktionsgeometrien werden durch eine empirische Funktion bewertet. Die besten Sonden werden in Merkmale vom Typ Donor, Akzeptor bzw. hydrophob

konvertiert. Auf ähnliche Weise lassen sich GRID-Interaktionsfelder[188] oder LUDI-Interaktionskarten[206, 207] mit diversen Filter- und Clustering-Strategien kombinieren, um Pharmakophormerkmale mit minimaler Energie zu selektieren.[208, 209, 210, 211] Eine Methode von Carlson et al. berücksichtigt Proteinflexibilität durch Momentaufnahmen des Proteins einer MD-Simulation.[212] Sie werden mit unterschiedlichen Sonden geflutet, die simultan minimiert werden. Nach jedem Optimierungsschritt werden Sonden ausgewählt, die Cluster bilden und fest an das Protein binden. Sie deuten auf markante Stellen in der Bindetasche hin. Mit einem ähnlichen MCSS (multiple copy simultaneous search)-Ansatz[213] können auch PHASE-Merkmale[183] generiert werden, indem vordefinierte Fragmente mit GLIDE[214] gedockt und bewertet werden.[215, 216] HS-PHARM verfolgt einen wissensbasierten Ansatz.[217] Anhand bekannter Zielstrukturen der sc-PDB[218, 219] wurden in einem maschinellen Lernverfahren charakteristische, an Bindungen beteiligten Atome ermittelt und in Fingerabdrücke kodiert. Bei gegebenen Apoprotein werden sie nach einem passenden Fingerabdruck durchsucht. Dessen kodierte Interaktionsatome dienen dann zur Erstellung der Pharmakophormerkmale.

Komplexbasierte Pharmakophormodellierung: Der Nachteil von Apoproteinansätzen ist, dass zumeist zu viele, energetisch nicht unterscheidbare Pharmakophormerkmale erhalten werden. Komplexbasierte Ansätze versuchen das Problem zu beheben, indem sie zur Selektion der Merkmale zumindest die Information eines Liganden heranziehen. Sie extrahieren die Merkmale, die komplementäre physikochemische Eigenschaften und ideale Interaktionsgeometrien aufweisen. Auf diese Weise schränkt auch POCKET V.2[220], eine Erweiterung des POCKET-Moduls, Pharmakophormerkmale anhand eines Protein-Ligand-Komplexes weiter ein. Allerdings reflektiert eine Pharmakophorhypothese, die anhand eines einzelnen Liganden ermittelt wurde, keine aussagekräftige Struktur-Aktivitäts-Beziehung. Aus diesem Grund existieren Methoden, die ligand- und strukturbasierte Modellierungmethoden vereinen und Pharmakophorhypothesen von einem Protein und mehreren Liganden oder mehreren Protein-Ligand-Komplexen ableiten: LIGANDSCOUT[179] leitet aus einem gegebenen Protein-Ligand-Komplex Merkmale von der Ligandstruktur ab. Die Merkmale werden bezüglich ihres Bindungspotentials mit Proteinmerkmalen bewertet und bei Einhaltung geometrischer Beschränkungen zum Pharmakophormodell hinzugefügt. Die so erhaltene Hypothese kann weiter verfeinert werden. Es können z. B. zusätzlich Merkmale für weitere bekannte Liganden erstellt und am initial erstellten Modell ausgerichtet werden oder Modelle mehrerer Komplexe vereint werden. Sind zusätzlich inaktive Moleküle bekannt, können über die Ausrichtung ihrer Merkmale auch Ausschlussmerkmale definiert werden. Anstatt Pharmakophormerkmale aus verschiedenen Modellen zu superpositionieren, ist es auch möglich, zuerst mehrere homologe Protein-Ligand-Komplexe anhand des Proteinrückgrats auszurichten.

Wird im Anschluss die Proteininformation verworfen, verbleiben die überlagerten nativen Bindungsmodi der unterschiedlichen Liganden. Ähnlich angeordnete Merkmale können dann über eine Clusteranalyse vereint werden, um ein gemeinsames Pharmakophormodell zu erhalten.[221] Strukturelle Interaktionsfingerabdrücke (SIFt)[222, 223, 224], die ein Interaktionsprofil eines Protein-Ligand-Komplexes binär kodieren, können dazu genutzt werden eine Pharmakophorhypothese zu erstellen. Eine Menge von SIFts kann dabei von mehreren Protein-Ligand-Komplexen oder auch von Docking-Posen verschiedener Liganden abgeleitet werden.[225] Merkmale der Liganden, die häufig auftretende Interaktionen der SIFts erklären, induzieren ein konserviertes Interaktionsmuster und somit ein gemeinsames Pharmakophormodell. FLAP (Fingerprints for Proteins and Ligands)[194, 226] ist ein Programm, das zum virtuellen Screening entwickelt wurde. Intern verarbeitet es in Fingerabdrücken kodierte Quadrupel von Pharmakophormerkmalen. Die Merkmale werden zuvor von Grid-Interaktionsfeldern abgeleitet und können sowohl für Moleküle als auch Proteine berechnet werden und zur Pharmakophormodellierung genutzt werden.[211]

3.4.3 Validierung von Pharmakophorhypothesen

Automatisch erstellte Modelle hängen entscheidend von den zur Verfügung gestellten Daten ab. Ist die Anzahl gegebener Liganden limitiert bzw. sind sie nicht ausreichend divers, können unter Umständen sehr restriktive, übertrainierte Modelle erzeugt werden. Im PBVS genutzt, sind sie nicht in der Lage, potentiell neue Liganden mit andersartigem Molekülgerüst zu identifizieren. Auch bei ausreichender Datenlage ist es nicht trivial zu entscheiden, welche Liganden bzw. Komplexe zur Modellierung herangezogen werden sollen. Unterschiedliche Trainingsmengen können zu unterschiedlichen Modellen führen, selbst wenn zur Modellierung dieselbe Methode genutzt wurde.[172] Die Qualität der Daten hat ebenso einen Einfluss auf das erstellte Modell. Insbesondere reagieren Modellierungsmethoden sensibel auf Wasserstoffausrichtungen, Protonierungszustände und tautomere Formen, da diese den Typ und die Lage der Merkmale vorgeben.[203, 227] Daher fordern die meisten Methoden entsprechend präparierte Liganden und Proteine. Strukturbasierte Modelle müssen zumeist manuell modifiziert werden, vor allem zur Reduzierung der Merkmale, um im virtuellen Screening eingesetzt werden zu können.[228] Ein weiteres Problem besteht darin, dass ligandbasierte Ansätze davon ausgehen, dass die bioaktive dreidimensionale Ligandstruktur einem energetisch optimalen Konformer entspricht. Liganden können jedoch auch in suboptimalen Konformationen[229, 230] und sogar in unterschiedlichen Modi binden. Um dieses Verhalten zu modellieren, müssen mehrere Hypothesen erstellt werden. Gewöhnlicherweise werden bereits mehrere

Modelle von den Methoden erstellt, die intern bewertet und entsprechend sortiert als Lösung präsentiert werden. Es ist die Aufgabe des Anwenders daraus passende zu wählen. Deshalb sollten automatisch erstellte Modelle stets validiert werden. Mögliche Ansätze zur Validierung sind die Bewertung verschiedener Hypothesen bezüglich ihrer statistischen Relevanz oder ihrer Vorhersageleistung in einem VS auf einem externen Datensatz.[231, 176, 177] Idealerweise wird die Hypothese durch experimentelle Testung potentiell neuer Liganden bewertet.[232] Nach Testung kann das Modell dann gegebenenfalls weiter verfeinert werden. Dafür erlauben die meisten Modellierungswerkzeuge einen Eingriff durch den Anwender. Bei ligand- und komplexbasierte Ansätzen müssen bzw. können eine Referenz zur Ausrichtung, Toleranzwerte, die Anzahl gemeinsamer Merkmale oder die Anzahl der durch die Hypothese beschriebenen Moleküle vorgegeben werden. Viele Software-Pakete zur Pharmakophormodellierung stellen zudem Editoren bereit, die es ermöglichen, einzelne Merkmale manuell zu verwerfen oder hinzuzufügen. Bei Ungewissheiten können Merkmale auch logisch ODER-verknüpft werden, um verschiedene alternative Merkmalszuordnungen während eines Merkmalsabgleichs in einem folgenden PBVS zu erlauben.

3.4.4 Anwendung im virtuellen Schreening

Sobald eine Hypothese ligand- oder strukturbasiert erstellt und validiert wurde, kann das Modell als Vorlage oder *Anfrage* genutzt werden, um Moleküle mit ähnlichen Merkmalen zu identifizieren. Solch ein pharmakophorbasiertes virtuelles Screening sieht sich mit folgender Aufgabe konfrontiert: Für jedes Molekül der Bibliothek muss festgestellt werden, ob es eine Merkmalskonfiguration besitzt, die durch die Anfrage beschrieben ist und mit dieser zur Deckung gebracht werden kann. Die Hypothese dient sozusagen als Schablone, um Merkmale von Molekülen einzupassen. Die zu lösende Aufgabe ist auch Bestandteil der ligandbasierten Modellierung, weshalb zum PBVS prinzipiell dieselben Algorithmen zur Anwendung kommen (vgl. Abschnitt 3.4.2). Allerdings muss dabei ausschließlich die Existenz einer gegebenen Konfiguration geprüft und geometrisch abgeglichen werden. Auf die Maximierung einer gemeinsamen Konfiguration kann verzichtet werden. Beim Merkmalsabgleich muss wiederum die Flexibilität der Moleküle berücksichtigt werden. Deshalb unterscheiden sich auch PBVS-Methoden dadurch, ob die Konformationen vorberechnet oder während des Merkmalsabgleichs erzeugt werden. Da das Screening allerdings mehr Moleküle abgleichen muss, gilt: je mehr vorberechnet werden kann, desto schneller lassen sich darauffolgenden Anfragen realisieren. Hauptsächlich haben sich deshalb Methoden wie MOE[169], Catalyst[233, 234, 235], Phase[183], Unity[236], LigandScout[199], Pharao[192] oder Pharmer[237] etabliert,

die alle vorberechnete Konformationen nutzen, um die Merkmale, der dann starren Konformationen, mit der Anfrage abzugleichen. Generell dient eine Pharmakophorhypothese im Screening als Filter. Ausschließlich das Vorhandensein und die Ausrichtung weniger Merkmale wird geometrisch evaluiert. Eine qualitative Bewertung der Resultate entfällt oder wird rudimentär, auf Basis der Abstandsquadrate der abgeglichenen Merkmale durchgeführt, sodass praktisch eine nicht priorisierte Submenge der ursprünglichen Bibliothek erhalten wird.

3.5 Vorbereitung eines virtuellen Screenings

LBVS, SBVS und PBVS-Verfahren verlassen sich darauf, dass das Anfragemolekül, Anfrageprotein bzw. Anfragepharmakophormodell und die verwendete Molekülbibliothek adäquat präpariert zur Verfügung gestellt werden. Art und Umfang der Präparierungsschritte hängt – neben der Qualität der gegebenen Daten – vor allem davon ab, welche Freiheitsgrade die genutzte Screening-Methode integriert. Unberücksichtigte Freiheitsgrade bieten Anlass für eine entsprechende Präparierung.

3.5.1 Aufbereitung der Molekülbliothek

3D-LBVS, SBVS und PBVS sind konformationsabhängige Verfahren. Damit sie effizient eingesetzt werden können, bietet sich eine Vorberechnung von Konformationen an. Mittlerweile ist man sich einig, dass auch Protomere in einem VS berücksichtigt werden müssen.[238, 239, 53] Es ist daher gängige Praxis geworden, in einem Präparierungsschritt die Molekülbibliothek mit diesen molekularen Zuständen anzureichern.

Konformergenerierung: Zur Erzeugung von Konformationen existieren diverse Methoden.[240, 241] Sie lassen sich entsprechend der Strategie zur Traversierung des Konformationsraum in deterministische und stochastische Methoden unterteilen. Zu den deterministischen Methoden gehören systematische Suchen[242, 243, 244, 245, 246, 247, 100, 169, 248]. Sie variieren Torsionen rotierbarer Bindungen gemäß vordefinierter, regelmäßiger Winkelwerte. Für eine effiziente Suche werden oft „teile-und-herrsche"-Strategien genutzt.[242, 249] Das Molekül wird fragmentiert, Konformationen werden für Fragmente erzeugt und vollständige Konformationen aus bereits berechneten Teilkonformationen zusammengesetzt. Dabei wird versucht, invalide Lösungen möglichst früh zu erkennen und nicht weiterzuverfolgen. Wenn die Diskretisierung der Torsionen allerdings zu fein gewählt wird, können systematische Suchen zu kombinatorischen Explosionen führen. Deshalb schränken wissensbasierte Ansätze[249, 250, 251,

100, 252, 253, 248, 254] die traversierten Torsionen gemäß der Winkel ein, wie sie typischerweise in Protein-Ligand-Komplexen beobachtet werden. Die Winkel und Ring-Template beruhen auf einer statistischen Analyse der experimentell bestimmten Strukturen und werden in Torsionswinkelbibliotheken verwaltet, um niederenergetische Konformere zu erzeugen.[100, 101, 255] Zufallssuchen variieren zufällig, über mehrere Iterationsschritte hinweg Atomkoordinaten oder Torsionen, bis die gewünschte Anzahl an Konformationen erzeugt wurde oder neue Konformationen nicht mehr gefunden werden können.[256, 257, 247, 169] Genetische Algorithmen kodieren dafür rotierbare Bindungen in einem Chromosom, wobei jedes Gen einen Freiheitsgrad repräsentiert. Ausgehend von einer zufällig gewählten Population werden durch Mutation- und Crossover-Operationen neue Individuen d. h. Konformationen erzeugt. Die Nachkommenschaft wird mit einer Fitnessfunktion bewertet, die z. B. Terme zur Bewertung der Torsionsspannung und Überlappungen enthält. Die besten Individuen werden zur weiteren Optimierung selektiert. Im Laufe des evolutionären Prozesses werden so sukzessiv bessere Konformationen vererbt.[258, 259, 260, 261, 262] Andere stochastische Verfahren variieren Konformationen, die durch Distanzmatrizen dargestellt sind. Sie variieren zufällig Distanzen zwischen Atomen und machen sich dabei zunutze, dass sich die Variablen nur in einem gewissen Bereich bewegen können.[263, 264, 265, 266, 200] Darüber hinaus können auch MD-Simulationen[267, 268], Monte-Carlo-Simulationen[251] oder Simulated-Annealing-Verfahren[269, 270] niederenergetische Konformere erzeugen. Diese Verfahren sind jedoch aufwändig und zum Prozessieren großer Molekülbibliotheken ungeeignet.

Protomergenerierung: Zur Generierung molekularer Zustände existieren verschiedene Werkzeuge, die ausgiebig oder systematisch Protonierungszustände und Tautomere enumerieren.[271, 272, 273, 274, 275, 169] Öffentlich verfügbare VS-Datenbanken, wie die ZINC-Datenbank (Zinc Is Not Commercial), sind sogar bereits entsprechend präpariert dem Anwender zur Verfügung gestellt.[276, 277] Prinzipiell können Zustandsenumeratoren Zustände entweder lokal oder global aufzählen.[54] Lokale Methoden weisen, basierend auf vordefinierten Regeln, spezifischen funktionellen Gruppen Zustände zu.[278, 279, 227, 273] Globale Methoden verteilen eine vordefinierte Anzahl an Protonen auf energetisch günstige Schweratome des gesamten Molekülgerüsts.[275] Beide Strategien erfordern eine Konfiguration durch den Anwender. Regeln müssen definiert bzw. die Anzahl zu verteilender Protonen festgelegt werden. Welche Zustände für ein VS jedoch überhaupt relevant sind, ist noch immer Gegenstand von Diskussionen. Studien der letzten Jahre untersuchten verschiedene Enumerierungsprotokolle und deren Einfluss auf Docking- und Screening-Resultate.[280, 281, 282, 283, 284, 285] Todorov *et al.* untersuchten den Einfluss von Zustandsvariationen des Liganden in Docking-Berechnungen,

die mit Gold[156, 130] durchgeführt wurden.[280] ten Brink und Exner führten Bindemodusvorhersagen mit Plants[132, 133, 134] und Gold[156, 130] durch. Sie enumerierten Protonierungszustände, Tautomere und Stereoisomere, die dazu genutzt wurden zuvor revidierte Zustände in Protein-Ligand-Komplexen des CCDC/Astex-Datensatzes zu rekonstruieren.[282] Eine Submenge des DUD (Directory of Useful Decoys)[286] wurde in einer Anreicherungsstudie mit AutoDock[128, 129] genutzt.[283] Milletti et al. fokussierten sich auf sieben Zielstrukturen des DUD-Datensatzes und verglichen die Anreicherungsleistung von Flap[194], Glide[140, 141, 287] und Gold auf Bibliotheken, die mit Tautomeren verschiedener Enumerierungsprotokolle angereichert wurden.[284] Da bei Berücksichtigung aller möglichen Protomere der Rechenbedarf extrem anstieg, wurden in den Studien generell weniger Zustände bevorzugt, um den VS-Prozess zu beschleunigen. Zudem wurde lediglich eine leicht verbesserte Anreicherungsleistung beobachtet, da die Bewertungsfunktionen schlechter aktive von inaktiven Molekülen unterscheiden konnten, wenn sie mit Posen vieler Zustände konfrontiert wurden. Da es zudem äußerst unwahrscheinlich ist, dass Liganden in einem energetisch ungünstigen Zustand binden, wurde im Allgemeinen die Empfehlung gegeben, die Bibliothek ausschließlich mit wenigen niederenergetischen Zuständen, die häufig in Wasser gelöst vorkommen, anzureichern.[288, 239, 289] Dies soll die seltenen Situationen abdecken, bei denen der komplexierte Zustand des Liganden nicht dem stabilsten, d. h. dem gelösten Zustand entspricht.

3.5.2 Aufbereitung der Proteinstruktur

Ein SBVS erfordert in der Regel mehrere Schritte zur Aufbereitung der Zielstruktur, die nicht Teil des Verfeinerungsprozesses von Röntgenkristallstrukturen sind.[290] Grundsätzlich müssen zumindest unaufgelöste Proteinatome zur Struktur hinzugefügt und Formalladungen und Bindungsordnungen bestimmt werden, sodass Docking-Algorithmen die Struktur korrekt interpretieren können. Teilweise sind Stick- und Sauerstoffatome in Kristallstrukturen nicht klar voneinander zu unterscheiden. Dies betrifft vor allem die Seitenketten von Asn, Gln und His, die bezüglich ihrer Form symmetrisch erscheinen, bezüglich ihrer elektrostatischen Eigenschaften jedoch asymmetrisch sind und tatsächlich um 180° gedreht vorliegen können. Wasserstoffe sind in Röntgenkristallstrukturen überhaupt nicht oder nur unzureichend aufgelöst. Aus diesem Grund sollten sie initial zur Struktur hinzugefügt bzw. erneut berechnet werden. Für die meisten der 20 proteinogenen Aminosäuren können Wasserstoffe anhand wohldefinierter Geometrien platziert werden. Für Cys, Ser, Thr und Tyr sind die Wasserstoffpositionen jedoch nicht eindeutig,

da sie Bestandteil terminaler, frei rotierbarer Hydroxyl- bzw. Thiolgruppen sind. Bei ionisierbaren Gruppen ergibt sich zudem die Frage, ob sie ihre neutrale oder geladene Form annehmen. Diese Variablen sind durch ihre direkte chemische Umgebung beeinflusst. Zur Bestimmung der tatsächlichen Anzahl und der Positionen von Wasserstoffen sollten sie deshalb im System betrachtet werden und ihr Zustand im Hinblick auf Etablierung eines energetisch günstigen, möglichst optimalen Wasserstofbrückennetzwerks im Protein bestimmt werden. Sind Kofaktoren und/oder ein Referenzligand gegeben, so sind sie Teil des Netzwerks und ihre Freiheitsgrade müssen ebenso optimiert werden. Auf Ligandseite ist die Zuordnung von Wasserstoffen und Bindungsordnungen jedoch wesentlich schwieriger, da der Ligand zumeist nicht einem von nur 20 möglichen wohldefinierten Teilstrukturen entspricht. Dennoch existieren verschiedene Methoden, die es ermöglichen Wasserstoffe in Protein-Ligand-Komplexen hinzuzufügen und auszurichten, Hybridisierungszustände der Atome zu bestimmen, Bindungsordnungen zuzuweisen und andere strukturelle Ungewissheiten aufzulösen.[291, 292, 293, 294, 295, 296, 297, 298, 299] Da diese Schritte für einen Docking-Erfolg so wichtig sind, finden sie sich oft als integrale Bestandteile in Präparierungsprotokollen wieder, die mitunter auch in der Lage dazu sind, entscheidende, bindungvermittelnde Wassermoleküle, die Koordination von Metallen und alternative bzw. fehlende Seitenketten zu modellieren. Dennoch ist es ratsam, die Präparierungsschritte zu verifizieren[300], da für einige Zielstrukturen die Präparierung das Resultat eines Screenings beeinflussen kann.[301] Die Proteinpräparierung ist daher ein äußerst interaktiver Prozess, der eine manuelle Intervention durch den Anwender erfordert.

3.6 LBVS, SBVS und PBVS im Vergleich

Unterschiedliche VS-Strategien haben sich ursprünglich aus der Situation heraus entwickelt, dass nicht immer Protein- und Ligandinformation für eine betrachtete Zielstruktur gegeben war. So wurde auf Basis der Datenlage ein geeignetes Screening-Verfahren zur Anwendung gewählt. Dieser Umstand hat sich in den letzten Jahren geändert. Heute sind häufig Proteinstrukturen aufgeklärt und bioaktive Moleküle bekannt. Der Anwender kann sich somit frei zwischen einem LBVS, SBVS oder ein PBVS entscheiden. Der Einsatz von 2D- und 3D-LBVS-Methoden resultiert allerdings in Hitlisten mit Verbindungen, die bezüglich der Ähnlichkeit zur Referenz priorisiert sind.[65] Vor allem bei topologischen LBVS-Ansätzen führt dies dazu, dass zur Referenz nahe Analoge in den oberen Rängen der Hitliste angereichert werden. Im VS ist man aber oft daran interessiert möglichst diverse Moleküle zu identifizieren, die als potentielle Leitstrukturen für spätere Optimierungen fungieren können. Solche Resultate können von einem SBVS

erwartet werden. Da sich die Ergebnisse von LBVS-Methoden von denen der SBVS-Methoden so unterscheiden, werden sie als „komplementär“ betrachtet und in Kombination gerne eingesetzt, um ein Gesamtbild von potentiellen Liganden (diverse und analoge) zu erhalten.[67] Ein Protein-Ligand-Docking kann sowohl die Komplementarität als auch die Gesamtheit kurz- und weitreichender Interaktionen bewerten und Moleküle priorisieren. Moleküle, die nicht in die Proteintasche eingepasst werden können, werden verworfen. Auf einer umfangreichen Bibliothek angewandt, führt die genauere Auswertung der Moleküle jedoch zu einem geringeren Durchsatz. Zudem modellieren die Bewertungsfunktionen zumeist nur Beiträge, von denen mit Gewissheit angenommen werden kann, dass sie in Protein-Ligand-Komplexen etabliert werden. Seltene Interaktionen und schwer einzuschätzende entropische Beiträge werden in der Regel vernachlässigt. Dies ist mitunter ein Grund, warum Docking-Methoden bei der Vorhersage der Bindungsenergie versagen und Schwierigkeiten haben, Posen nahe der bioaktiven Konformation vorherzusagen.[302] Aufgrund der simplen Beschreibung der Hypothesen lässt sich ein PBVS oft effizienter als ein SBVS lösen und ist deshalb in größerem Umfang anwendbar. Allerdings produziert es höhere falsch-positiv-Raten, da die Beschränkungen durch wenige Merkmale im Vergleich zu einem vollständigen Abbild der Proteinbindetasche weniger selektiv sind. Die Erweiterung der Pharmakophorbeschreibung durch Merkmale, die die Molekülform beschreiben, kann falsche Vorhersagen reduzieren.[303] Zudem können nicht typisierte Ausschlussmerkmale eine Restriktion durch den Proteinraum simulieren. Allerdings sind sie meist so „löchrig“, dass es möglich ist, Moleküle in Hohlräumen zu platzieren.[174]

3.7 Pharmakophorgeleitetes Docking

Es konnte gezeigt werden, dass eine Kombination von pharmakophor- und strukturbasiertem virtuellen Screening zu deutlich verbesserten Anreicherungsraten führt.[210, 304, 305] Ist sowohl Protein- als auch Ligandinformation gegeben können somit unterschiedliche Strategien vereint werden und helfen, die Schwächen eines Ansatzes durch die Stärken eines anderen auszugleichen. Prinzipiell könnten dafür die verschiedenen Ansätze konkateniert werden. Eine sukzessives Ausführen der Methoden hat aber den entscheidenden Nachteil, dass die gegebene Referenzinformation auch nur nacheinander ausgewertet werden kann. Dagegen kann eine Vereinigung der Ansätze innerhalb eines Werkzeugs, wie sie von *pharmakophorgeleiteten Docking-Methoden* realisiert wird, vollständig auf die gegebene Information zurückgreifen und sie zugleich verarbeiten. Gegebene Pharmakophormodelle können so den Docking-Prozess dieser Methoden lenken. Pharmakophorgeleitete Docking-Methoden unterscheiden sich hauptsächlich darin, wie

die Zusatzinformation des Pharmakophormodells im Docking-Prozess verarbeitet wird. Die Pharmakophorinformation kann dazu verwendet werden, die Bewertungsfunktion anzupassen oder den Suchraum der Docking-Methode systematisch einzuschränken:

Anpassung der Bewertungsfunktion: Manche Methoden wie GEMDOCK[306], SP-DOCK[307] und GOLD[308] modifizieren anhand einer Pharmakophorhypothese ausschließlich die Bewertungsfunktion der zugrundeliegenden Docking-Methode. Zusätzliche, korrigierende Terme bewerten die Ähnlichkeit der Posen zur Hypothese und geben ihnen ein höheres Gewicht wenn sie die Hypothese erfüllen.

Systematische Einschränkung des Suchraums: FLEXX-PHARM[309], eine Erweiterung von FLEXX, verwirft Teillösungen während des inkrementellen Aufbaus des Liganden, sobald es nicht mehr möglich ist, dass die Gesamtlösung die gegeben Merkmale der Hypothese erfüllen kann. Dadurch wird der Suchraum effektiv eingeschränkt und indirekt neue Posen präsentiert, denen zuvor, aufgrund schlechterer Bewertung im gewöhnlichen Docking, keine Beachtung geschenkt wurde.

Die Anpassung der Bewertungsfunktion kann eine verbesserte Vorhersage des Bindungsmodus und eine verbesserte Anreicherungsleistung erreichen. Die systematische Einschränkung des Suchraums bietet darüber hinaus die Möglichkeit, den Docking-Prozess zu beschleunigen. Dadurch wird die Methode effizient im VS auf umfangreichen Bibliotheken anwendbar. Neben den synergetischen Effekten bezüglich der Vorhersagequalität und Effizienz legt ein pharmakophorgeleiteter Docking-Ansatz zusammen mit dem äußerst interaktiven Prozess der Pharmakophormodellierung außerdem den Grundstein für einen VS-Prozess, der sich durch den Anwender steuern lässt.

3.8 Protomerintegration

Während es üblich geworden ist, durch Präparierung Protomere in Molekülbibliotheken bereitzustellen, wird die Proteinstruktur den strukturbasierten Methoden nur in einem initial aufbereiteten Zustand zur Verfügung gestellt. Eine einmalig präparierte Proteinstruktur kann aber ausschließlich einen komplementären Molekülzustand erkennen. Wurden die Wasserstoffkoordinaten des Proteins zuvor im Kontext eines bereits bekannten Liganden bestimmt, so ist der Proteinzustand ein Abdruck des Referenzliganden bezüglich seiner Wasserstoff- und Elektronenpaarpositionen. In einem VS ist man jedoch mit unterschiedlichen Liganden konfrontiert. Ein anderer Ligand kann prinzipiell die chemische Umgebung des Bindungspartners ändern und somit auch im Protein einen anderen Zustand induzieren. Der Einfluss derartiger Proteinzustandsänderungen auf Docking- und Screening-Resultate ist bislang allerdings nur wenig untersucht. Kim *et*

al. präparierten verschiedene Rezeptormodelle des RmlC-Enzyms, die in einer Ensemble-Screening-Studie verwendet wurden.[310] Sie beobachteten eine Abhängigkeit der Anreicherungsleistung von den initial modellierten Histidinzuständen und forderten, die Freiheitsgrade in Docking-Berechnung miteinzubeziehen. Im Allgemeinen können makromolekulare und molekulare Zustände im SBVS mittels zweier unterschiedlicher Herangehensweisen simultan berücksichtigt werden. Die *naive Vorgehensweise* präpariert Protein- und Molekülzustände, die dann in einem *Ensembleansatz* bewertet werden. Die Alternative ist es, Protomere als zusätzliche Freiheitsgrade des Docking-Problems anzusehen (vgl. Abschnitt 3.3.2), die in einem *integrierten Ansatz* vorhergesagt werden.

Naive Vorgehensweise (*Ensembleansatz)*: Sind Molekülzustände bereits in einer Bibliothek bereitgestellt, dann kann man in Analogie zur Präparierung der Molekülbibliothek verfahren (vgl. Abschnitt 3.5.1), um im SBVS auch variable Proteinzustände zu berücksichtigen. Dafür werden ebenfalls präparierend alternative Protomere des Proteins enumeriert. Für jeden Proteinzustand wird dann ein Screening-Lauf initiiert. Die resultierenden Hitlisten werden im Anschluss vereint. Die Vereinigung kann dadurch bewerkstelligt werden, dass für jedes identifizierte Molekül nur die beste Bewertung aus allen angereicherten Protein-Ligand-Zustandskombinationen gewählt wird. Die ausgewählten Bewertungen werden dann erneut sortiert. Abbildung 3.4 stellt dieses Verfahren im Falle eines Liganden schematisch dar. Es ist unter der Bezeichnung *Ensemble-Docking* bekannt und wird im Allgemeinen dazu verwendet, Proteinflexibilität abzubilden.[94, 95] Wird nach diesem Schema eine ganze Molekülbibliothek evaluiert, so spricht man von einem *Ensemble-Screening*.

Integrierter Ansatz: Protomere des Proteins und des Liganden können aber auch als Freiheitsgrade des Docking-Problems betrachtet werden mit dem Ziel, während der Platzierung komplementäre Zustände beider Komponenten vorherzusagen. Nur eine Publikation beschreibt die Integration derartiger Freiheitsgrade in den Docking-Prozess. eHits[311, 125], das eine Placement-and-Linking-Strategie verfolgt (vgl. Abschnitt 3.3.4), platziert Fragmente unter Berücksichtigung der Protonierungszustände von Protein und Ligand. Protein und Ligandfragmente werden von eHits durch typisierte Oberflächenpunkte dargestellt, die starke und schwache Donoren, starke und schwache Akzeptoren oder hydrophobe Merkmale repräsentieren. Unter der Annahme, dass Atome mit freien Elektronenpaaren protoniert werden und ihre Funktion ändern können, erhalten Punkte in diesen Bereichen Donor- und Akzeptortypen. Bei der Platzierung der Fragmente anhand vordefinierter Orientierungen werden die Typen individuell ausgewertet und derjenige selektiert, der eine Interaktion etabliert. Der Ansatz

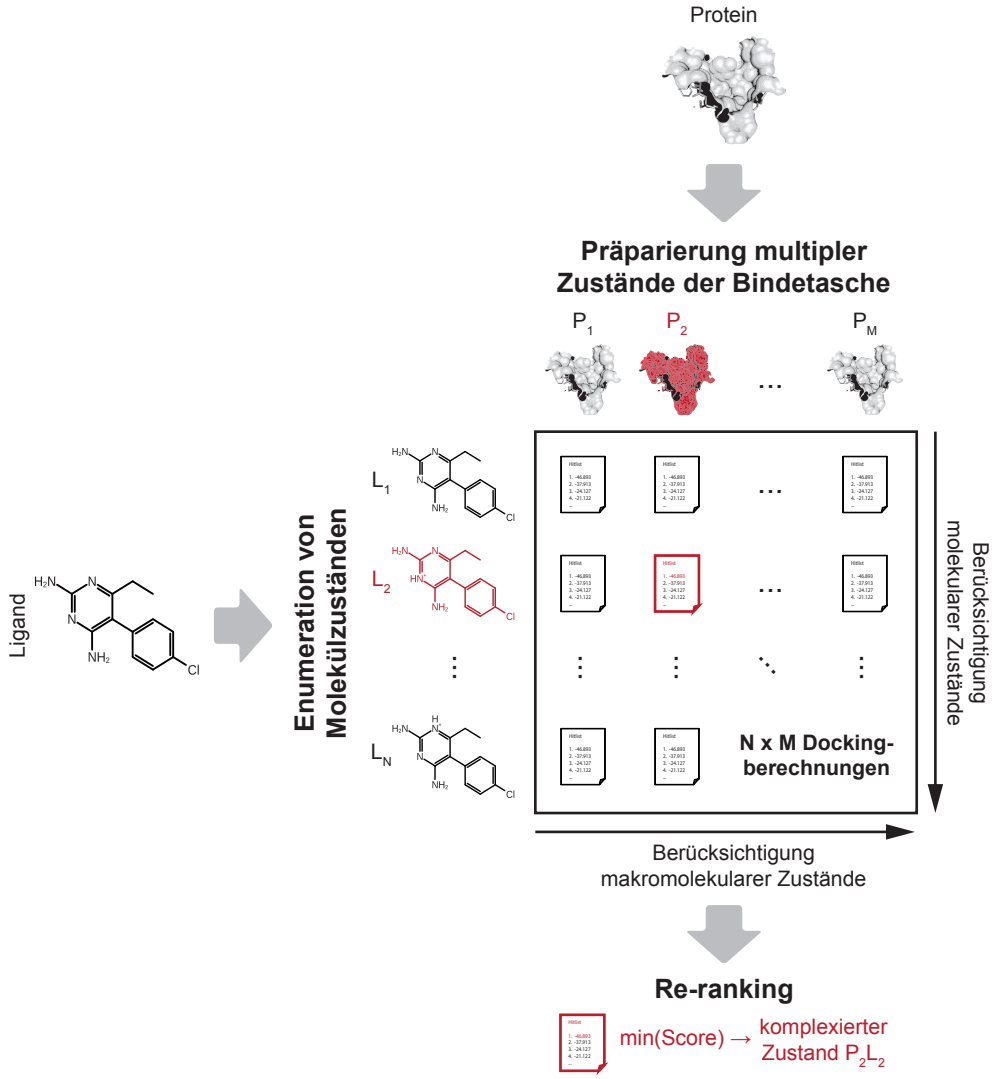


Abbildung 3.4: Ensemble-Docking zur Berücksichtigung von Protein- und Ligandzuständen: Vor der Docking-Berechnung werden Zustände von Protein und Ligand enumeriert. Jede mögliche Zustandskombination wird separat gedockt. Die Kombination, die die beste Bewertung erzielt definiert den Zustand im vorhergesagten Komplex.

von eHits modelliert so alle möglichen Protonierungszustände. Ein Nachteil ist allerdings, dass durch die lokale Handhabung der Zustände extreme, äußerst unrealistische und energetisch ungünstige Protonierungen forciert werden.

Grundsätzlich integrieren Docking-Methoden (bis auf eHits) keine Protomerfreiheitsgrade. Molekülbibliotheken werden zwar oft inklusive Molekülzustände bewertet,

ein Ensembleansatz zur Bewertung von Proteinzuständen wird in der Regel aber nicht angewandt. Die Ursache hierfür ist, dass die Strategie vor allem bei Zielstrukturen mit vielen alternativen Proteinzuständen kombinatorisch explodiert und somit nicht mehr effizient ausgeführt werden kann. Werden Proteinzustände missachtet oder besitzt ein Protein nur einen Zustand, bringt aber auch die mittlerweile gängige Präparierung der Molekülbibliothek mit Zuständen (vgl. Abschnitt 3.5.1) entscheidende Nachteile für den VS-Prozess: Moleküldateiformate können nur einen einzelnen Zustand eines Moleküls repräsentieren. Protomere unterscheiden sich aber lediglich in der Position von wenigen Wasserstoffen und im Typ weniger Bindungen. Eine Enumeration von Protomeren expandiert somit nicht nur Bibliotheken, sondern reichert verstärkt redundante Molekülinformation in Bibliotheken an, nämlich die bezüglich der Zustandsvariation invarianten Molekülbereiche. Besonders in Kombination mit gleichzeitiger Enumeration von Konformeren erfordert dies unnötigerweise gesteigerte Speicherkapazitäten und führt dazu, dass im Screening wiederholt fast identische Moleküle initialisiert und interne Molekülrepräsentation aufgebaut werden. In einem SBVS werden dann wiederholt, fast identische Docking-Berechnungen durchgeführt, die zu teils identischen Docking-Lösungen führen und die Bewertung der Verbindungen maßgeblich erschweren. Wird analog dazu auf Proteinseite verfahren, potenzieren sich diese Effekte noch. Eine interne Traversierung des *Protomerraums* kann die Genauigkeit der Vorhersagen steigern, überflüssige Berechnungen vermeiden und den Einsatz im groß angelegten VS ermöglichen. Vor allem gewährleistet es aber konsistente Vorhersagen, unabhängig von den gegebenen Zuständen, die letztendlich ein und dasselbe Molekül bzw. Protein repräsentieren.

4 Integrierte Methoden

Dieses Kapitel beschreibt die in cRAISE integrierten Modelle und Methoden. Die Komponenten wurden in den letzten Jahren an der Universität Hamburg als eigenständige Werkzeuge entwickelt und realisieren Aufgaben wie das Lesen und Schreiben gängiger Molekül- und Proteindateiformate (Naomi[3, 312, 313]), das Verwalten und Analysieren molekularer Bibliotheken (Mona[314]), sie bilden die Flexibilität kleiner Moleküle ab (Confect[255]) oder optimieren das Wasserstoffbrückennetzwerk eines Protein-Ligand-Komplexes (Protoss[315]). Ihre Basisfunktionalitäten wurden in einer Softwarebibliothek – der Naomi-Bibliothek – vereint. Diese löste nach und nach die bis dato etablierte *Flex*-Bibliothek* ab. Im Zuge dieser Neuentwicklungen entstand auch cRAISE. Es baut auf den Konzepten seiner Flex*-basierten Vorgängerversionen TrixX[316] und TrixX-BMI[2] auf, verbindet jedoch die Naomi-basierten Komponenten und nutzt oder erweitert deren Modelle.

4.1 Naomi: Molekül- und Proteininitialisierung

Jeder Berechnung auf molekularen Daten geht eine Molekülinitialisierung voraus. Sie interpretiert die gegebenen Daten und baut eine interne Molekülrepräsentation auf. Auch wenn identische Moleküle in unterschiedlichen Formaten oder Formen gegeben sind, muss die Initialisierung sicherstellen, dass die Daten so interpretiert werden, dass sie zur selben internen Repräsentation führen. Nur dann können konsistente Folgeberechnungen gewährleistet werden. Diese zentrale Anforderung stand im Fokus der Entwicklung von Naomi[3, 312], das als Werkzeug zur Konvertierung zwischen gängigen molekularen Dateiformaten dient. Durch die Wichtigkeit seiner Funktion für alle weiteren Entwicklungen ist Naomi der Kern und zugleich Namensgeber der Naomi-Bibliothek. cRAISE repräsentiert intern Moleküle und Proteine gemäß des Naomi-Models und greift an vielen Stellen auf die dadurch gegebene chemische Information zurück.

4.1.1 Informationsgehalt molekularer Daten

Die im computergestützten Wirkstoffentwurf verwendeten Daten sind diversen Ursprungs. Sie folgen verschiedenen Formaten, die einen unterschiedlichen Informationsgehalt aufweisen. Molekulare Dateiformate wie das Structure Data File[317] (SDF), das Tripos Sybyl MOL2 Format[318] (MOL2) oder das Protein Data Bank Format[319] (PDB) enthalten kartesische Koordinaten zur dreidimensionalen Repräsentation von Molekülen. Andere Formate wie das Simplified Molecular Input Line Entry System[320] (SMILES), die SYBYL Line Notation[321] (SLN) oder der IUPAC International Chemical Identifier[322] (InChI) kodieren Moleküle in Form von Zeichenketten zur zweidimensionalen, topologischen Repräsentation. Im strukturbasierten Design ist die dreidimensionale Darstellung unumgänglich, somit beschränkt sich cRAISE, ohne eine vorbereitende Koordinatengenerierung, auf die Verwendung der SDF-, MOL2- und PDB-Formate. Tabelle 4.1 stellt die wesentlichen Unterschiede bezüglich der enthaltenen Information der 3D-Formate gegenüber. Das SDF-Format ist zum Austausch kleiner Moleküle konzipiert, die MOL2- und PDB-Formate erlauben auch die Speicherung von Proteindaten. Dabei beschreibt das PDB-Format hauptsächlich die Konnektivität zwischen den Atomen und spezifiziert selten Bindungstypen. Im SDF wird regelmäßig und im PDB fast ausschließlich auf die Angabe von Wasserstoffen verzichtet. Das SDF-Format annotiert Formalladungen, MOL2 und PDB dagegen kaum. Als Folge dieser Unterschiede kann das SDF-Format nur eine lokalisierte Valenzstruktur eines Moleküls darstellen. Es ist somit möglich, mehrere Resonanzformen[1] zur Beschreibung eines delokalisierten Systems anzugeben. Das MOL2-Format kann äquivalente Resonanzformen typischer funktioneller Gruppen durch eine einzige delokalisierte Repräsentation darstellen. Es ist dennoch möglich, die Delokalisierung funktioneller Gruppen verschiedenartig zu beschreiben.

Tabelle 4.1: Unterschiede im Informationsgehalt gängiger 3D-Moleküldateiformate.

Format	Datenart	Bindungstyp	H-Atome	Ladung	Delokalisierung
SDF	Molekül	ja	selten	ja	nein
MOL2	Molekül/Protein	ja	oft	selten	möglich
PDB	Molekül/Protein	selten	selten	nein	nein

[1] Moleküle werden häufig als Valenzstrukturen (Lewis-Strukturen)[323] dargestellt. Allerdings beschreiben sie die Bindungsverhältnisse oft nicht zutreffend. Bindungen können von Einfach-, Doppel- oder Dreifachbindungen abweichen. Elektronen und Partiallladungen sind nicht eindeutig Atomen zuzuordnen, sondern liegen delokalisiert vor. Deshalb beschreibt man das Molekül mit mehreren *Resonanzstrukturen* (mesomere Grenzstrukturen). Eine Resonanzstruktur ist eine Valenzstruktur, die keinen wirklichen molekularen Zustand repräsentiert. Vielmehr ist der wirkliche Zustand ein Zwischenzustand zwischen den Resonanzformen. Resonanzstrukturen sind somit Hilfsmittel zur Behebung der Mängel einer unzureichenden Molekülrepräsentation.

4.1.2 Das NAOMI-Modell

NAOMI[3] stellt das chemische Model zur internen Repräsentation von Molekülen bereit und gewährleistet die konsistente Initialisierung aus Moleküldateien. Entsprechend des NAOMI-Modells ist ein *Molekül* ein ungerichteter Graph, dessen Knoten Atome und dessen Kanten Bindungen repräsentieren. Jedem Atom sind, neben seinen kartesischen Koordinaten, hierarchisch drei Klassen chemischer Information zugeordnet:

Element: Das Element ist über sein Elementsymbol e eindeutig charakterisiert und entspricht einem der Elemente des Periodensystems. Anhand des Elementsymbols sind Elementname, Ordnungszahl, relative Atommasse, Van-der-Waals-Radius, kovalenter Radius, die Kennzeichnung als Metall und die Anzahl der Valenzelektronen bestimmt.

Valenzzustand: Der Valenzzustand ist ein Tupel $v = (e, b_1, b_2, b_3, c)$. e bezeichnet das Element, b_1, b_2 und b_3 die Anzahl der Einfach-, Doppel- und Dreifachbindungen und c die Formalladung. Anhand der Information ist die Anzahl freier Elektronen ersichtlich und ob das Atom Teil eines konjugierten oder aromatischen Systems sein kann.

Atomtyp: Der Atomtyp enthält Information, die die nähere Atomumgebung betrachtet. Er weist dem Atom eine Idealgeometrie gemäß der VSEPR-Theorie[324] (Valence Shell Electron Pair Repulsion) zu, die dessen Hybridisierungszustand reflektiert. Er klassifiziert Atome als konjugiert oder aromatisch und markiert Atome in delokalisierten Elektronensystemen. Außerdem wird das Atom als potentieller Wasserstoffbrückendonor oder -akzeptor ausgezeichnet. Der Atomtyp kann nur anhand einer validen Valenzzustandsfolge eines Moleküls abgeleitet werden.

Die zentrale Annahme des NAOMI-Modells und der darauf beruhenden Initialisierungsroutine ist, dass jedes n-atomige Molekül zumindest eine valide Valenzstruktur d. h. eine *Valenzzustandsfolge* $V = (v_1, v_2, \cdots, v_n)$ besitzen muss. Da die Valenzstruktur nur eine Grenzform des Moleküls darstellt, kann ein Molekül durch mehrere äquivalente Valenzzustandsfolgen repräsentiert sein. Die initialisierte Valenzzustandsfolge ist durch die Resonanzform der Eingabe bedingt. Von einer einzelnen Valenzstruktur kann jedoch darauf geschlossen werden was die Darstellung eigentlich meint. Diese Information ist auf Atomtypebene repräsentiert. Unterschiedliche Valenzzustandsfolgen eines Moleküls sind auf Atomtypebene durch identische Atomtypfolgen beschrieben. Der Atomtyp charakterisiert somit eindeutig die physikochemischen Eigenschaften eines Atoms, unabhängig vom gegebenen Valenzzustand. Der Unterschied zwischen Element, Valenzzustand und Atomtyp ist auch in Abbildung 4.1 am Beispiel des Carboxylats erklärt.

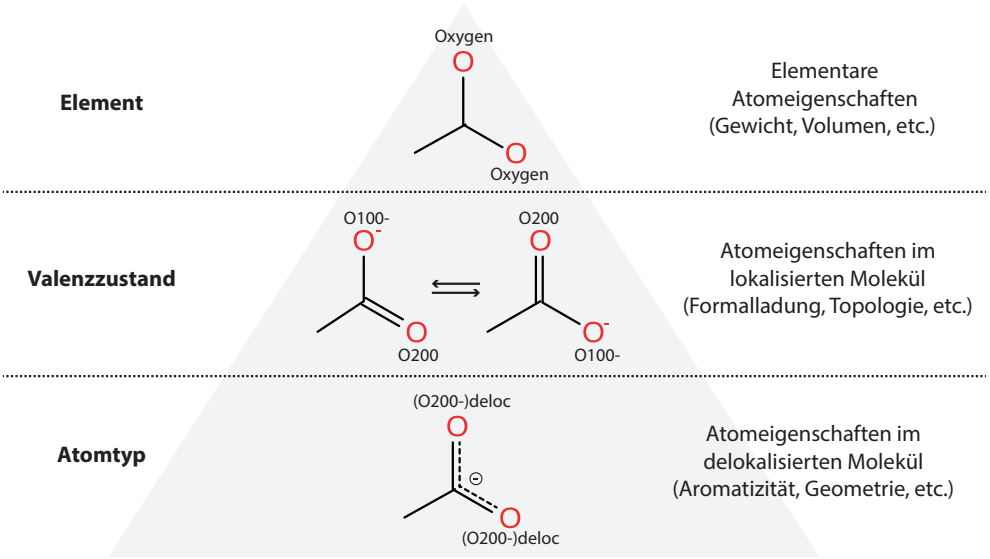


Abbildung 4.1: Naomi-Molekül: Auf Elementebene werden Eigenschaften anhand der Elementinformation und auf Valenzzustandsebene anhand der lokalisierten Valenzstruktur abgeleitet. Die Valenzzustandsrepräsentation ist nicht eindeutig. Nur die Atomtypebene beschreibt eindeutig die physikochemischen Eigenschaften eines Moleküls.

4.1.3 Molekülinitialisierung

Die Initialisierung gliedert sich entsprechend oben gegebener Reihenfolge und Anzahl an Informationsklassen in drei Phasen, die sukzessive durchlaufen werden. Sie sammeln die in der Moleküldatei bereitgestellten Element-, Bindungstyp- und Konnektivitätsdaten sowie Formalladungen, interpretieren sie entsprechend der Formatspezifikation, leiten weitere notwendige Informationen ab und annotieren sie schließlich an den Atomen:

Zuweisung des Elements: Aus den Konnektivitätsdaten wird ein Molekülgraph gebaut, an den die Elemente und initiale Bindungstypen zugewiesen werden. Dabei werden die Elemente des Periodensystems und Einfach-, Doppel-, Dreifach- und aromatische Bindungstypen unterstützt. Bei undefinierten Atom- und Bindungstypen muss der Initialisierungsprozess abgebrochen und das Molekül verworfen werden, da diese Information für die folgenden Initialisierungsschritte notwendig ist.

Zuweisung des Valenzzustands: In diesem Schritt wird versucht, eine valide Valenzstruktur des Moleküls zu erzeugen. Die Valenzstruktur ist genau dann valide, wenn es möglich ist, jedem Atom einen Valenzzustand zuzuweisen und aromatische Bindungen zu lokalisieren. Sind Formalladungen, Wasserstoffe und lokalisierte Bindungsordnung gegeben, so kann der Valenzzustand direkt bestimmt werden. Bei Fehlen von Ladungen

oder mehrerer Eigenschaften wird zusätzliche Information, z. B. der Sybyl-Atomtyp, aus den Ursprungsdaten herangezogen, um eine eindeutige Interpretation zu gewährleisten. Kann für ein Atom so kein Valenzzustand definiert werden, wird die direkte Atomumgebung nach typischen, vordefinierten Korrekturmustern durchsucht. Der korrespondierende Valenzzustand wird zugewiesen und die Bindungsordnungen und Valenzzustände der Umgebung adaptiert. Schlägt die Zuweisung fehl, so wird das Molekül als fehlerhaft identifiziert und verworfen. Bei Erfolg werden die Bindungen aromatischer Systeme lokalisiert. Da für Moleküle mit kovalent gebundenen Metallen keine klaren Valenzregeln definiert sind, akzeptiert die Initialisierung nur einfache Metallionen.

Zuweisung des Atomtyps: Die Atomtypzuweisung generiert eine delokalisierte Beschreibung des Moleküls. Dieser Schritt ist unabhängig von den gegebenen Daten, da die notwendige Information in der bis dahin erstellten internen Molekülrepräsentation zur Verfügung steht. Zunächst werden die Ringe des Moleküls identifiziert[325] und als delokalisiert markiert wenn sie alternierende Einfach- und Doppelbindungen enthalten und die Anzahl delokalisierter Elektronen die Hückel-Regel[326] erfüllt. Zur Identifizierung von Resonanzformen wird das Molekül anhand seiner konjugierten Systeme partitioniert. Die Systeme sind aus den Valenzzuständen und der Ringinformation erkennbar. Für jedes System wird überprüft, ob Atompaare ihre Formalladung tauschen können, alle Resonanzformen werden aufgezählt und Atome mit delokalisierten Ladungen markiert. Schließlich werden anhand der Valenzzustände, der Information über das konjugierte System und der Delokalisierung passende, vordefinierte Atomtypen gewählt.

Die Molekülinitialisierung aus dem PDB-Format weicht von der Standardroutine ab. Aufgrund fehlender Bindungstypen kann keine definierte Bindungsordnung am Molekülgraph annotiert und direkt kein Valenzzustand ermittelt werden. NAOMI begegnet dem Problem indem es die lokale Umgebung von Atomen anhand der gegebenen Koordinaten geometrisch bewertet.[312] Potentielle Valenzzustände werden daraufhin überprüft, wie gut sie die gegebene Atomgeometrie erklären. Der bestbewertete Valenzzustand wird zugewiesen. Nach erfolgreicher Zuweisung kann der Rest der Initialisierungsroutine durchlaufen und eine vollständige interne Repräsentation des Moleküls erstellt werden.

4.1.4 Proteininitialisierung

Das Fehlen von Bindungstypen in PDB-Strukturen ist der Tatsache geschuldet, dass das Format primär zur Darstellung makromolekularer Proteinkomplexe dient. Sie können sich aus mehreren Proteinketten zusammensetzen. Aufgrund mangelnder Auflösung der

Daten können Proteinketten auch Kettenbrüche aufweisen. Proteinkomplexe und speziell Protein-Ligand-Komplexe enthalten neben Polypeptidketten auch Liganden, Kofaktoren, Metalle und Wasser. Um Proteine und Komplexe intern zu repräsentieren sind innerhalb der Naomi-Bibliothek weitere Strukturen definiert. Ihre Initialisierung[312] beruht wesentlich auf dem Vorgehen zur Molekülinitialisierung.

Residue: Eine Residue ist ein Naomi-Molekül mit unvollständigem Molekülgraph. Sie repräsentiert eine Aminosäure mit offenen Valenzen. Eine Residue besitzt einen Typ, der einem der 20 proteinogenen Standardaminosäuren entspricht oder undefiniert ist. Zudem annotiert sie ihre Position innerhalb der Aminosäuresequenz und die Kette in der sie zu finden ist. Bei Initialisierung wird für jede Residue ein Molekülgraph erstellt und aus den gegebenen Daten Elemente, Typ und Position identifiziert. Anhand vordefinierter Mustern der Standardaminosäuren wird die Bindungsordnung für jedes Atom ermittelt. Dadurch lassen sich Valenzzustände und Atomtypen bestimmen. Fehlende Wasserstoffe werden entsprechend der Atomgeometrie hinzugefügt. Unaufgelöste Atome werden als solche gekennzeichnet. Atome des N- und C-Terminus erhalten einen Atomtyp mit offener Valenz. Aminosäuren, die nicht einer der Standardaminosäuren entsprechen, werden mit der koordinatenbasierten Routine für PDB-Moleküle initialisiert.

Proteinmolekül: Ein Proteinmolekül ist eine Menge miteinander verbundener Residuen, die einen abgeschlossenem Molekülgraph ohne offenen Valenzen bilden. Proteinmoleküle entsprechen nicht den nach dem PDB-Format definierten Proteinketten. Sie beschreiben lediglich zusammenhängende Residuen. Bei ihrer Initialisierung wird jede Residue (außer im Fall eines Kettenbruchs oder eines Kettenendes) mit ihrer vor- und nachfolgenden Residue verbunden. Atome mit offenen Valenzen erhalten einen Atomtyp, sodass die Verbindung der Residuen einer Peptidbindung entspricht.

Protein: Ein Protein besteht aus ein oder mehreren Proteinmolekülen, die in ihrer Gesamtheit alle Polypeptide des gegebenen Proteinkomplexes repräsentieren. Bei Initialisierung werden alle Proteinmoleküle im Protein vereint.

Komplex: Im Komplex werden Protein und Moleküle vereint. Liganden, Kofaktoren, Metallionen und Wasser werden durch Naomi-Moleküle repräsentiert. Metallionen und Wassermoleküle sind als solche gekennzeichnet und vom Rest separiert.

4.2 Naomi: Repräsentation molekularer Zustände

Naomi kann auch als Werkzeug zur Kanonisierung (Konvertierung unterschiedlicher Molekülformen auf eine eindeutige Repräsentation), zur Normalisierung (Erzeugung eines molekularen Grundzustands) und zur Generierung von Protonierungszuständen und

Tautomeren genutzt werden.[313] cRAISE nutzt direkt die Information, die durch das zugrundeliegende *Model der Valenzzustandskombinationen (VSC-Modell)* gegeben ist, um zu entscheiden, ob ein Molekül über alternative Tautomere und Protonierungszustände verfügt und welche Atome durch mögliche Zustandsänderungen betroffen sind.

4.2.1 Das Model der Valenzzustandskombinationen

Nach Initialisierung eines NAOMI-Moleküls unterscheiden sich unterschiedliche Formen eines Eingabemoleküls lediglich in den Valenzzuständen ihrer Atome und in der Verteilung der Bindungsordnungen. Ihre Molekülgraphen sind identisch. Um zu erkennen, ob es sich bei zwei Molekülen um alternative Zustände handelt, können deshalb Valenzzustandsfolgen oder *Valenzzustandskombinationen* (VSC) verglichen werden. Tabelle 4.2 stellt die Unterschiede einzelner Valenzzustände dar, die nach dem VSC-Modell in den Folgen auftreten können. Eine Valenzzustandsänderung kann wie folgt klassifiziert wer-

Tabelle 4.2: Mögliche Arten von Valenzzustandssubstitutionen in Valenzzustandsfolgen und deren charakteristische Eigenschaften. ± bezeichnet eine Änderung, 0 keine Änderung der Eigenschaft des substituierten Valenzzustands (nach Urbaczek et al.[313]).

				Beispiel	
Substitution	Doppelbindungen	Bindungen	Ladung	Donor	Akzeptor
Protonierungsartig	0	±	±	O200	O100-
Resonanzartig	±	0	±	O100-	O010
Tautomerartig	±	±	0	O200	O010

den: Eine *protonierungsartige Substitution* ändert die Anzahl der Bindungen und die Ladung des Valenzzustands. Eine *resonanzartige Substitution* ändert die Anzahl der Doppelbindungen und die Ladung des Valenzzustands. Eine *tautomerartige Substitution* ändert die Anzahl der Doppelbindungen und die Anzahl der Bindungen des Valenzzustands. Die am Austausch beteiligten Valenzzustände werden als *Donoren* (höhere Anzahl an Einfachbindungen) und *Akzeptoren* (niedere Anzahl an Einfachbindungen) bezeichnet. Unterschiedliche Formen eines Moleküls können durch Mehrfachsubstitutionen von Valenzzuständen in Valenzzustandsfolgen herbeigeführt werden (vgl. Tabelle 4.3). Die Anzahl der Substitutionen wird durch $\Delta(D \rightarrow A)$ bzw. $\Delta(A \rightarrow D)$ bezeichnet. Sie gibt an, wie häufig ein Donor durch einen Akzeptor bzw. ein Akzeptor durch einen Donor des entsprechenden Typs ersetzt wird, um unterschiedliche Arten von Molekülformen zu erhalten. Kekulestrukturen unterscheiden sich nur in der Verteilung der Bindungsordnung, nicht aber in der Valenzzustandsfolge. Unterschiedliche Ionisierungs- bzw. Protonierungszustände werden mit protonierungsartigen Substitutionen erzeugt.

Sie unterscheiden sich in der Anzahl der Donor- und Akzeptorsubstitutionen. Ist die Anzahl gleich, werden Protonen innerhalb des Moleküls transferiert. Die Gesamtladung bleibt jedoch unverändert. Entspricht die Anzahl der Substitutionen von Donoren nicht der von Akzeptoren, so wird die Gesamtladung des Moleküls geändert und ein anderer Ionisierungszustand erhalten. Tautomere und Mesomere erfahren nur tautomer- bzw. resonanzartige Substitutionen. Die Anzahl an Donor- und Akzeptorsubstitutionen bleibt hierbei erhalten. Andere Substitutionen erzeugen unterschiedliche Redoxformen.

Tabelle 4.3: Molekülformen durch Valenzzustandssubstitutionen (Urbaczek et al.[313]).

Molekülform	Valenzzustandssubstitution	Bedingung
Kekule	keine	keine
Ionisierung	Protonierung	$\Delta(D \to A) \neq \Delta(A \to D)$
Protonierung	Protonierung	$\Delta(D \to A) = \Delta(A \to D)$
Mesomer	Resonanz	$\Delta(D \to A) = \Delta(A \to D)$
Tautomer	Tautomer	$\Delta(D \to A) = \Delta(A \to D)$
Redox	Resonanz	$\Delta(D \to A) \neq \Delta(A \to D)$
	Tautomer	$\Delta(D \to A) \neq \Delta(A \to D)$

4.2.2 Erzeugung molekularer Zustände

Substitutionen in Valenzzustandsfolgen erlauben es, intern Molekülzustände zu enumerieren. Die Menge der Zustände wächst jedoch exponentiell mit der Größe eines Moleküls. Zudem sind nicht alle Zustände chemisch stabil und praktisch nie in Protein-Ligand-Komplexen zu finden. Eine simple Enumeration ist daher aus informatischer und chemischer Sichtweise ungünstig. Naomi begegnet diesem Umstand, indem es das Problem der Zustandsenumeration in einem „teile-und-herrsche“-Ansatz auf möglichst kleine, unabhängige Subprobleme reduziert, diese einzeln löst und die Teillösungen zu vollständigen Lösungen rekombiniert. Die Lösungen werden dabei bewertet, um ausschließlich chemisch sinnvolle Zustände zu erzeugen. Die Generierung von molekularen Zuständen gliedert sich dementsprechend in die Partitionierung des Moleküls in Zonen (*Multizustandszonen*), die Selektion alternativer Valenzzustände in Multizustandszonen, die Generierung valider Teilzustandsfolgen, die Bewertung valider Teilzustandsfolgen und die Rekombination gut bewerteter Teilzustandsfolgen zu vollständigen Lösungen.

Definition 4.2.1 (Potentiell konjugiertes System)**.** *Ein potentiell konjugiertes System besteht aus zusammenhängenden Atomen, die keine sp^3-Hybridisierung aufweisen, d. h. keiner ihrer Atomtypen besitzt eine tetrahedrale Geometrie.*

Definition 4.2.2 (Funktionelle Gruppe)**.** *Eine funktionelle Gruppe besteht aus zusammenhängenden Atomen, die kein Wasserstoff, kein aliphatischer Kohlenstoff sind und keine Kohlenstoff-Kohlenstoff-Mehrfachbindung enthalten. Atome konjugierter Ringsysteme sind ebenso nicht Teil funktioneller Gruppen, außer wenn das Ringatom über eine exozyklische Doppelbindung an ein Heteroatom gebunden ist oder selbst ein Heteroatom ist, das über eine exozyklische Bindung an ein weiteres Heteroatom gebunden ist.*

Definition 4.2.3 (Alternativer Valenzzustand)**.** *Ein Valenzzustand v eines Atoms a ist ein alternativer Valenzzustand v', wenn die Anzahl der Schweratomnachbarn von a nicht größer als die Anzahl der Bindungen $\sum_{i=1}^{3} b'_i$ in v' ist, d. h. wenn die Topologie von a erhalten bleibt.*

Identifizierung von Multizustandszonen: Die Aufteilung in unabhängige Teilprobleme muss gewährleisten, dass bei Rekombination weder valide Molekülzustände ausgelassen, noch unnötige oder invalide zusammengesetzt werden. Aus chemischer Sicht sind Molekülzustände prinzipiell Zustandsänderungen einzelner funktioneller Gruppen und konjugierter Molekülbereiche. Protonierungszustände und Tautomere unterscheiden sich nur durch Variationen innerhalb dieser Zonen. Um unabhängige Teilprobleme zu generieren, muss die Partitionierung demnach abgeschlossene, konjugierte Systeme und funktionelle Gruppen innerhalb des Moleküls identifizieren. Die Heuristik von NAOMI deduziert diese Zonen ausgehend von der Gesamtstruktur durch sukzessiven Ausschluss irrelevanter Atome und Bindungen nach den Definitionen 4.2.1 und 4.2.2. Ihr Ziel ist es, Atome und Bindungen auszuschließen, die nie durch Zustandsänderungen betroffen sind und über Zustände hinweg konstant bleiben. Die Ableitung der Zonen ist eindeutig, da sie nur auf der Auswertung von Atomtypeigenschaften beruht und unabhängig von der initial gegebenen Valenzstruktur des Eingabemoleküls ist.

Selektion alternativer Valenzzustände: Innerhalb einer Multizustandszone können unabhängig valide Zustandsfolgen generiert werden. Hierfür werden Valenzzustände protonierungs-, resonanz- oder tautomerartig substituiert (vgl. Tabelle 4.2). Für jeden Valenzzustand und jede Substitutionsart sind dafür vordefinierte, mögliche Zustände vorgehalten. Sie können einem Atom zugewiesen werden, wenn nach Definition 4.2.3 die Topologie des Atoms erhalten bleibt. Um molekulare Formen eines bestimmten Typs zu erhalten, müssen die Substitutionen zudem die in Tabelle 4.3 gestellten Bedingungen erfüllen. Dieser generische Ansatz zur Auswahl eines neuen Valenzzustands kann insbesondere bei funktionellen Gruppen zu einer beträchtlichen Menge alternativer Zustände führen. Für sie wird daher ein heuristischer Ansatz zur Selektion alternativer Valenzzustände genutzt: Für funktionelle Gruppe wird geprüft, ob sie in einer vordefinierten

Liste veränderlicher Gruppen enthalten ist. Ist dies der Fall, werden vordefinierte Valenzzustände selektiert. Ist dies nicht der Fall, so ist die Gruppe eine zusammengesetzte funktionelle Gruppe (z. B. Triphosphat), die in Subgruppen unterteilt wird. Sie werden individuell überprüft und bekommen dann alternative Valenzzustände zugewiesen. Je nach Art der geforderten Substitution sind spezifische alternative Valenzzustände definiert. So besitzen beispielsweise symmetrische funktionelle Gruppen keine alternativen Tautomerzustände. Sie sind also bezüglich dieser Substitution invariant, da eine derartige Substitution zu äquivalenten Molekülrepräsentationen führen würde.

Enumeration valider Valenzzustandsfolgen: Die Enumeration valider Valenzzustandsfolgen ist ein Backtracking-Verfahren, das sukzessiv Teilzustandsfolgen zu einer vollständigen Folge zusammensetzt und Teillösungen verwirft, sobald erkennbar ist, dass deren Kombination zu einem invaliden Molekülzustand führt. Die Atome von Multizustandszonen werden hierfür in einer definierten Ordnung prozessiert. Zunächst werden Zustände terminaler Atome, dann benachbarter Atome und zuletzt der restlichen Atome variiert. Atome außerhalb der Bereiche bleiben durch ihren einzigen Valenzzustand repräsentiert. Im zugrundeliegenden Suchbaum entsprechen einzelne Knoten einer Zuweisung eines Valenzzustands, innere Knoten repräsentieren Teilzustandsfolgen und Blätter stellen vollständig generierte Zustandsfolgen dar. Ausgehend von der Wurzel werden die möglichen Zustände eines Atoms betrachtet. Eine Substitution und der dadurch aufgespannte Teilbaum werden verworfen, wenn die erhaltene Valenzzustandsfolge chemisch invalide ist. Hierfür finden folgende Überprüfungen statt:

- Bei terminalen Atomen ist die Zuweisung eines Valenzzustands äquivalent mit der Zuweisung einer neuen Bindungsordnung. Die Zuweisung ist kompatibel, wenn die Anzahl der Bindungen die Anzahl benachbarter Atome nicht übersteigt.
- Bei einem Blatt wird geprüft, ob eine valide Verteilung der Bindungsordnungen erreicht werden kann. Dies trifft nicht zu, wenn eine ungerade Anzahl von Valenzzuständen mit ungerader Anzahl an Mehrfachbindungen festgestellt wird.
- Der Oxidationszustand wird durch die Anzahl der Donoren bestimmt. Unterscheidet sich dieser vom initialen Zustand, so kann die Lösung verworfen werden.

Bei Erfolg werden die Bindungsordnungen rekursiv zugewiesen und man erhält eine vollständige Valenzzustandsfolge, die einen validen Molekülzustand repräsentiert.

Bewertung valider Valenzzustandsfolgen: Die Valenzzustandsfolgen sind chemisch valide, können jedoch instabile Tautomere, unwahrscheinliche Protonierungszustände und unangemessene Resonanzformen darstellen. Um solche Zustände ausschließen zu können, werden die Valenzzustandsfolgen bewertet. Die Bewertung S_{VSC} einer Folge

ergibt sich hierbei aus der Summe der Einzelbewertungen von Teilzustandsfolgen:

$$S_{\mathrm{VSC}} = \sum S_{\mathrm{ring}} + \sum S_{\mathrm{subst}} + \sum S_{\mathrm{group}} \tag{4.1}$$

S_{ring} bewertet die Valenzzustandsfolge eines konjugierten Rings, S_{subst} die Valenzzustandsfolge eines Ringsubstituenten und S_{group} die Valenzzustandsfolge einer funktionellen Gruppe. Letztere wird gegebenenfalls weiter in Subgruppen unterteilt, die dann individuell bewertet werden. Für jede dieser Komponenten erhält ein bevorzugter Referenzzustand die bestmögliche Bewertung von 100. Im Fall konjugierter Ringe ist er eine Teilzustandsfolge, die nicht zu exozyklischen Doppelbindungen führt. Ringsubstituenten erhalten die bestmögliche Bewertung, wenn die zugehörige Substruktur lediglich eine exozyklische Einfachbindung aufweist. Funktionelle Gruppen bzw. Subgruppen sind vorzugsweise neutral und entsprechen der stabilsten tautomeren Form. Vom Referenzzustand abweichende alternative Teilzustandsfolgen werden entweder gleich oder schlechter bewertet. Die Referenzstrukturen und deren alternative Valenzzustandsfolgen sind einschließlich zugehöriger Bewertungen in einer Datenbank hinterlegt. Sie enthält insgesamt 252 Einträge (113 Ringe, 121 Subgruppe, 18 Substituenten). Die Bewertungen wurden anhand von Zustandspaaren abgeleitet, für die bevorzugte molekulare Formen aus Experimenten und theoretischen Berechnungen bekannt sind. Für seltene Substrukturen sind keine Bewertungsdaten hinterlegt. Sie werden mit einem generischen Ansatz bewertet. Die Referenz eines konjugierten Rings, Ringsubstituent oder einer funktionellen Gruppe erhält *per se* eine Bewertung von 80 und eine Valenzzustandssubstitution wird mit P bestraft, wenn sie zu einer ungünstigen Änderung führt (vgl. Tabelle 4.4):

$$S_{\mathrm{generic}} = \max(0, 80 - \sum P) \tag{4.2}$$

Tabelle 4.4: Generische Bestrafungsterme P bei Substitutionen in konjugierten Ringen, Ringsubstituenten und funktionellen Gruppen (nach Urbaczek et al.[313]).

Substruktur	Substitution führt zu	P
Ring	nicht aromatischen Ring	20
Ring	Einzelladung im Ring	20
Ring	Mehrfachladung im Ring und Substituenten	80
Ring	Drei aufeinanderfolgende Donoren im Ring	80
Substituent	Substituent mit exozyklischer Doppelbindung	20
Substituent	Einzelladung in Substituent	20
Substituent	Mehrfachladung in Substituent	80
Funktionelle Gruppe	Mehrfach positive Ladung in Gruppe	80

4.3 Protoss: Wasserstoffbrückennetzwerkoptimierung

Protoss[315] ist ein Werkzeug zur Vorhersage der wahrscheinlichsten Lage von Wasserstoffatomen in einem Protein-Ligand-Komplex. Es versucht uneindeutige Wasserstoffpositionen zu bestimmen, indem es sie innerhalb eines Wasserstoffbrückennetzwerk betrachtet und dieses optimiert. In cRAISE ist die Methode Teil der Proteinpräparierung und wird zur Optimierung der Wasserstoffe von Posen im Screening eingesetzt. Die Naomi-basierte Implementierung greift auf das in Abschnitt 4.1 vorgestellte Naomi-Modell und die in Abschnitt 4.2 vorgestellte Repräsentation molekularer Zustände zurück. Algorithmisch basiert es auf seiner Flex*-basierten Vorgängerversion[297].

4.3.1 Initiale Wasserstoffpositionen

Bei Initialisierung eines Proteins bzw. eines Protein-Ligand-Komplexes gemäß der Naomi-Routine (vgl. Abschnitt 4.1.4) werden fehlende Wasserstoffpositionen hinzugefügt. Die Proteininitialisierung weist einem Atom mit dem Atomtyp auch eine Idealgeometrie gemäß der VSEPR-Theorie zu. Mit dem Wissen über den dadurch reflektierten Hybridisierungszustand des Atoms und dessen kovalent gebundenen Schweratomnachbarn werden fehlende Wasserstoffe identifiziert und deren Positionen bestimmt. So induziert beispielsweise die trigonal-planare Geometrie eines sp^2-hybridisierten Kohlenstoffatoms mit zwei Schweratomnachbarn ein Wasserstoffatom, das in Richtung der unbelegten Ecke des hierbei aufgespannten regelmäßigen Dreiecks zu finden ist. Es gibt allerdings Fälle bei denen die Postion eines Wasserstoffes anhand der Atomgeometrie und der lokalen Schweratomumgebung nicht eindeutig bestimmbar ist. Dies betrifft vor allem die Wasserstoffe terminaler Schweratome (Alkohole, primäre Amine), aber auch sekundäre Amine in zyklischen oder azyklischen Strukturen. Bei Wassermolekülen ist die relative Anordnung der Wasserstoffpositionen durch die Initialisierungsroutine zwar eindeutig, die Orientierung des Moleküls ist jedoch zufällig. Existieren für Protein und Ligand verschiedene Tautomere und Protonierungszustände, so weist die Initialisierung lediglich einen wahrscheinlichen Grundzustand und die entsprechende Lage der Wasserstoffe zu.

4.3.2 Enumeration alternativer Wasserstoffpositionen

Auf Basis initialer Wasserstoffpositionen und des Grundzustands identifiziert Protoss Substrukturen in Protein und Ligand mit variablen Wasserstoffpositionen. Dies sind terminale funktionelle Gruppen, die eine Rotation des terminalen Wasserstoffs ermöglichen, sowie Wassermoleküle, Proteinresiduen mit Amid- und Imidazolgruppen, die

Aminosäure-Flips ermöglichen und Substrukturen, für die ein alternativer Protonierungszustand oder ein alternatives Tautomer existiert. Im Folgenden werden sie zusammenfassend als *variable Regionen* und ihre Freiheitsgrade als *Modi* bezeichnet. Jeder enumerierte Modus spiegelt eine mögliche Ausrichtung der Wasserstoffe einer variablen Region wider. Wasserstoffe terminaler funktioneller Gruppen sind durch 12 Modi abgebildet, die ihre freie Rotation um die Hauptachse der terminalen Schweratombindung diskretisieren. Die Orientierung der Wasserstoffatome von Wassermolekülen wird über 60 gleichmäßig im Raum verteilte Orientierungen der tetrahedralen Geometrie des Sauerstoffatoms repräsentiert. Residuen, die Aminosäure-Flips ermöglichen, sind in Röntgenkristallstrukturen nicht eindeutig beschrieben. Es besteht die Möglichkeit, dass ihre funktionelle Gruppe inklusive Wasserstoffe tatsächlich um 180° gedreht vorliegt. Sie werden durch zwei korrespondierende Modi beschrieben. Die Modi von Gruppen mit Protonierungszuständen oder Tautomeren entsprechen möglichen validen Valenzzustandsfolgen, der in Abschnitt 4.2 vorgestellten assoziierten Multizustandszonen. Sie sind entsprechend ihrer Bewertung vom wahrscheinlichsten zum unwahrscheinlichsten Zustand sortiert.

4.3.3 Bewertung des Wasserstoffbrückennetzwerks

Um das beste Wasserstoffbrückennetzwerk zu identifizieren muss PROTOSS für eine gewählte Menge von Modi intra- und intermolekulare polare Interaktionen erkennen und bewerten. Hierfür stattet es jeden Modus mit potentiellen Interaktionsoberflächen aus, die mit der Lage eines Wasserstoffatoms (Donoroberfläche) und freien Elektronenpaaren (Akzeptoroberfläche) assoziiert sind. PROTOSS registriert einzelne Wasserstoffbrücken und Metallinteraktionen, indem es die Lage zweier komplementärer Interaktionsoberflächen geometrisch bewertet. Die Bewertung berücksichtigt im Wesentlichen die Distanz der Schweratome und die Abweichung des Wasserstoffs vom freien Elektronenpaar. Zudem werden Donor-Donor, Donor-Metal und Akzeptor-Akzeptor-Kontakte identifiziert. Sie liefern einen destabilisierenden, repulsiven Beitrag. Statische funktionelle Gruppen (ein Modus) sind selbst nicht Teil der Optimierung. Sie sind jedoch ebenso mit potentiellen Interaktionsflächen ausgestattet, da ihre Interaktionen mit einem Modus zur Güte des Wasserstoffbrückennetzwerkes beitragen. Für eine gewählte Menge M von Modi ergibt sich die Gesamtbewertung des Wasserstoffbrückennetzwerks $S_{\text{total}}(M)$ durch:

$$S_{\text{total}}(M) = \sum_{m \in M} S_{\text{base}}(m) + \sum_{m,n \in M} (S_{\text{interactions}}(m,n) + S_{\text{repulsion}}(m,n)) \tag{4.3}$$

$S_{\text{base}}(m)$ repräsentiert hierbei die Basisbewertung eines Modus m, die Wasserstoffbrücken zu allen statischen Gruppen bewertet und bei einem Protonierungszustand oder

Tautomer dessen relative Stabilität gemäß Gleichung 4.2 angibt. $S_{\text{interactions}}(m, n)$ bewertet die polaren Interaktionen und $S_{\text{repulsion}}(m, n)$ die repulsiven Wasserstoffkontakte zwischen den Modi zweier variablen Regionen in der gewählten Menge.

4.3.4 Wasserstoffbrückennetzwerkoptimierung

Protoss optimiert das Wasserstoffbrückennetzwerk, indem es die Menge M der Modi bestimmt, die die Gesamtbewertung $S_{\text{total}}(M)$ minimiert. Abhängigkeiten zwischen den Modi unterschiedlicher variabler Regionen werden durch einen Graphen modelliert:

- Jede variable Region wird durch einen Knoten repräsentiert und ihre Modi werden inklusive ihrer Basisbewertung an den Knoten annotiert.
- Zwei Knoten werden über eine Kante verbunden, wenn es für mindestens eine Kombination ihrer Modi möglich ist, eine Interaktion zu bilden.
- Für jede Kante des Graphen wird eine Matrix vorgehalten, die die paarweisen Bewertungen zwischen den Modi der variablen Regionen enthält.
- Der Graph wird auf einen bzw. mehrere Bäume reduziert: Der Knoten eines zusammenhängenden Subgraphen mit höchstem Knotengrad wird als Wurzel ausgezeichnet. Zweifach-Zusammenhangskomponenten werden auf einen Knoten reduziert, der Referenzen und Kanten der ersetzten Knoten erhält.

Auf dem Graph wird das Wasserstoffbrückennetzwerk mittels eines dynamischen Programmieransatzes optimiert. Hierbei können die einzelnen Bäume unabhängig voneinander prozessiert werden, da sich zwischen ihren variablen Regionen keine Interaktionen etablieren. Bäume, die nur aus einer Wurzel bestehen, werden direkt ausgewertet. Für sie muss lediglich die Basisbewertung $S_{\text{base}}(m)$ für jeden Modus m bestimmt und der Modus mit minimaler Bewertung selektiert werden. Die anderen Bäume werden in einer *post-order* Reihenfolge traversiert. Für jeden besuchten Knoten wird die optimale Lösung des zugrundeliegenden Teilbaums berechnet und gespeichert. Dafür werden für jeden Modus m des aktuell besuchten Knotens zwei Werte nach Gleichung 4.3 bestimmt: Die Basisbewertung und die Summe über die besten Bewertungen seiner Kinder. Beide Werte werden addiert und als beste Bewertung des Modus m am gegenwärtigen Knoten annotiert. Die post-order Traversierung gewährleistet, dass die optimale Bewertung der Teilbäume der Kinder bereits berechnet wurde, sobald der Elternknoten besucht wird. Werden mehre optimale Modi identifiziert, so wird der erste Modus erhalten. Dies geschieht im Fall von Protonierungszuständen, Tautomeren und Flips. Bei Modi frei rotierbarer Wasserstoffe wird die Torsionsmedian aller optimaler Modi gewählt. Zyklische

Abhängigkeiten sind durch reduzierte Knoten erkennbar. Bei ihnen kann der dynamische Programmieransatz nicht fortgesetzt werden. Deshalb wird der referenzierte Knoten mit den wenigsten Modi entfernt. Dies wird solange fortgesetzt bis der Zyklus aufgelöst ist. Die dann azyklische Komponente kann im dynamischen Programmieransatz weiter bearbeitet werden. Für die entfernten Knoten werden alle Kombinationen ausgewertet, indem der Baum für jeden Modus erneut traversiert wird. Abschließend werden die Wasserstoffkoordinaten der variablen Regionen gesetzt, indem die Wasserstoffpositionen der optimalen Modi auf den Protein-Ligand-Komplex übertragen werden.

4.4 CONFECT: Konformergenerierung

Im Allgemeinen führt cRAISE ein starres Docking durch und bildet keine Molekülflexibilität ab. Dennoch kann es als flexible Docking-Methode fungieren, indem die bereitgestellte Molekülbibliothek mit Konformationen angereichert wird. cRAISE integriert hierzu CONFECT[254], eine wissensbasierte Methode, die mittels häufig vorkommender Torsionswinkel kleiner Moleküle[255] den eigentlich kontinuierlichen Konformationsraum auf die wesentlichen Konformere reduziert und generiert. Die Konformergenerierung mit CONFECT ist ähnlich zum Flex*-basierten TrixX Conformer Generator (TCG)[253], der in TRIXX-BMI zur Anwendung kam. Allerdings verwendet CONFECT die folgend beschriebene *Torsionsbibliothek* und prozessiert intern NAOMI-Moleküle.

4.4.1 Die Torsionsbibliothek

Die Torsionsbibliothek registriert und kategorisiert die Verteilung von Torsionswinkeln kleiner Moleküle in einem gegebenen Datensatz. Sie stellt eine Sammlung sogenannter *Torsionssignaturen* dar. Eine Torsionssignatur umfasst ein *Torsionsmuster*, das die Umgebung einer rotierbaren Bindung beschreibt. Es ist ein SMARTS-Ausdruck[327], der mindestens die vier Atome, über die sich ein Torsionswinkel definieren lässt, kodiert und optional die nähere Umgebung der rotierbaren Bindung spezifiziert. Zudem umfasst eine Torsionssignatur eine *Liste typischer Torsionswinkel* und die gesamte Torsionsverteilung in Form eines *Torsionshistogramms*. Ein Beispiel für ein Torsionshistogramm ist in Abbildung 4.2 dargestellt. Es speichert die absoluten Häufigkeiten im Datensatz eingenommener Torsionswinkel der Bindungsart, die durch das Torsionsmuster definiert ist. Relative Häufigkeiten können anhand der absoluten Häufigkeiten jederzeit abgeleitet werden. Das Histogramm unterteilt den Torsionsraum einer Bindung gleichmäßig in Intervalle von jeweils 10°. Ein Histogramm-Peak wird in einem Intervall verzeichnet,

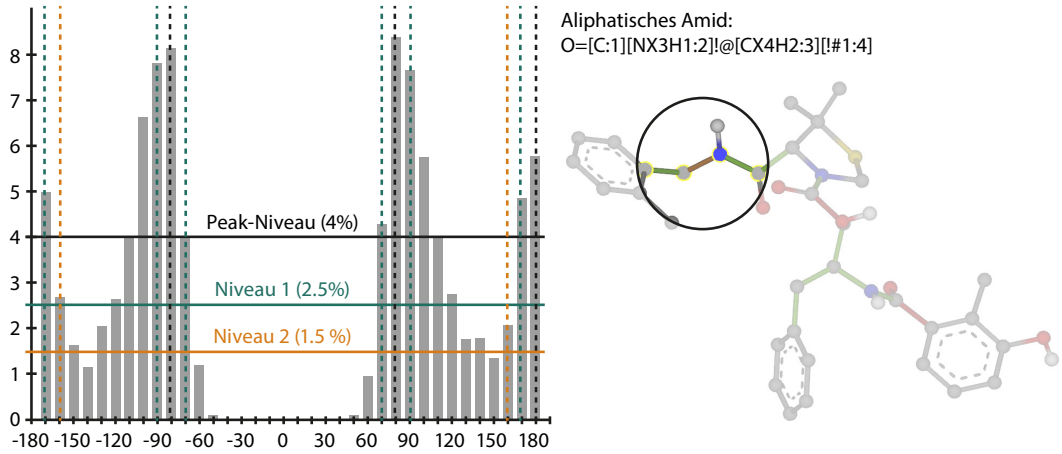


Abbildung 4.2: Torsionshistogramm einer rotierbaren Bindung: Die Bindung fällt in die Kategorie aliphatischer Amide. Der SMARTS-Ausdruck kodiert die vier Atome um die Bindung. Typische Torsionswinkel sind bei -80°, 80° und 180° in der Torsionsdatenbank registriert. Seltener treten Abweichungen von den optimalen Torsionswinkel bei den ersten und zweiten Toleranzniveaus auf. Die Abweichungen des ersten und zweiten Toleranzniveaus fallen für die Peaks bei -80° und 80° zusammen.

das mehr als 4 % aller beobachteten Torsionswinkel der Signatur umfasst. Sein Optimum spiegelt einen typischen Torsionswinkel wider. Zusätzlich sind für jeden Peak zwei Toleranzniveaus bestimmt, deren untere bzw. obere Grenzen zwei erweiterte Torsionsbereiche definieren. Das erste Toleranzniveau umfasst gerade 2,5 %, das zweite Toleranzniveau gerade 1,5 % der Datenpunkte. Die Toleranzbereiche beschreiben die Breite eines Peaks und somit relativ vom idealen Torsionswinkel häufig und selten eingenommene Abweichungen. Mit vordefinierten Torsionsmustern können Histogramme eines gegebenen Moleküldatensatz abgeleitet und in der Torsionsbibliothek verwaltet werden. Die Konformationen, die in dieser Arbeit verwendet wurden, basieren auf einer Torsionsbibliothek[255], deren Torsionswinkelverteilungen aus der Cambridge Structural Database[328] abgeleitet wurden.

4.4.2 Konformergenerierung

Mit Hilfe einer Torsionsbibliothek kann jeder rotierbaren Bindung eindeutig eine Torsionssignatur zugeordnet und für sie typische Torsionswinkel entnommen werden. Confect generiert Konformationen indem es diese Winkel sukzessiv im Molekül einstellt:

Komponentenzerlegung: Ein kanonisiertes Molekül wird an azyklischen Einfachbindungen zwischen Schweratomen geschnitten und in Komponenten zerlegt. Einfachbin-

dungen zu terminalen Methyl-, Triflouromethyl- oder Nitril-Gruppen sind davon ausgeschlossen, da eine Rotation um solche Achsen zu redundanten Konformationen führt. Die Zerlegung des Moleküls in Komponenten ist im Komponentenbaum hinterlegt. Er definiert, an welchen Bindungen Molekülteile rotiert werden können. Seine Knoten entsprechen den Komponenten, seine Kanten den rotierbaren Bindungen zwischen den Komponenten. Die Wurzel des Baums wird mit dem Floyd-Warshall-Algorithmus bestimmt, der die zentrale Komponente des Moleküls identifiziert. Die Kinder eines Knotens sind entsprechend der kanonisierten Bindungen des Moleküls sortiert. Jeder Kante im Baum wird eine Torsionssignatur aus der Torsionsbibliothek zugeordnet. Jedem ihrer typischsten Torsionswinkel wird ein Wert zugewiesen, der der relativen Häufigkeit des Winkels bezüglich aller beobachteten Torsionen der Signatur entspricht. Gegebenenfalls erhalten Torsionswinkel des ersten und zweiten Toleranzniveaus einen anteiligen Wert.

Generierung von Konformationen: An der Wurzel beginnend wird auf dem Komponentenbaum eine Breitensuche durchgeführt. Sie prozessiert die Komponenten gemäß der Kanonisierung. Bei Besuch eines Knotens werden Teilkonformationen generiert, indem alle an der eingehenden Kante annotierten Torsionswinkel eingestellt werden. Dies erweitert den Komponentenbaum zu einem Konformationsbaum, d. h. jeder Winkel erzeugt einen zusätzlichen Knoten. Vater und Kinder eines neuen Knotens entsprechen Teilkonformationen der vorherigen bzw. nachfolgend prozessierten Komponente. Die erzeugten Knoten werden in eine Prioritätswarteschlange einsortiert. Die Sortierreihenfolge ist durch die Summe der relativen Torsionshäufigkeiten der Gesamtkonformation bestimmt. Für bereits eingestellte Torsionen wird die Summe exakt ermittelt, für noch nicht eingestellte wird sie mit Werten der bestmöglichen Torsionen willentlich überschätzt. Die Prioritätswarteschlange bestimmt die beste Teilkonformation, für die die nächste Komponente erweitert und bearbeitet wird. An einem Blatt angekommen ist eine vollständige Konformation erzeugt. Die Prioritätswarteschlange wird abgearbeitet bis alle Konformationen generiert sind oder ein Abbruchkriterium erreicht wird.

Beschränkung der Konformationen: Die maximale Anzahl generierter Konformationen ist von CONFECT beschränkt. Ist diese Anzahl erreicht, wird folgend kein weiterer Komponentenknoten erweitert. Da immer zuerst die Teillösung erweitert wird, die optimal bezüglich der Torsionsbewertung ist, ist eine Bestensuche realisiert. Sie garantiert, dass die statistisch häufigsten Konformationen bei Abbruch bereits erzeugt wurden.

Da die Komponentenzerlegung nur an azyklischen Bindungen erfolgt, können Komponenten flexible Ringstrukturen enthalten. Flexible Ringe mit bis zu neun Ringatomen werden von CONFECT erkannt. Die zugehörigen Ringkomponenten werden dann, in Kombination zu den Torsionen der Eingangsbindung, mit einmalig vorberechneten

Ringkonformationen erweitert. Makrozyklen mit mehr als neun Schweratomen behalten die gegebene Initialkonformation bei. Generell kann das Einstellen von Torsionen zu invaliden Konformationen mit intramolekularen Atomkollisionen führen. Daher wird während der Traversierung durch den Konformationsbaum überprüft, ob eine Torsion zu Überlappungen von Atomen führt. Im Fall einer starken Kollision wird die Teilkonformation verworfen und nicht weiterverfolgt. Konformationen mit leichten Kollisionen können optional mit einem Kraftfeld optimiert und behoben werden.

4.4.3 Qualitätsstufen

Die Größe des durchmusterten Konformationsraums $K_1(m)$ eines Moleküls m ist durch die Anzahl der Torsions-Peaks $|P(t_i)|$ beschränkt, wobei $t_i \in T$ ein Torsionsmuster einer flexiblen Bindungen ist und i über alle rotierbaren Bindungen des Moleküls iteriert:

$$|K_1(m)| = \prod_{i=1}^{|T|} |P(t_i)| \tag{4.4}$$

Zuzüglich Toleranzwinkel des ersten Toleranzniveaus erhöht sich die Zahl auf maximal

$$|K_2(m)| = 3^{|T|} \cdot \prod_{i=1}^{|T|} |P(t_i)| \tag{4.5}$$

und zuzüglich der Toleranzwinkeln des zweiten Toleranzniveaus auf maximal

$$|K_3(m)| = 5^{|T|} \prod_{i=1}^{|T|} |P(t_i)| \tag{4.6}$$

Konformationen. Enthält ein Molekül flexible Ringe, so multipliziert sich die Zahl mit der Anzahl vorab berechneter Ringkonformationen. Für große und flexible Moleküle mit frei rotierbaren Bindungen, die durch Histogramme mit $|P(t)| = 12$ charakterisiert sind, ist die theoretisch mögliche Anzahl an Konformationen sehr hoch. Auch wenn kollidierende Konformationen den Raum reduzieren, lassen sich die verbleibenden Konformationen oftmals nicht mehr effizient prozessieren. Deshalb reduziert Confect den Konformationsraum automatisch in Fällen, in denen $|K_1(m)|$ 100 000 Konformationen übersteigt. Es nähert dann eine freie Rotation durch sechs gleichmäßig verteilte Torsionswinkel an. Übersteigt $|K_1(m)|$ die Grenze von 1 000 000 Konformationen, werden Rotationen auf drei gleichmäßig verteilte Torsionswinkel reduziert. Zudem bietet Confect mehrere Qualitätsstufen an, über die sich die Größe und Abdeckung des Konformationsraumes benutzterdefiniert steuern lassen (vgl. Tabelle 4.5). Sie legen fest, welche

Winkelabweichungen zusätzlich durchmustert werden und limitieren die Anzahl erzeugter Konformationen, die in der Lösungswarteschlange vorgehalten werden. Die Lösungen sind entsprechend relativer Torsionshäufigkeiten sortiert. Confect bietet optional an, sie auf Basis der Torsion Fingerprint Deviation (TFD)[329] oder der Root Mean Square Deviation (RMSD) zu clustern, um die Anzahl generierter Konformationen weiter zu reduzieren und lediglich repräsentative Lösungen anzubieten.

Tabelle 4.5: Confect-Qualitätsstufen: QS 1 durchmustert den Konformationsraum gemäß der Peaks P der Histogramme. QS 2 und QS 3 erweitern den Raum durch Winkelabweichungen δ_1 und δ_2, der ersten und zweiten Toleranzniveaus. QS 21, QS 4, QS 41 und QS 5 stellen zusätzlich halbe Winkelabweichungen ein. Small und Extra Small reduzieren, intern den Raum indem freie Rotationen durch sechs bzw. drei gleichmäßig verteilte Torsionswinkel angenähert wird (nach Schärfer et al.[254]).

Qualitätsstufe	Torsionswinkel ($p \in P$)	Rotation	Limit
QS 0	p	12	1
QS 1	p	12	250
QS 2	p, $p \pm \delta_1$	12	500
QS 21	p, $p \pm \delta_2$	12	500
QS 3	p, $p \pm \delta_1$, $p \pm \delta_2$	12	1000
QS 4	p, $p \pm \delta_1$, $p \pm \frac{\delta_1}{2}$	12	1000
QS 41	p, $p \pm \delta_2$, $p \pm (\delta_1 + \frac{\delta_2 - \delta_1}{2})$	12	1000
QS 5	p, $p \pm \delta_1$, $p \pm \delta_2$, $p \pm \frac{\delta_1}{2}$, $p \pm (\delta_1 + \frac{\delta_2 - \delta_1}{2})$	12	2000
Small	p	6	100 000
Extra Small	p	3	1 000 000

4.5 Mona: Mengenoperationen auf Molekül datenbanken

Mona[314] ist eine Entwicklung, um Mengen von Molekül daten interaktiv zu präparieren und zu visualisieren. Ihre Intension, dabei mit großen Datensätzen zu arbeiten, stellt ähnliche Anforderungen wie die Datenverwaltung im VS. Mona verwaltet Molekülinformation, die weit über die topologische Beschreibung hinausgeht. Dies erlaubt den Erhalt dreidimensionaler Daten. cRAISE nutzt die zugrundeliegende Molekül datenbank zur Speicherung und Verwaltung der Molekülbibliothek und erweitert sie zur Verwaltung seiner Screening-Resultate. Einige Funktionalitäten von Mona finden bei der Duplikaterkennung und der Filterung von Molekülen Anwendung.

4.5.1 Mona

Mona unterstützt Aufgaben wie die Analyse und Filterung von Moleküldaten anhand ihrer physikochemischen Eigenschaften, das Identifizieren von Duplikaten oder die Substruktursuche. Dafür verwaltet es gegebene Moleküldaten in einer relationalen SQL-Datenbank, auf der alle Operationen arbeiten. Jeder Verwendung von Mona geht initial der Import der als SMILES, SDF oder MOL2 bereitgestellten Moleküldaten in die Moleküldatenbank voraus. Während des Imports wird eine eindeutige topologische Beschreibung generiert, die es erlaubt, Duplikate in den gegebenen Daten zu erkennen. Typische chemieinformatische Aufgaben können interaktiv durch klassische mathematische Mengenoperation realisiert werden. Dafür kreiert Mona Molekülmengen, auf denen die Operationen ausgeführt werden. Die Resultate einzelner Operationen werden u. a. durch zweidimensionale Molekülstrukturbilder und Histogramme über die Moleküleigenschaften visualisiert. Auch Mona arbeitet auf dem zugrundeliegenden Naomi-Model, um konsistent Moleküle formatunabhängig handhaben zu können.

4.5.2 Die Molekülzeichenkette

Mona repräsentiert Moleküle durch eine Molekülzeichenkette, den *kanonisierten Mol-String*. Er erfüllt zwei Eigenschaften, die die Erkennung von Duplikaten und den raschen und konsistenten Aufbau interner Molekülrepräsentationen gewährleisten:

- Der kanonisierte Mol-String beschreibt eindeutig die Topologie eines Moleküls.
- Der Mol-String enthält alle Daten, um ein valides Naomi-Molekül zu erzeugen.

Die Eindeutigkeit des Mol-Strings wird gewährleistet, indem importierte Moleküle kanonisiert und die Atome und Bindungen definiert sortiert werden.[313] Für ein kanonisiertes Molekül wird ein Mol-String erstellt, der ein Naomi-Molekül serialisiert. Er kodiert Eigenschaften, die notwendig sind, um erneut eine vollständige Repräsentation aufzubauen. Dafür umfasst der Mol-String eine Folge von Valenzzuständen und die zugehörige Folge von Bindungstypen. Die Information genügt, um einen Molekülgraph mit annotierten Valenzzuständen zu erstellen, von dem Atomtypen abgeleitet werden können (vgl. Abschnitt 4.1.3). Mona betrachtet Moleküle genau dann als identisch, wenn die Valenzzustandsfolgen ihrer kanonischen Mol-Strings identisch sind. Dies ist der Fall, wenn die Topologie der Moleküle übereinstimmt. Zur dreidimensionalen Molekülrepräsentation werden kartesische Koordinaten der kanonisierten Atome (die *Konformation*) in Form eines Koordinaten-BLOBs (Binary Large Objekt) in der Moleküldatenbank gespeichert. Die Konformation ist jedoch nicht Bestandteil des Mol-Strings.

4.5.3 Die Moleküldatenbank

Im Mona-Kontext findet der Begriff des *Moleküls* eine andere Anwendung. Er wird dazu verwendet, um zwischen dem einmaligen Vorhandensein und dem mehrmaligen Vorkommen eines Eingabemoleküls im gegebenen Datensatz zu unterscheiden.

Molekül: Ein Molekül ist eine Entität, die das Vorhandensein eines Eingabemoleküls im Datensatz registriert und dessen Eigenschaften beschreibt. Jedes Molekül ist über seinen *Molekülschlüssel* eindeutig identifizierbar.

Instanz: Eine Instanz ist einem Molekül zugeordnet und ist eine Entität, die das Vorkommen eines Eingabemoleküls registriert. Für jedes Molekül existiert mindestens eine Instanz. Mehrere Instanzen repräsentieren Duplikate bzw. Ausprägungsformen eines Eingabemoleküls, dessen Eigenschaften sich nur geringfügig von seinem assoziierten Molekül unterscheiden. Jede Instanz wird über ihren *Instanzschlüssel* identifiziert.

Molekülmenge: Molekülmengen sind Mengen verschiedener Molekülentitäten. Sie werden durch eine Liste von Molekülschlüsseln repräsentiert.

Diese Begriffe sind im Datenbankschema der Mona zugrundeliegende Moleküldatenbank reflektiert, das in Abbildung 4.5.3 dargestellt ist.

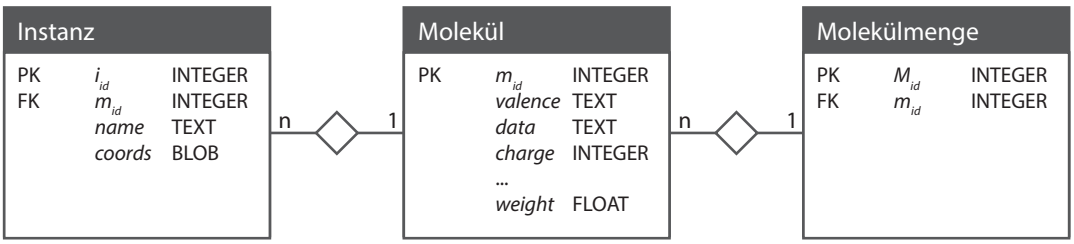


Abbildung 4.3: Schema der Moleküldatenbank.

Bei chemieinformatischen Anwendungen können Molekülduplikate Folgeberechnungen stören und Ergebnisse verfälschen. Aus diesem Grund wurde in Mona der Instanzbegriff eingeführt. Eingabemoleküle werden von Mona als Instanzen interpretiert, wenn sie ein Duplikat eines bereits registrierten Moleküls darstellen. Je nach Anwendung können sich die Kriterien unterscheiden, die darüber entscheiden wann ein Molekül als Duplikat eines anderen zu interpretieren ist. Als Molekülduplikate können Resonanzformen, Konformationen, Stereoisomere, Tautomere oder Protonierungszustände aufgefasst werden. Da die Identifizierung von Molekülen und ihren Instanzen auf einer kanonisierten, topologischen Beschreibung von Eingabemolekülen basiert, werden Resonanzformen und Konformationen von Mona immer als Instanzen eines Moleküls

interpretiert. Tautomere und Protonierungszustände werden als separate Molekülentitäten aufgefasst. Dasselbe gilt für Stereoisomere deren Stereozentren nicht ausreichend beschrieben sind (bei SMILES). Allerdings erlaubt Mona dem Benutzer zu steuern, ob Tautomere, Protonierungszustände und Stereoisomere während des Importprozesses als Duplikate betrachtet werden sollen. Dafür kann während des Imports angeben werden, ob ein Molekül zunächst in einen Grundzustand (d. h. in ein kanonisches Tautomer, in die neutralisierte Form oder ohne Berücksichtigung der Stereochemie) überführt werden soll. In diesem Fall werden aus Tautomeren, Protonierungszuständen und Stereoisomeren Instanzen erzeugt. Die Instanz eines Moleküls beschreibt also nicht nur echte Duplikate. Vielmehr ist sie ein Mittel, um eine im Datensatz beobachtete Ausprägungsform des Moleküls darzustellen. Zum Export der präparierten Daten in ein molekulares Dateiformat kann ausgewählt werden, ob nur auf die Molekül-assoziierten Daten der zuerst registrierten Instanz, oder auf die Daten aller Instanzen, zurückgegriffen werden soll.

4.5.4 Operationen auf Molekülmengen

Mona realisiert einen Arbeitsschritt, indem es ein oder mehrere Eingabemolekülmengen verarbeitet und als Resultat eine Ausgabemolekülmenge produziert. Prinzipiell unterstützt Mona zwei Arten von Operationen auf Molekülmengen:

Klassische mathematische Mengenoperationen arbeiten auf mehreren Eingabemengen und resultieren in einer Ausgabemenge. Sie werten die Molekülschlüssel der Molekülmengen und somit lediglich die Topologie der Moleküle aus und realisieren ihre Vereinigung, ihren Schnitt oder ihr Komplement. Diese Operationen sind mittels einfach formulierten SQL-Anweisungen umgesetzt und sind effizient auch auf großen Datensätzen durchführbar. Die Aufteilung von Molekülen nach unterschiedlichen Kriterien lässt sich mit diesen Operationen ebenfalls bewerkstelligen.

Filterung und Selektion operieren auf einer Eingabemenge und resultieren in einer Ausgabemenge. Die erzeugte Molekülmenge schließt die Moleküle der Ursprungsmenge aus, die gegebene Kriterien nicht erfüllen. Die Kriterien für den Ausschluss können entweder benutzerdefinierte molekulare Eigenschaften oder die visuelle Selektion von Molekülen sein. Die Moleküleigenschaften, für die Filterkriterien definiert werden können, sind in Tabelle 4.6 gegeben. Eine Kombination von Kriterien wird über eine Filterkette realisiert. Die Filterkette wird dabei durch logische UND-Verknüpfung einzelner Kriterien definiert. Für Filterketten können *Toleranzen* angegeben werden. Toleranzen definieren die Anzahl der Filter einer Filterkette, die notwendiger Weise erfüllt sein müssen. Solche Filterketten können über Mona definiert, als *Molekülprofil* in eine XML-Datei

geschrieben und wiederholt verwendet werden. Die Effizienz einer Filterkette ist abhängig von den Arten der ausgewählten Filter. Filter, die Moleküleigenschaften wie die Existenz eines Elements oder einer funktionellen Gruppe bewerten, sind schnell, da die entsprechenden Werte vorberechnet und für Existenz- oder Bereichsanfragen optimiert in der Moleküldatenbank als Attribute hinterlegt sind. Solche Filter können direkt in SQL-Anweisungen zur Datenbankabfrage umgesetzt werden. SMARTS-Filter sind die aufwendigsten Filter, da sie zunächst den Aufbau der Moleküle über den Mol-String erfordern, um dann den SMARTS-Ausdruck[327] auszuwerten. Toleranzen haben ebenso Einfluss auf die Geschwindigkeit der Filterung, da sie mehre Datenbankanfragen erfordern.

Tabelle 4.6: Mögliche Moleküleigenschaften einer Profildefinition. Die Eigenschaften werden entweder für Bereichsanfragen oder für Existenzanfragen in der Moleküldatenbank registriert und können zur Filterung von Molekülmengen genutzt werden.

Anfragentyp	Moleküleigenschaften
Bereich	Gesamtladung, molekulares Gewicht, Volumen, polare Oberfläche (TPSA), Oktanol-Wasser-Verteilungskoeffizient (PLogP), Anzahl der Schweratome, Anzahl der Heteroatome, Anzahl potentieller Wasserstoffbrückendonoren und -akzeptoren, Anzahl aromatischer Atome, Anzahl der Halogenatome, Anzahl der Bindungen, Anzahl rotierbarer Bindungen, Maximum aufeinanderfolgender rotierbarer Bindungen, Anzahl der Ringsysteme, Anzahl individueller Ringe, Anzahl aromatischer Ringe, maximale Ringgröße, maximale Ringsystemgröße und die Anzahl der Stereozentren
Existenz	Alle chemischen Elemente des Periodensystems, ein vordefiniertes Molekülmuster (SMARTS), typische funktionelle Gruppen (Alkohole, Ether, Ketone, Aldehyde, Ester, Amine, Amide, Amidine, Guanidine, Azide, Nitrile, Pyrrole, Furan, Thiophen, Phenyl, Pyridine)

4.6 TrixX(-BMI): Strukturbasiertes virtuelles Screening

TrixX[316] und TrixX-BMI[2] sind die Flex*-basierten Vorgängerversionen von cRAISE. Im Folgenden sind die Konzepte, das Interaktionsmodell, der Deskriptor sowie die deskriptorbasierten Docking- und Screening-Abläufe von TrixX und TrixX-BMI vorgestellt. Das Hauptaugenmerk ist auf wesentliche Unterschiede zwischen den beiden Versionen und auf Aspekte gelegt, die cRAISE von ihnen abgrenzt.

4.6.1 Der zweigeteilte virtuelle Screening-Prozess

Die zentrale Idee der TrixX-Entwicklungen ist eine Zweiteilung des VS-Prozesses:

Präparierungsphase: Jegliche molekulare Information, die in den Screening-Phasen benötigt werden könnte, wird einmalig aus der Molekülbibliothek abgeleitet und persistent gespeichert. Die Speicherung erfolgt derart, dass auf individuelle molekulare Eigenschaften direkt und rasch zugegriffen werden kann.

Screening-Phase: Die Screening-Phase extrahiert ausschließlich die notwendige Information aus der präparierten Bibliothek, die für ein gegebenes Anfrageprotein passend ist. Der Zugriff auf unnötige Information wird vermieden.

Die Zerlegung des Prozesses hat den Vorteil, dass die Screening-Phase den Großteil der Molekülbibliothek nicht auswerten muss. Dies ermöglicht die effiziente Prozessierung umfangreicher Bibliotheken. Die Strategie setzt jedoch voraus, dass die Bibliothek statisch ist und sich inhaltlich selten ändert. Unter dieser Voraussetzung, kann die molekulare Information einmalig abgeleitet und entsprechend ihrer Eigenschaften kategorisiert verwaltet werden. Die persistent gespeicherte Information kann wiederholt genutzt werden, um potentielle Liganden für verschiedenartige Zielproteine zu identifizieren.

4.6.2 Der TrixX-Deskriptor

Die Umsetzung der Strategie stellt zwei entscheidende Anforderungen an die Repräsentation der molekularen Information:

- Da die Repräsentation persistent gehalten und aus großen Molekülbibliotheken abgeleitet werden soll, ist eine möglichst kompakte Molekülrepräsentation notwendig.
- Aus der Information muss ersichtlich sein, ob ein Molekül prinzipiell mit dem Zielprotein interagieren kann und in dessen aktive Bindetasche passt.

Diese Anforderungen legt die Nutzung pharmakophorartiger Deskriptoren nahe, die die relative räumliche Anordnung potentieller Interaktionsstellen und die Passform der Moleküle beschreiben. Der TrixX-*Deskriptor* kodiert diese Eigenschaften. Er wurde in TrixX grundlegend und in TrixX-BMI weiterentwickelt. Letzterer hat in Form des RAISE-Deskriptors, mit geringfügigen Anpassungen, bis heute Bestand. Der TrixX-Deskriptor dient nicht nur zur kompakten Speicherung der molekularen Information

(*Molekül*deskriptoren), sondern wird auch zum zielgerichteten Zugriff auf diese verwendet. Dafür werden pharmakophorartige Anfragebedingungen gestellt, die von einem gegebenen Protein abgeleitet werden (*Anfragedeskriptoren* oder *Proteindeskriptoren*). TrixX identifiziert potentielle Liganden mittels Protein-Ligand-Docking, das über einen Vergleich vorberechneter Protein- und Molekül deskriptoren realisiert wird. Komplementäre Paare der Deskriptoren sollen hierbei möglichst nur potentielle Posen induzieren, die Interaktionen zum Zielprotein etablieren und so die Lage affiner Liganden reflektieren. Dementsprechend müssen komplementäre Deskriptoren Interaktionen zwischen Protein und Ligand widerspiegeln. Wie eine Interaktion definiert ist, wird durch das *Interaktionsmodell* beschrieben. Die Berechnung der TrixX-Deskriptoren erfolgt gemäß dieses Modells. TrixX und TrixX-BMI basieren auf demselben Interaktionsmodell, das jedoch vom Naomi-basierten Interaktionsmodell von cRAISE abweicht.

4.6.3 Das TrixX-Interaktionsmodell

Das Interaktionsmodell von TrixX und TrixX-BMI vereinfacht das Interaktionsmodell von FlexX[121], das auf den Arbeiten von Böhm[207, 206] und Klebe [330] beruht. Funktionelle Gruppen sind in Protein und Ligand mit *Interaktonsgeometrien* versehen (vgl. Abbildung 4.4). Indem der Typ, die relative Anordnung und Orientierung der Interaktionsgeometrien im Raum überprüft wird, kann für zwei funktionelle Gruppen das Potential zur Bildung einer Interaktion eingeschätzt werden. Nach dem TrixX-Interaktionsmodell etablieren zwei funktionelle Gruppen eine Interaktion, wenn

- die Interaktionstypen der funktionellen Gruppen kompatibel sind und
- die Interaktionszentren jeweils auf der Interaktionsoberfläche des Partners liegen. Dies ist der Fall, wenn zwei ihrer Interaktionsstellen nahezu räumlich zur Deckung kommen und die Interaktionsrichtungen entgegengesetzt orientiert sind.

Interaktionsgeometrien (vgl. Abbildung 4.5) besitzen einen *Interaktionstyp*, der von der Art *Donor*, *Akzeptor*, *Metall* oder *Hydrophob* sein kann. Sind die Typen zweier Gruppen kompatibel, so besteht die Möglichkeit zur Interaktion. Die Konstellation der Typen entscheidet, welche Art von Interaktion etabliert wird. Wasserstoffbrücken sind möglich beim Aufeinandertreffen von Donor- und Akzeptortypen. Hydrophobe Kontakte werden durch Paare von hydrophoben Typen etabliert. Eine Koordinierung durch ein Metallion erfordert einen Akzeptortyp auf Molekülseite und einen Metalltyp auf Proteinseite. Eine Interaktionsgeometrie ist durch ein *Interaktionszentrum*, einen *Interaktionsradius* und eine *Interaktionsoberfläche* charakterisiert. Die Interaktionsoberfläche ist Teil einer Kugeloberfläche, die um das Interaktionszentrum mit Interaktionsradius gegeben

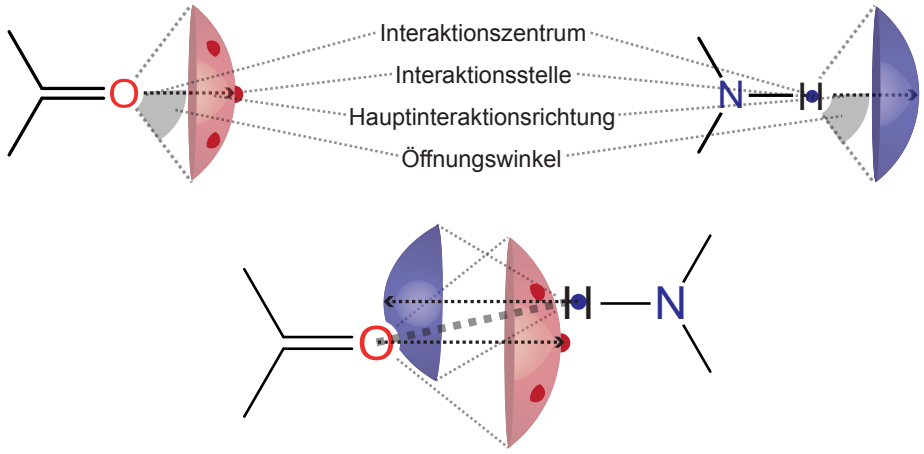


Abbildung 4.4: Interaktionsmodell: Kompatible und passend angeordnete Interaktionsgeometrien eines Proteinakzeptors (rot) und eines Moleküldonors (blau).

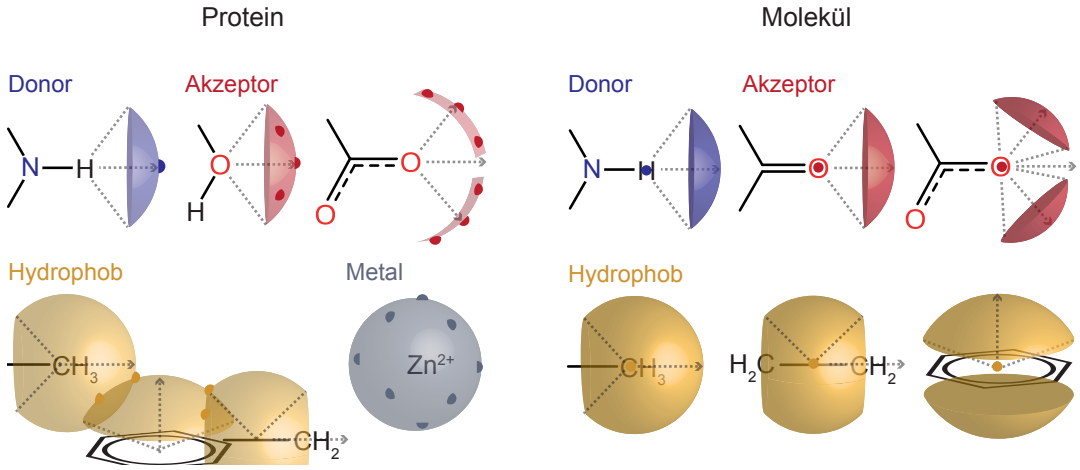


Abbildung 4.5: Interaktionsgeometrien im Protein (links) und Molekül (rechts).

ist. Auf Proteinseite entspricht sie einer der FlexX-Geometrien und kann die Form einer Kugeloberfläche, einer Kugelkappe, einer abgeschnittenen Kugelkappe oder eines sphärischen Rechtecks annehmen. Auf Molekülseite sind die Interaktionsgeometrien von TrixX und TrixX-BMI stets rotationssymmetrisch (vgl. Tabelle 4.7). Die Ausrichtung und Ausdehnung einer Geometrie ist über eine *Hauptinteraktionsrichtung* und einen *Öffnungswinkel* spezifiziert. Interaktionsoberflächen des Proteins sind auf ihren wasserzugänglichen Teil beschränkt und werden durch Punkte approximiert, die sich aus einem Clustering fein diskretisierter FlexX-Interaktionsoberflächen ergeben.[316] Die Punkte repräsentieren *Interaktionsstellen*, die eine Interaktionsstelle eines korrespondierenden

Interaktionspartners erwarten. Während Interaktionsstellen des Proteins immer auf den Interaktionsoberflächen liegen, liegen die Interaktionsstellen auf Ligandseite immer auf dem Interaktionszentrum und somit in der Regel auf Atomzentren. Hydrophobe Interaktionsstellen von Ethylgruppen und aromatischen Ringen bilden die Ausnahme. Ihre Interaktionsstellen sind auf der Bindung bzw. im Ringzentrum lokalisiert.

Tabelle 4.7: Rotationssymmetrische Interaktionsgeometrien von Molekülen.

Typ	Subtyp	Radius (Å)	Öffnungswinkel (°)	Toleranz (°)
Donor	Geladen	1.8	50	40
	Ungeladen	1.9	50	40
Akzeptor	Carboxylat	1.8	30 (2 Kegel)	50
	Andere	1.9	60	50
Metall	Alle	1.9	360	360
Hydrophob	Aromatischer Ring	4.5	70 (2 Kegel)	360
	Methyl	4.0	150	360
	Ethyl	4.0	30-50	360
	Halogen	4.8	150	360

4.6.4 TrixX-Deskriptoren im Vergleich

TrixX-Deskriptoren werden auf Basis von Interaktionsgeometrien abgeleitet. Moleküldeskriptoren kodieren pharmakophorartige Eigenschaften in einer kompakten Form. Proteindeskriptoren beschreiben korrespondierende Eigenschaften der aktiven Bindetasche. Zusammengefasst besitzt der TrixX-Deskriptor die folgenden Eigenschaften:

Dreieckseiten: Der TrixX-Deskriptor spiegelt ein Drei-Punkt-Pharmakophor wider. Die Punkte entsprechen Interaktionsstellen aus drei assoziierten Interaktionsgeometrien. Die relative, räumliche Anordnung der Interaktionsstellen ist durch Verbinden der Punkte zu einem Dreieck beschrieben. Lediglich die Seitenlängen des Dreiecks sind im Deskriptor kodiert. Prinzipiell erzeugt jedes mögliche Triplett von Interaktionsstellen aus unterschiedlichen Geometrien des Proteins bzw. eines Moleküls einen TrixX-Deskriptor. Allerdings dürfen die Dreiecksseiten eine gewisse Länge nicht überschreiten.

Deskriptortyp: Jeder Dreieckspunkt besitzt einen Typ, der dem Interaktionstyp der assoziierten Geometrie entspricht. Der Typ des Deskriptors ist durch eine kanonisierte Folge der Interaktionstypen der Dreiecksbasis bestimmt. Der Typ gibt an, welche Konstellation von Interaktionen ein Protein- bzw. Molekülbereich eingehen kann.

Interaktionsrichtungen: Jeder Dreieckspunkt besitzt eine Interaktionsrichtung, die der Interaktionsrichtung der assoziierten Geometrie entspricht. Die Interaktionsrichtung gibt an, in welcher Orientierung eine Interaktion gebildet werden kann.

Bulk: Der sogenannte *Bulk* beschreibt die molekulare Ausdehnung bzw. den Hohlraum einer Bindetasche. Er ist auf die lokale Umgebung um die Deskriptorbasis beschränkt.

TrixX und TrixX-BMI kodieren diese Deskriptoreigenschaften unterschiedlich. TrixX spezifiziert Interaktionsrichtungen über Eulerwinkel. TrixX-BMI beschreibt dagegen eine Richtung über kartesische Koordinaten des Einheitsvektors. Durch Kanonisieren der Dreiecke (mittels Sortierung der Interaktionstypen und Seitenlängen) wird in beiden Versionen ein definiertes lokales Koordinatensystem bezüglich der Dreiecksbasis bestimmt, mit dem sich die Orientierungen im Raum beschreiben lassen. Ein weiterer Unterschied besteht in der Repräsentation des Bulks (vgl. Abbildung 4.6). Der TrixX-Deskriptor untersucht nur sehr lokal, ob sich das molekulare Volumen über den Bereich seiner Dreiecksbasis erstreckt. Die Dreiecksbasis wird nach außen erweitert und das vergrößerte Dreieck oberhalb und unterhalb der Basis positioniert. Für jedes erweiterte Dreiecke kodieren ein Bit im Zentrum, drei Bits an den Ecken und je neun Bits an den Seiten, ob sich das Molekül über das extrudierte Dreieck erstreckt. Insgesamt repräsentieren somit 2×31 Bits den Bulk des TrixX-Deskriptors. Diese Beschreibung spiegelt jedoch kaum die Molekülform wider. Der TrixX-BMI-Deskriptor approximiert daher die Form des Moleküls anhand von 80 annähernd gleichmäßig verteilten Bulk-Strahlen. Die Strahlen resultieren aus einem verfeinerten Ikosaeder, der am lokalen Koordinatensystem des Dreiecks ausgerichtet ist. 80×15 Bits kodieren die Länge der Strahlen.

4.6.5 Deskriptorbasiertes Protein-Ligand-Docking

Mit vorberechneten Protein- und Molekküldeskriptoren lässt sich ein Protein-Ligand-Docking wie folgt realisieren:

1. Vergleiche alle Protein- mit allen Molekküldeskriptoren und identifiziere nach Definition 4.6.1 Paare von kompatiblen Deskriptoren (*Deskriptortreffer*).
2. Berechne für jeden Deskriptortreffer die Transformation des Molekül- auf den Proteindeskriptor.
3. Wende jede Transformation auf das Molekül an.
4. Bewerte die dadurch erzeugten Posen.

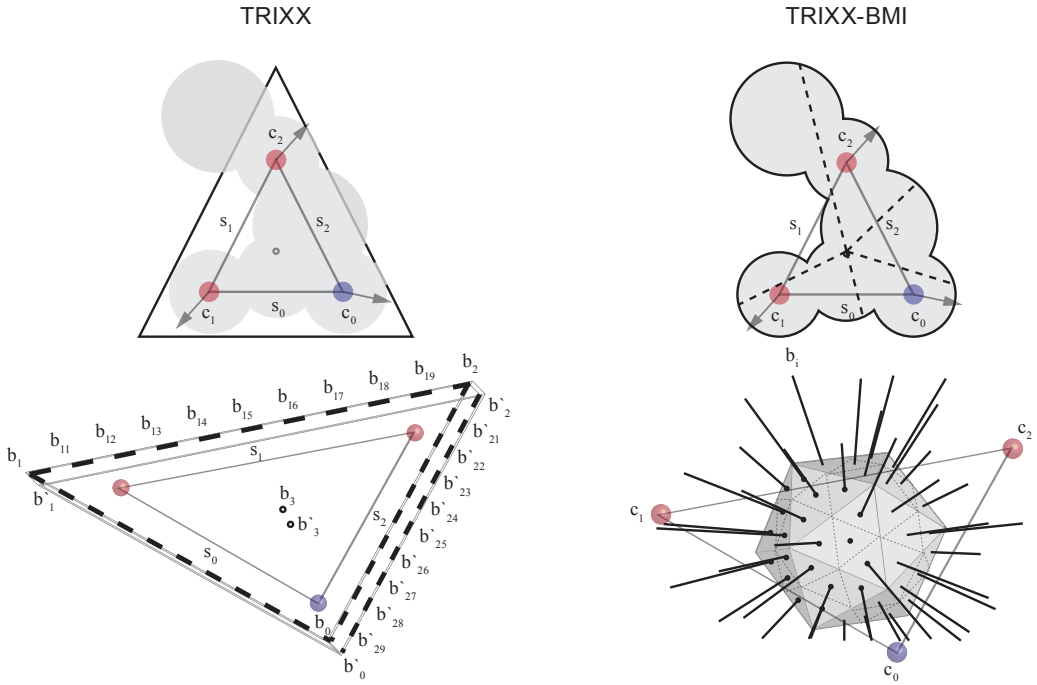


Abbildung 4.6: Kodierung des Bulks: TrixX kodiert den Bulk über eine extrudierte Dreiecksbasis (links), TrixX-BMI über 80 an der Dreiecksbasis ausgerichtete Strahlen.

Definition 4.6.1 (Deskriptortreffer)**.** *Ein Paar von Protein- und Moleküldeskriptoren ist ein Deskriptortreffer, wenn ihre Deskriptortypen kompatibel sind, die Dreiecksseiten paarweise nahezu gleich lang sind, die assoziierten Interaktionsrichtungen entgegengesetzt orientiert sind und der Protein-Bulk stets größer als der Molekül-Bulk ist.*

Alle TrixX-Entwicklungen basieren auf dieser Platzierungsmethode. Sie unterscheiden sich jedoch in welcher Art und Weise Molekülflexibilität gehandhabt wird. TrixX verfolgt bei der Platzierung einen Placement-and-Linking-Ansatz (vgl. Abschnitt 3.3.4). Dabei wird das Molekül in sich teilweise überlappende und eine gewisse Anzahl an Interaktionsstellen aufweisende Fragmente zerlegt. Für jedes Fragment werden präparierend Teilkonformationen erzeugt, anhand derer Deskriptoren abgeleitet werden. Der Deskriptorabgleich identifiziert Deskriptortreffer, die dann Fragmentplatzierungen generieren. Kompatible Fragmentplatzierungen werden zu Molekülplatzierungen verbunden, wenn die Distanz zwischen ihnen innerhalb einer gewissen Toleranz liegt. Der Fragment-Linking-Ansatz ist sehr ressourcenschonend, da für identische Fragmente, Konformationen und Deskriptoren nur einmalig berechnet und gespeichert werden müssen. Das Vorgehen kann jedoch Posen mit invaliden Bindungslängen und -winkel produzieren. Zudem

kann ein Ligand nur dann erfolgreich identifiziert werden, wenn zuvor alle Fragmente erfolgreich platziert und zugleich minimale Abstände zwischen den Teilplatzierungen eingehalten wurden. Das Risiko nur Platzierungen zu erhalten, die nicht wieder miteinander verbunden werden können, ist relativ hoch. TrixX-BMI gab den Placement-and-Linking-Ansatz zugunsten eines inkrementellen Aufbaus auf (vgl. Abschnitt 3.3.4). Es fragmentiert initial nur sehr große Moleküle mit mehr als acht rotierbaren Bindungen. Für die Fragmente und kleinere Moleküle werden Konformationen und Deskriptoren generiert. Aus einem Deskriptortreffer kann dann meist direkt eine vollständige Pose erzeugt werden. Im Fall eines Fragmenttreffers werden die Konformationen der restlichen Fragmente sukzessive angebaut, bewertet und so zu vollständigen Posen erweitert. Dieses Vorgehen hat den Vorteil, dass inkompatible Fragmentplatzierungen vermieden und nur valide Posen erhalten werden. Die grobe Fragmentierung verlagert die Modellierung der Molekülflexibilität jedoch auf den integrierten Konformergenerator[253]. Zudem müssen deutlich mehr Moleküldeskriptoren erzeugt werden.

4.6.6 TrixX-Arbeitsabläufe im Vergleich

Da eine Überlagerung inkompatibler Deskriptoren zu Posen mit ungünstig ausgerichteten Interaktionen, repulsiven Interaktionen oder sterischen Kollisionen zwischen Protein und Molekül führen würde, unterstützen die TrixX-Entwicklungen einen zielgerichteten Zugriff auf ausschließlich komplementäre Moleküldeskriptoren. Ermöglicht wird dies, indem Moleküldeskriptoren entsprechend ihrer Attribute in einer Indexstruktur verwaltet werden. Sowohl TrixX, TrixX-BMI als auch cRAISE verwalteten ihre Deskriptoren der Molekülbibliothek in einer Indexstruktur. Proteindeskriptoren werden dazu genutzt, Anfragen an den Index zu stellen und nur passende Moleküldeskriptoren zu extrahieren ohne inkompatible auswerten zu müssen. Die Nutzung eines Index erfordert den zweigeteilten virtuellen Screening-Prozess: In der Präparierungsphase werden die Deskriptoren der Molekülbibliothek abgeleitet und in die Indexstruktur überführt. In der Screening-Phase werden aus Proteindeskriptoren Anfragen formuliert und entsprechende Deskriptortreffer extrahiert. Die Prozessabläufe beider TrixX-Versionen sind in Abbildung 4.7 gegenübergestellt. Die Indexierungs- und Screening-Phasen unterscheiden sich durch die geänderte Handhabung der Molekülflexibilität. TrixX berücksichtigt nur Fragmente, die noch nicht innerhalb der Molekülbibliothek beobachtet wurden und erzeugt Deskriptoren nur für diese. Dadurch werden identische Deskriptoren weitgehend vermieden und ein kompakter Index erzeugt. Die grobe Fragmentierung von TrixX-BMI erhöht die Chance für Redundanzen im Deskriptorindex, da gleiche und sehr ähnliche Deskriptoren in verschiedenen Fragmenten bzw. Molekülen auftreten können.

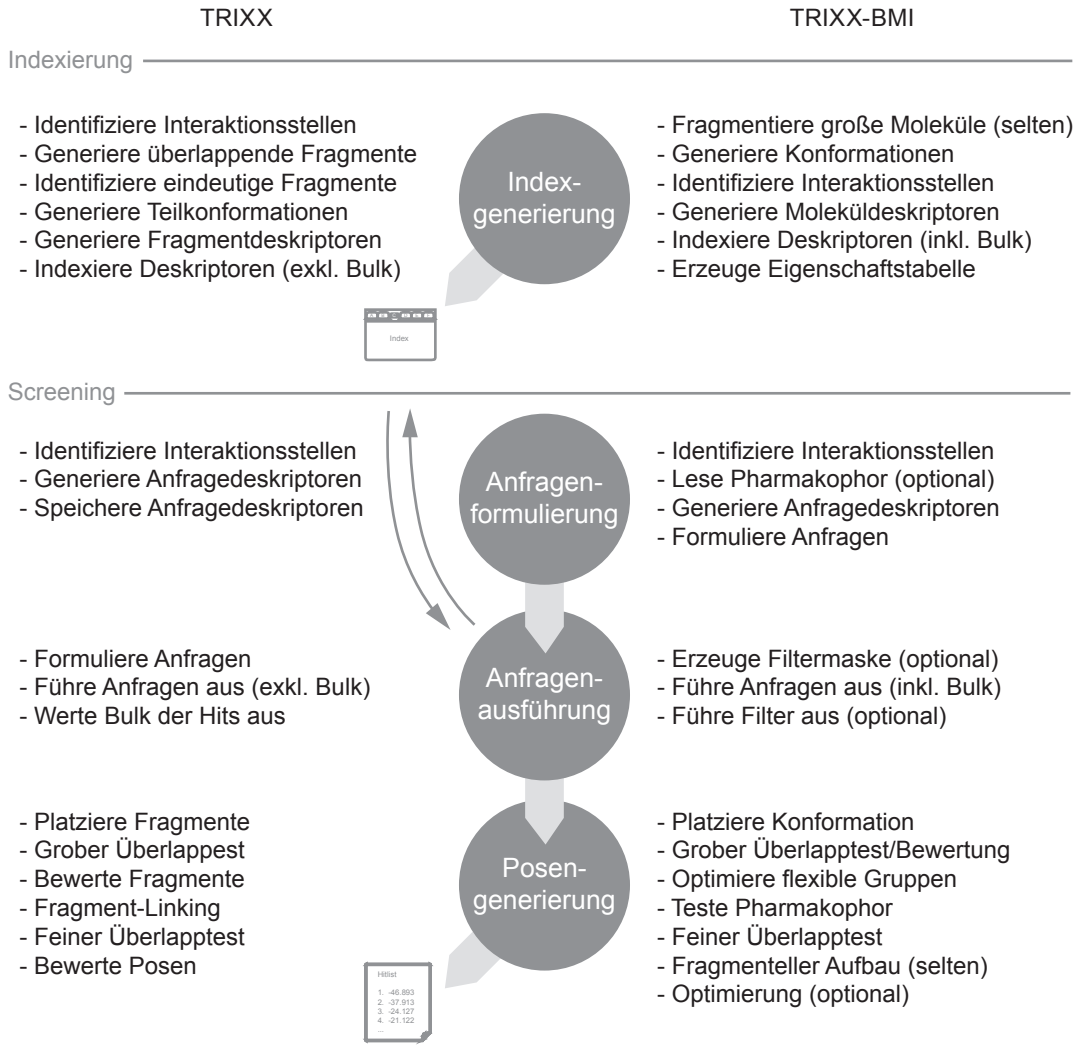


Abbildung 4.7: Gegenüberstellung der Arbeitsabläufe von TrixX und TrixX-BMI.

TrixX-BMI nimmt dies in Kauf und entgegnet dem dadurch gesteigerten Speicherbedarf mit einer Komprimierung der Deskriptoren. Ein weiterer Unterschied besteht in der Auswertung der Bulk-Beschreibung. TrixX indexiert den Bulk wegen seiner hohen Dimensionalität nicht. Als Folge kann er nicht zur Formulierung von Anfragebedingungen genutzt werden. Der Bulk wird deshalb nach Erhalt der Deskriptortreffer evaluiert. TrixX-BMI verfeinert die Bulk-Repräsentation und integriert ein effizientes Indexierungssystem, das in der Lage ist, die hochdimensionierten Daten handhaben und entsprechend hochdimensionierte Anfragen unterstützen zu können (siehe Abschnitt 4.7).

4.6.7 Geleitetes SBVS mit TrixX-BMI

Im Vergleich zu TrixX stellt TrixX-BMI zusätzliche Funktionalität bereit. Es ermöglicht benutzerdefinierte, pharmakophorartige Randbedingungen in der Screening-Phase aufzustellen. Hierfür nutzt es Pharmakophordefinitionen wie sie auch von FlexX-Pharm[309] (vgl. Abschnitt 3.7) unterstützt werden. Mit ihnen kann die Erfüllung spezifischer Interaktionen und räumlicher Inklusions- und Exklusionsmerkmale gefordert werden. Die Merkmale können zudem über SMARTS[327] mit Molekülbereichen assoziiert werden. Mehrere Pharmakophormerkmale werden mit Bool'schen Operatoren zu einer Pharmakophordefinition kombiniert. Sie wird zum einen, analog zu der Vorgehensweise von FlexX-Pharm, als Filter bei der Platzierung der Posen genutzt. Zum anderen wertet TrixX-BMI die Pharmakophormerkmale bereits auf Ebene der Anfragedeskriptoren aus. Es schränkt die Anfragen auf solche ein, die das Potential haben Posen zu produzieren, die die Merkmale erfüllen können. Dazu werden die Proteininteraktionsstellen den vordefinierten Pharmakophormerkmalen zugeordnet. Nur Anfragedeskriptoren, für die mindestens zwei Ecken Inklusionen zugeordnet sind und keine Ecke besitzen, die einem Exklusionsmerkmal zugeordnet sind, werden für das Screening genutzt. Ist nur ein Inklusionsmerkmal gegeben oder würden weniger als 50 Anfragedeskriptoren generiert, so muss nur ein Pharmakophormerkmal von den Anfragedeskriptoren erfüllt sein. Je nach Komplexität kann eine Pharmakophordefinition eine wesentliche Beschleunigung der Screening-Phase ermöglichen. Außerdem bietet TrixX-BMI die Möglichkeit, die Molekülbibliothek während des Screening-Prozesses nach wenigen molekularen Eigenschaften zu filtern. Es unterstützt die Filterung der Moleküle nach dem Molekulargewicht, der Anzahl an Wasserstoffbrückendonoren und -akzeptoren, die Anzahl enthaltener Ringe, die Anzahl rotierbarer Bindungen und dem logP. So kann beispielsweise eine einfache Filterung der Moleküle nach Oprea's Leitstrukturkriterium (vgl. Abschnitt 2.5.2) umgesetzt werden. Durch die Etablierung eines Molekülfilters, können Moleküle mit ungewollten Eigenschaften vom Screening ausgeschlossen werden, ohne dass eine erneute Indexierung der meist umfangreichen Molekülbibliothek notwendig wird.

4.6.8 Berücksichtigung variabler Wasserstoffpositionen

Die Dreiecksbasis der TrixX/TrixX-BMI-Deskriptoren beruht auf den Interaktionsgeometrien der als statisch betrachteten funktionellen Gruppen. Diese sind durch die bereitgestellten Protein- und Moleküldaten einmalig spezifiziert und verbleiben während des Docking-Prozesses unverändert. Die Lage und Art der Interaktionsstellen hängt vom Vorhandensein, der Position und der Orientierung von Wasserstoffen ab, die durch die

gegebenen Daten bereitgestellt sind. Eine Zustandsänderung von Protein und Ligand bei der Bindung kann jedoch die Lage und den Typ der Interaktionsgeometrien ändern. TrixX-BMI verfolgt einen ersten, stark vereinfachten Ansatz, um funktionelle Gruppen mit frei rotierbaren Wasserstoffen zu modellieren. Dafür werden die entsprechenden Interaktionsgeometrien als rotierbar markiert. Während des Deskriptorabgleichs wird dann der Richtungsabgleich an diesen Stellen ignoriert und schließlich werden die Wasserstoffe in den Posen nachträglich angepasst. Diese indirekte Art der Modellierung wird nur auf Ligandseite realisiert. Frei rotierbare Proteininteraktionsgeometrien werden nicht berücksichtigt. Eine Änderung der Anzahl oder der Position der Wasserstoffe, wie sie durch (De-)Protonierungen und Tautomerisierung der Moleküle/Protein auftritt, wird weder von TrixX noch von TrixX-BMI modelliert.

4.7 FastBit: Persistente Bitmap-Indizes

Die Daten molekularer Bibliotheken, die im VS durchsucht werden, sind äußerst umfangreich. Während der Suche können sie nicht im Hauptspeicher gehalten, sondern müssen im Sekundärspeicher verwaltet werden. Um eine effiziente Screening-Phase zu gewährleisten, sind deshalb einige Kriterien bezüglich der Datenverwaltung zu beachten:

- Die Anzahl an I/O-Operationen ist so gering wie möglich zu halten, um die Anzahl der Speicherseiten zu minimieren, die in den Hauptspeicher transferiert werden müssen. Deshalb ist ein sequenzielles Abtasten der Molekülbibliothek zur Identifizierung von Liganden in der Screening-Phase zu vermeiden.
- Bei hochskalierten Screening-Anwendungen unterscheidet sich der Inhalt der genutzten Molekülbibliotheken kaum. Zudem ändert die Screening-Phase die Daten nicht. Es muss daher nur ein effizienter, lesender Datenzugriff unterstützt werden.
- Für die Beschreibung der Molekülform muss es möglich sein, Deskriptoren mit einer hochdimensionalen Bulk-Repräsentation zu verwalten.
- Zugleich müssen die durch Definition 4.6.1 gegebenen, hochdimensionierten Anfragen zur Suche nach Deskriptortreffern formuliert werden können.

Die ersten beiden Kriterien werden durch eine Verwaltung der Daten mittels Indexstrukturen erfüllt. Die Wahl des Indexierungssystem hängt von der Art der verwalteten Daten und der Art der Anfragen ab. RAISE-Deskriptoren legen die Verwendung von Bitmap-Indizes nahe, wie sie im FastBit-System[331] realisiert sind. FastBit ist ein von John Wu, am Berkeley Lab entwickeltes Bitmap-Indexierungssystem. Es unterstützt

die Verwaltung und Anfrage hochdimensionierter Daten und wurde bereits von TrixX-BMI zur Verwaltung seiner Deskriptoren und zur Unterstützung von Anfragen auf die Indexstruktur integriert. cRAISE nutzt diese Komponente weiterhin. Es ist die einzig in cRAISE integrierte Methode, die nicht Bestandteil der Naomi-Bibliothek ist.

4.7.1 Bitmap-Indizes

Bitmap-Indizes[332] sind besonders für Szenarien geeignet, in denen primär ein lesender Zugriff auf hochdimensionierte Daten unterstützt werden muss und komplexe Anfragen beantwortet werden sollen. Ein Bitmap-Index etabliert einen Index, indem für jede Dimension ein einzelner Bitmap erzeugt wird. Er gliedert die Daten nicht horizontal, entsprechend der einzelnen Dateneinträge, sondern vertikal, sodass die Daten spaltenweise entsprechend ihrer Dimensionen organisiert sind. Mehrdimensionale Anfragen können auf Bitmaps effizient durch die Auswertung einfacher Bool'scher Ausdrücke beantwortet werden. Dafür werden die Bitmaps der auszuwertenden Dimensionen über Bool'sche Operationen verknüpft. Das Resultat ist ein Bitvektor, der Dateneinträge markiert, für die die gestellte Anfrage positiv beantwortet werden kann. Ein Beispiel für einen Bitmap-Index und dessen Zugriff ist schematisch in Abbildung 4.8 dargestellt.

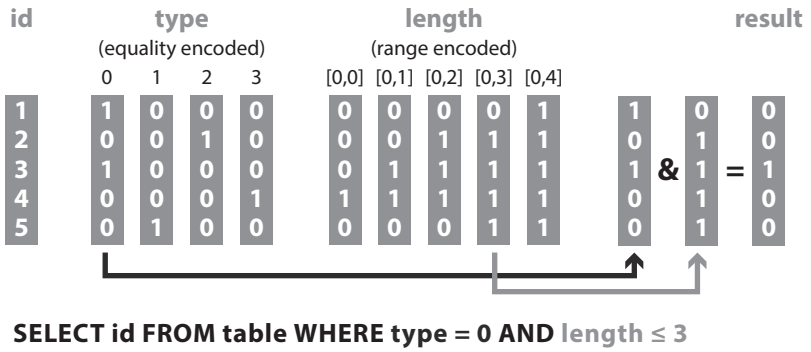


Abbildung 4.8: Die Anfrage verknüpft den Bitmap des Typs 0 mit dem Bitmap, der eine Länge bis zu 3 kodiert. Beide Bedingungen sind nur für den dritten Eintrag erfüllt. Das entsprechende Bit ist im Ergebnisvektor gesetzt. Die Typinformation ist für Gleichheitsanfragen kodiert, die Längeninformation für Bereichsanfragen.

4.7.2 Binning und Anfragenauswertung

Bitmap-Indizes haben den Nachteil, dass sie einen Mehraufwand an Speicherplatz benötigen, um Attribute hoher Kardinalität zu verwalten. Die hohe Kardinalität resultiert aus der Kodierung der meist kontinuierlichen Datenwerte mittels Bitmaps. FastBit

entgegnet diesem Problem, indem es ein Binning-Schema anwendet und einen Index auf Basis einer grobkörnigen Repräsentation der Daten erstellt. Die Grenzen der Bins können dabei entweder durch den Benutzer definiert oder durch eine statistische Analyse der Daten abgeleitet werden. FastBit wertet Anfragen aus, indem es die granularen Bitmap-Indizes nutzt, um zunächst Lösungskandidaten zu generieren. Diese Kandidaten werden dann individuell überprüft, indem das System die exakten Rohdaten heranzieht, um zu entscheiden, ob ein Kandidat die Anfragebedingung erfüllt oder nicht. Dieses Schema zur Auswertung von Anfragen kann auch auf zwei Ebenen hierarchisch angewendet werden. Dadurch wird die Granularität des Index erhöht und die Anzahl an Kandidaten reduziert.

4.7.3 Bitmap-Kodierung

Um unterschiedliche Arten von Anfragen, wie z. B. Gleichheits- oder Bereichsanfragen, durch eine Auswertung Bool'scher Ausdrücke beantworten zu können, kann die Bitmap-Kodierung angepasst werden (vergl. dazu Abbildung 4.8). Die Wahl für ein bestimmten Kodierungsschemas senkt die Anzahl an Bits, die während der Anfrage ausgewertet müssen.[333] Wenn eine Dimension primär für Bereichsanfragen genutzt wird, müssen die Bitmaps entsprechend kodiert werden. Eine Bereichsanfrage kann dann beantwortet werden, indem nur ein Bit ausgewertet werden muss. Wären Bitmaps in diesem Fall fälschlicherweise für Gleichheitsanfragen kodiert, müsste eine Bereichsanfrage die Bitmaps innerhalb des Bereichs mit einem logischen ODER kombinieren. Das resultiert im schlimmsten Fall in einer Auswertung aller möglichen Kombinationen an Dimensionen. FastBit unterstützt diverse Kodierungsschemata. Für cRAISE ist jedoch nur die Kodierung für Gleichheits- und Bereichsanfragen relevant.

4.7.4 Komprimierung der Bitmaps

Eine Möglichkeit den gesteigerten Speicherbedarf von Bitmap-Indizes zu begegnen besteht darin, die einzelnen Bitmaps zu komprimieren. Die Nutzung von Standardkompressionsalgorithmen würde jedoch zu einer gesteigerten Laufzeit führen. Jedes auszuwertende Bitmap müsste dekomprimiert werden, um eine Anfrage zu beantworten. Eine Komprimierung der Bitmaps muss daher sicherstellen, dass bei der Anfragenprozessierung die Bool'schen Operationen wie UND, ODER und NICHT mittels standardmäßiger CPU-(central processsing unit)-Operationen auf den komprimierten Bitmaps realisiert werden können. FastBit nutzt eine spezielle Komprimierungsmethode, den sogenannten Word Aligned Hybrid (WAH) Code.[334] Dieser erlaubt es, Standardoperationen ohne eine Dekomprimierung der Bitmaps auszuführen. Folgen gleichwertiger

Bits werden kompakt repräsentiert, indem der Wert der gleichwertigen Bits, gefolgt von der Länge der Bitfolge, kodiert wird. Folgen unterschiedlicher Bits verbleiben unkomprimiert. Die zentrale Idee von WAH ist eine Gruppierung der Bits innerhalb komprimierter Bitmaps, sodass eine Gruppe CPU-Wortlänge − 1 Bits beinhaltet, z. B. 31 Bits auf einer 32-Bit CPU. Damit können gleichwertige Bitfolgen bis zu einer Länge von 2^{30} durch 32 Bits dargestellt werden. Ein zusätzliches Bit-Flag wird dazu benötigt, um Komprimierung anzuzeigen. Eine komprimierte Gruppe wird mit einer führenden 1 versehen. Unkomprimierte Wörter werden eins-zu-eins gespeichert und mit einer führenden 0 versehen. Durch diese Gruppierung der komprimierten Bits wird gewährleistet, dass einzelne Gruppen somit exakt in ein Wort der CPU passen. Auf diesen können weiterhin die Standardoperationen der CPU ausgeführt werden, um die Anfragen direkt auf den Bitmaps auszuführen ohne dafür zuvor die Bitmaps dekomprimieren zu müssen.[335, 336]

5 cRAISE: Überblick

cRAISE ist eine Methode zum strukturbasierten virtuellen Screening. Es baut auf den Konzepten seiner Flex*-basierten Vorgängerversionen TrixX und TrixX-BMI auf, repräsentiert intern jedoch Moleküle und Zielproteine gemäß des Naomi-Modells. Durch Naomi und der darauf basierenden, in Kapitel 4 vorgestellten Methoden wurden während der Entwicklung von cRAISE eine Vielzahl neuartiger Funktionalitäten zugänglich. In cRAISE integriert, kommen sie zum Teil weiterentwickelt zur Anwendung. Dadurch ist eine strukturbasierte virtuelle Screening-Methode entstanden, die Moleküle und Protein, auf Basis einer einheitlichen Repräsentation, konsistent handhabt. Sie schafft außerdem einen erweiterten Rahmen zum Eingriff in den Screening-Prozess durch den Anwender und bietet die Möglichkeit, vielfältig auf das Screening-Resultat einzuwirken. Zudem realisiert cRAISE, effizient und simultan, Zustandsänderungen von Protein und Ligand während des Screening-Prozesses zu berücksichtigen. Dieses Kapitel stellt die Arbeitsabläufe vor, die cRAISE hierfür durchführt. Es bietet einen Überblick über den grundlegenden, strukturbasierten VS-Prozess, den geleiteten VS-Prozess unter gegebenen Randbedingungen und den VS-Prozess bei Berücksichtigung molekularer und makromolekularer Zustände. Zudem wird die Integration und Modifikation der Naomi-basierten Komponenten aufgezeigt.

5.1 Strukturbasiertes virtuelles Screening

Wie seine Vorgängerversionen ist cRAISE eine zweigeteilte Prozedur. In Abbildung 5.1 ist der grundlegende Ablauf beider Phasen dargestellt. In der Präparierungsphase werden Konformere und RAISE-Deskriptoren für eine gegeben Molekülbibliothek generiert. Die Konformationen werden in einer Moleküldatenbank, die Deskriptoren in einem zugehörigen Bit-komprimierten Index gespeichert. Beide können wiederholt genutzt werden und verbleiben in den darauffolgenden VS-Läufen unverändert. Die Screening-Phase

leitet RAISE-Deskriptoren von einer gegebenen Proteinbindetasche ab, anhand derer SQL-ähnliche Anfragen formuliert werden. Unter Verwendung des Bitmap-Index detektieren die Anfragen passende Moleküldeskriptoren. Die Konformere der Deskriptortreffer werden aus der Moleküldatenbank extrahiert, in die aktive Bindetasche mittels Überlagerung der Deskriptoren platziert und schließlich bewertet. Die beste Pose eines Moleküls wird in die finale Trefferliste des Screening-Laufes eingetragen.

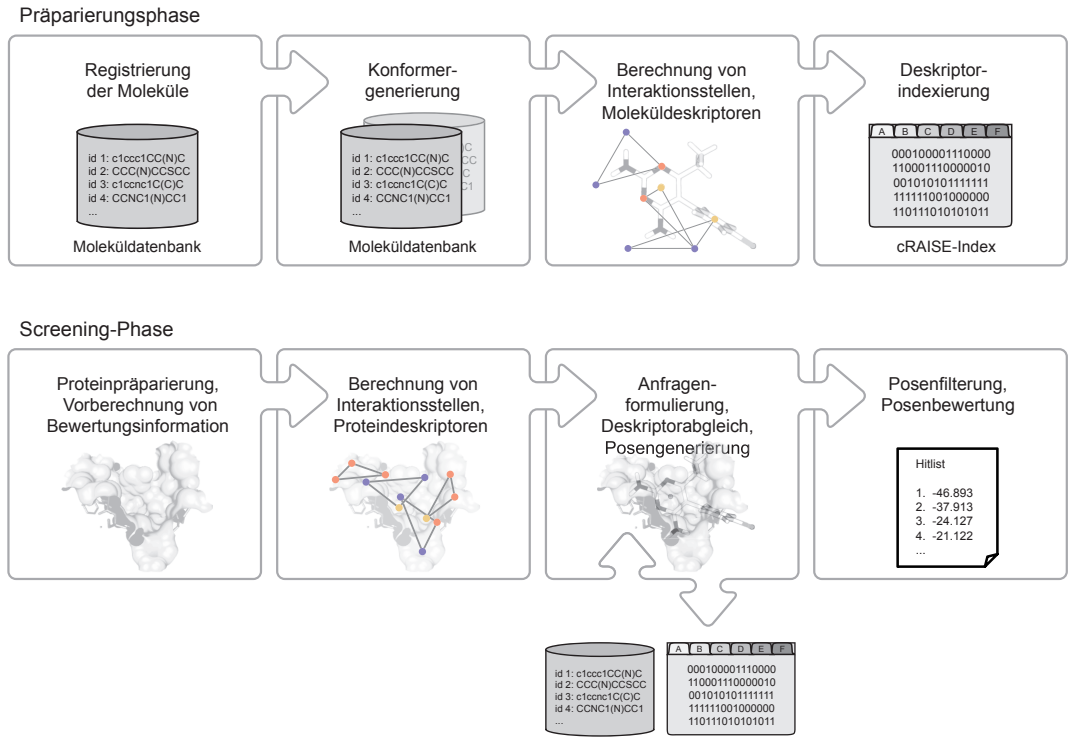


Abbildung 5.1: Grundlegende Arbeitsabläufe der Präparierungs- und Screening-Phase

Die wesentlichen Konzepte des Deskriptors, der deskriptorbasierten Docking-Methode und des indexbasierten VS-Prozesses beruhen auf den in den Abschnitten 4.6 und 4.7 vorgestellten Arbeiten. In cRAISE sind sie überarbeitet und in Hinblick auf der in Abschnitt 4.1 vorgestellten Naomi-Repräsentation von Molekülen und Proteinen angepasst. Auf Basis des Naomi-Modells wird in dieser Arbeit ein vereinfachtes Interaktionsmodell postuliert, das als Grundlage zur Berechnung von Interaktionstellen und RAISE-Deskriptoren dient, aber auch zur Bewertung von Docking-Lösungen herangezogen wird. In der Präparierungsphase importiert cRAISE initial Moleküle aus einer gegebenen Molekülbibliothek in eine Moleküldatenbank. Die Moleküldatenbank unterscheidet sich hierbei nur geringfügig von der in Mona genutzten Datenbank, die in Ab-

schnitt 4.5 vorgestellt wurde. In cRAISE ist sie hinsichtlich einer Verwaltung von Konformationsdaten und für einen möglichst raschen Zugriff auf diese optimiert. Zudem ist die Moleküldatenbank durch Zusatzinformationen erweitert, um eine rasche Posenintialisierung während der Screening-Phase zu gewährleisten. Die Docking- bzw. Screening-Lösungen werden ebenso in einer Variante dieser Datenbank hinterlegt. Zur Generierung von Konformationen nutzt cRAISE den in Abschnitt 4.4 vorgestellte Algorithmus und die Torsionsbibliothek von Confect. cRAISE führt jedoch eine neue Qualitätsstufe und ein anderes Clustering-Verfahren ein, um den Anforderungen im strukturbasierten VS umfangreicher Molekülbibliotheken gerecht zu werden.

5.2 Geleitetetes strukturbasiertes virtuelles Screening

Ein erster Entwurf eines pharmakophor- und eines molekülprofilgeleiteten VS-Prozesses wurde bereits in TrixX-BMI konzipiert. Bei der Entwicklung von cRAISE wurden diese Konzepte weiterverfolgt und maßgeblich erweitert. So entstand eine sehr flexible Methode, die ein Screening unter benutzerdefinierten Randbedingungen vielfältig und effizient unterstützt. Die Methode macht sich die in der Moleküldatenbank registrierten Moleküleigenschaften zunutze, die in der Präparierungsphase, während des Molekülimports *per se* detektiert und gespeichert werden. Auf die Präparierungsphase haben gegebene Randbedingungen sonst keinen Einfluss. Profildefinitionen wie sie durch Mona spezifiziert werden, können mit cRAISE verarbeitet werden, um die Bibliothek während des Screenings zu filtern. Für ein Screening unter solchen Randbedingungen wurde die grundlegende Screening-Phase von cRAISE modifiziert. Der Ablauf einer derart geleiteten Screening-Phase ist in Abbildung 5.2 dargestellt.

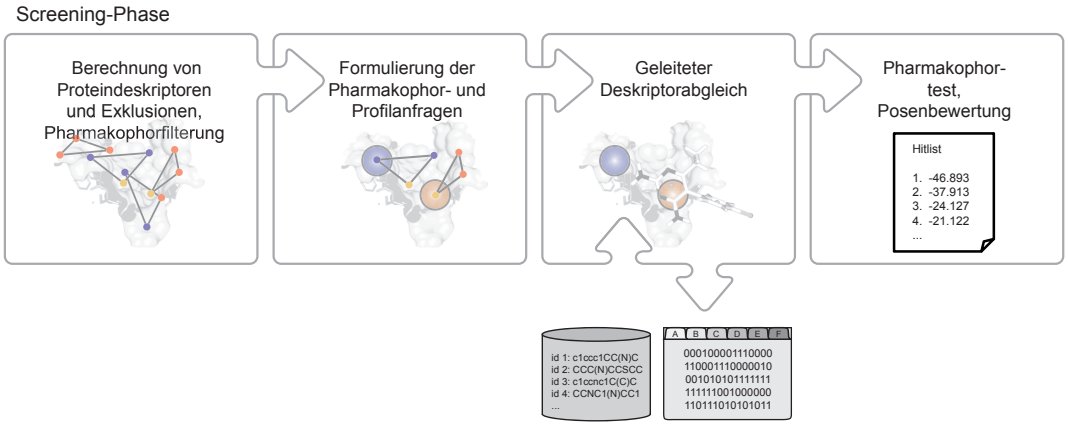


Abbildung 5.2: Screening-Phase bei gegebenen Randbedingungen

Die gegebene Zusatzinformation wird nicht nur zur Selektion der geforderten Ergebnisse, sondern auch zugunsten einer Beschleunigung des Prozesses genutzt. Ein Molekülprofil wird direkt in der Anfrage kodiert, sodass Moleküle, die eine gegebene Anforderung nicht erfüllen, nie aus der Datenbank aufgebaut werden müssen. Optional kann cRAISE auch eine Pharmakophorhypothese nutzen, um ähnlich wie TrixX-BMI die Anfragedeskriptoren plausibel zu reduzieren, sodass nur Posen erzeugt werden, die die Pharmakophordefinition erfüllen können. Auf Deskriptorebene kann die Erfüllung eines Pharmakophors jedoch nur in einer sehr lokalen Umgebung überprüft werden. Deswegen werden die Pharmakophormerkmale während der Posenprozessierung erneut auf ihre globale Einhaltung überprüft. Posen werden vor der eigentlichen Bewertungsphase verworfen, wenn sie der gegebenen Hypothese widersprechen. Für die Verwendung in cRAISE wurde in dieser Arbeit eigens ein Merkmalsschema entworfen, das Pharmakophordefinitionen durch ungerichtete Level 4 und gerichtete Level 3 Merkmale (vgl. Abschnitt 3.4.1) unterstützt. Es bietet die Möglichkeit, verschiedene Arten von Inklusions- und Exklusionsmerkmalen zu spezifizieren und verbindet sie mittels einer simplen Logik zu einer vollständigen Pharmakophordefinition.

5.3 Screening makro- und molekularer Zustände

Um verschiedene Protein- und Molekülzustände während der Screening-Phase zu betrachten, wurde der grundlegende Ablauf wie folgt modifiziert: Die Präparierungsphase generiert zunächst einen Grundzustand für jedes Molekül. Lediglich dieser Zustand wird in der Moleküldatenbank registriert und dafür verwendet, Konformere zu generieren. Diese normalisierten Konformere dienen als Ausgangspunkt, um RAISE-Multizustandsdeskriptoren zu erzeugen. Jeder dieser Deskriptoren entspricht einem bestimmten Molekülzustand. Die Molekküldeskriptoren werden in die Indexstruktur überführt. Die Screening-Phase erzeugt initial einen Grundzustand der gegebenen aktiven Bindetasche, von dem ebenso RAISE-Multizustandsdeskriptoren abgeleitet werden. Jeder dieser Deskriptoren spiegelt einen Zustand des Proteins wider. Unter Verwendung der zuvor erstellten Indexstruktur werden die Anfragedeskriptoren dazu genutzt, um passende Molekküldeskriptoren zu identifizieren. Der Abgleich von Multizustandsdeskriptoren selektiert implizit den passenden Zustand von Protein und Ligand. Die Konformere der Treffer werden aus der Moleküldatenbank extrahiert, im Grundzustand des Moleküls initialisiert und mittels Deskriptorüberlagerung in die aktive Bindetasche platziert. Im Beisein beider Bindungspartner werden die Protein- und Ligandzustände explizit adaptiert. Wasserstoffe werden entfernt, hinzugefügt und ausgerichtet, um das Wasserstoffbrückennetzwerk des erzeugten Komplexes zu optimieren. Mit explizit

adaptierten Wasserstoffkoordinaten und Bindungstypen wird jede Pose eines Moleküls individuell bewertet und die Beste wird der Trefferliste des VS-Laufes hinzugefügt. Die Präparierungs- und Screening-Phasen unter Berücksichtigung molekularer und makromolekularer Zustände, sind in Abbildung 5.3 skizziert.

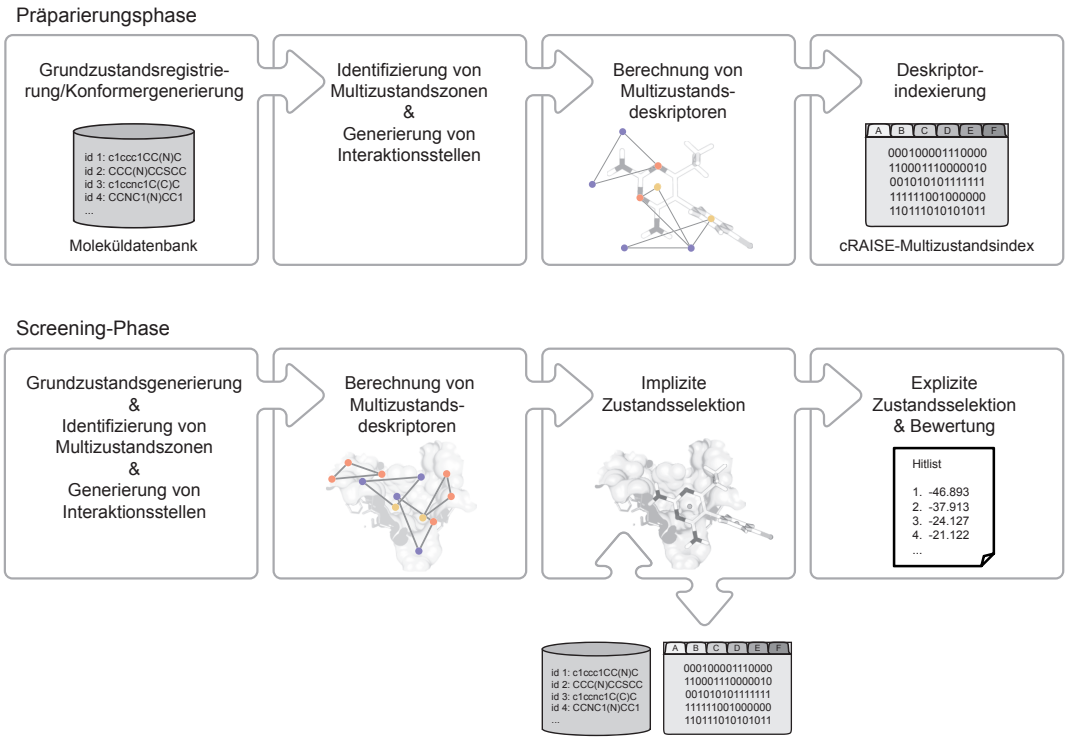


Abbildung 5.3: Präparierungs- und Screening-Phase unter Berücksichtigung molekularer und makromolekularer Zustände

Zur Generierung von Multizustandsdeskriptoren identifiziert cRAISE Molekül- und Proteinbereiche, die verschiedene Zustände einnehmen können. Die Funktionalität hierfür beruht auf dem in Abschnitt 4.2 vorgestellten VSC-Modell und ist in der Naomi-Bibliothek bereitgestellt. Zur expliziten Zustandsselektion ist Protoss (vgl. Abschnitt 4.3) in die Posenprozessierung von cRAISE integriert. Beide Komponenten sind in cRAISE bezüglich der Enumerierung möglichst stabiler Tautomere und Protonierungszustände konfiguriert. Das Multizustands-Docking von cRAISE ist unabhängig von gegebenen molekularen und makromolekularen Eingabezuständen. Dagegen ist die Detektion der Moleküleigenschaften und die Konformergenerierung jedoch vom gegebenen Zustand beeinflusst. Dies macht es notwendig, Moleküle vor ihrem Import in die Moleküldatenbank zu normalisieren. Das Erzeugen eines wahrscheinlichen molekularen

Grundzustands wird durch Naomi ermöglicht. Protoss wird zur Normalisierung der aktiven Bindetasche genutzt. An dieser Stelle hat es vor allem die Aufgabe strukturelle Ungereimtheiten wie Amid-Flips aufzulösen, die cRAISE nicht als echte Zuständsänderungen erachtet und deswegen nicht als zusätzliche Freiheitsgrade modelliert.

6 Modelle

Dieses Kapitel befasst sich mit der Modellierung der cRAISE-Arbeitsabläufe und stellt in den Abschnitten

- Das cRAISE-Interaktionsmodell
- Potentielle Interaktionsstellen
- Der RAISE-Deskriptor
- Molekülpräparierung
- Rezeptorpräparierung
- Berechnung von Molekül- und Proteindeskriptoren
- Molekül- und Konformerverwaltung
- Der Deskriptorindex
- Der indexbasierte Deskriptorabgleich
- Die cRAISE-Bewertungsfunktion
- Die Bewertungshierarchie
- Parallelisierung

zunächst die Basiskomponenten vor. Zur Modellierung von Randbedingungen wurden einzelne dieser Komponenten angepasst. Die Modifizierungen sind in den Abschnitten

- Molekülprofile
- cRAISE-Pharmakophorhypothesen

beschrieben. Das Kapitel schließt mit dem Abschnitt

- Modellierung makro-/molekularer Zustände

das auf die Erweiterung des Interaktionsmodells, die Berechnung von Multizustandsdeskriptoren, die Etablierung von Grundzuständen, den Deskriptorabgleich mit Multizustandsdeskriptoren und auf die Anpassung der Bewertungshierarchie zur Modellierung von molekularer und makromolekularer Zustandswechsel eingeht.

6.1 Das cRAISE-Interaktionsmodell

Ein Interaktionsmodell definiert unter welchen Bedingungen eine Interaktion zwischen Protein und Ligand detektiert wird. Innerhalb von cRAISE bildet das Modell die Grundlage zur Deskriptorberechnung, des Deskriptorabgleichs und der Bewertung von Posen. Es dient dabei zur systematischen Einschränkung des Suchraums während der Molekülplatzierung und zur groben Einschätzung der Bindungsaffinität in Abhängigkeit zur Lage des Moleküls. In beiden Fällen wird durch einen paarweisen Abgleich potentieller Interaktionsstellen von Ligand und Protein zwei Arten von Interaktionen identifiziert: gerichtete, hydrophile Interaktionen und ungerichtete, hydrophobe Kontakte (vgl. Abbildung 6.1). Der Abgleich bewertet die Typen, die Distanz und die Orientie-

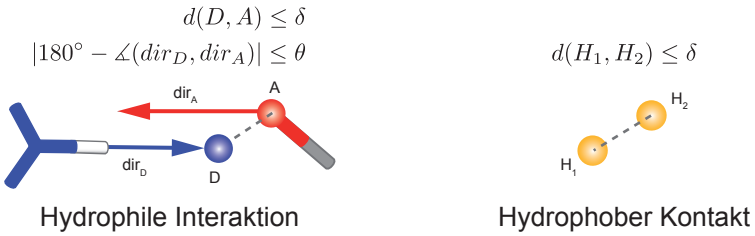


Abbildung 6.1: Identifizierung hydrophiler und hydrophober Interaktionen.

rung zweier Interaktionsstellen. Sind ihre Typen komplementär, kommen sie annähernd räumlich zur Deckung und besitzen entgegengesetzt orientierte Richtungen, so wird eine Interaktion erkannt. Um eine hydrophile Interaktion zu detektieren, muss eine Interaktionsstelle vom Typ *Donor* auf eine Interaktionsstelle vom Typ *Akzeptor* treffen. Bei hydrophoben Kontakten müssen beide Stellen den Typ *Hydrophob* aufweisen. Da hydrophobe Kontakte ungerichtet sind, entfällt bei ihrer Detektion ein Richtungsabgleich.

Das Interaktionsmodell von cRAISE lehnt sich stark an das in Kapitel 4.6 vorgestellte Modell der TrixX-Versionen an, das hydrophile Interaktionen und hydrophobe Kontakte zunächst nicht weiter spezifiziert. Hydrophile Interaktionen umfassen so nicht nur Wasserstoffbrücken, sondern auch Salzbrücken und Interaktionen mit Metallen (vgl. Abschnitt 2.4). Hydrophobe Kontakte umfassen zudem auch aromatische Wechselwirkungen. Ein wesentlicher Unterschied besteht jedoch in der Lage potentieller Interaktionsstellen. Während TrixX auf Ligandseite die Interaktionsstellen auf Interaktionszentren platzierte, die im Falle von Donoren, der Wasserstoffkoordinate und im Fall von Akzeptoren, der assoziierten Schweratomkoordinate entsprachen, wurden auf Rezeptorseite die Interaktionsstellen stets auf die klassischen FlexX-Interaktionsoberflächen

platziert. Bei einem Abgleich der Interaktionsstellen konnten so gewinkelte, hydrophile Interaktionen entstehen. cRAISE forciert mit seiner Modellierung dagegen lineare Interaktionsgeometrien wie sie für Wasserstoffbrücken typisch sind. Rotationssymmetrische Abweichungen von dieser Idealgeometrie werden durch Abweichung im Richtungsabgleich implizit modelliert.

6.2 Potentielle Interaktionsstellen

Potentielle Interaktionsstellen (vgl. Abbildung 6.2) weisen einen von drei möglichen Typen auf: *Donor*, *Akzeptor* oder *Hydrophob*. Akzeptorstellen sind stets auf dem assoziierten Schweratomzentrum eines Wasserstoffbrückenakzeptors lokalisiert. Donorstellen verweilen auf einer vom Schweratomzentrum eines Wasserstoffbrückendonors entfernten Position, die das Schweratom eines komplementären Interaktionspartners erwartet. Akzeptorstellen können mehrere Interaktionsrichtungen besitzen. Sie reflektieren die Ausrichtung freier Elektronenpaare. Donorstellen besitzen stets nur eine Interaktionsrichtung, die der Orientierung eines Wasserstoffes entspricht. Hydrophobe Interaktionsstellen sind im Fall kleiner Moleküle entlang aliphatischer Ketten auf Schweratomen, auf Bindungen und auf aliphatischen und aromatischen Ringzentren positioniert. Im Fall des Proteins sind sie im Volumen der aktiven Bindetasche zu finden.

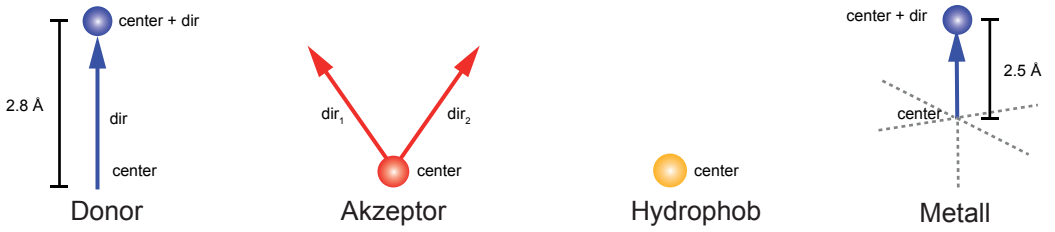


Abbildung 6.2: Arten potentieller Interaktionsstellen.

Potentielle Interaktionsstellen werden direkt oder indirekt anhand ein oder mehrerer Atome eines voll initialisierten NAOMI-Moleküls oder Proteins abgeleitet (vgl. Kapitel 4.1). Hierbei wird auf die chemische Information der Atomtypen zurückgegriffen.

Donorinteraktionsstellen werden bei Atomtypen, die das Atom als potentiellen Wasserstoffbrückendonor kennzeichnen, generiert. Sie werden 2,8 Å entfernt vom Schweratomzentrum entlang einer idealisierten Wasserstoffbrücke platziert und besitzen eine assoziierte Interaktionsrichtung, die der gegebenen oder gegebenenfalls neu berechneten Wasserstofforientierung entspricht. Welche und wie viele Donorstellen ein Atom ableitet,

ist bei hinzugefügten Wasserstoffen von der VSEPR-Geometrie (linear, trigonal-planar, tetrahedral) und der Anzahl gebundener Schweratomnachbarn abhängig.

Metalldonoren sind ausschließlich auf Rezeptorseite zu finden, da das Naomi-Modell nur die Initialisierung einfacher Metallionen unterstützt. In Metalloproteinen koordinieren zum Ion benachbarte Akzeptoren das Metall derart, dass eine wohldefinierte Geometrie etabliert wird (tetrahedral, trigonal-bipyrimidal oder oktahedral). Freie Koordinationsstellen sind für den Ligand zugänglich. Für sie werden für alle Geometrien, die die Anordnung benachbarter Akzeptoren erklären könnten, jeweils eine Donorstelle 2,5 Å entfernt vom Zentrum des Ions erzeugt. Die Interaktionsrichtung entspricht der Richtung zur freien Koordinationsstelle.

Akzeptorinteraktionsstellen werden bei Atomtypen, die das Atom als potentiellen Wasserstoffbrückenakzeptor auszeichnen, generiert. Für jeden Akzeptor wird genau eine Interaktionsstelle erzeugt, die auf dem Zentrum des assoziierten Schweratoms verbleibt. Sie kann eine oder mehrere assoziierte Interaktionsrichtungen besitzen, die idealen Orientierungen freier Elektronenpaare entsprechen. Welche und wie viele Interaktionsrichtungen ein Akzeptoratom ableitet, ist von dessen VSEPR-Geometrie und der Anzahl gebundener Schwer- und Wasserstoffatomnachbarn abhängig.

Hydrophobe Interaktionsstellen kleiner Moleküle werden wie folgt identifiziert:

1. Eine Stelle wird auf ein Schweratomzentrum platziert, wenn das Atom
   - den Atomtyp *C400* und drei oder mehr Kohlenstoff-, Halogen- und Wasserstoffnachbarn besitzt
   - den Atomtyp *C210* und drei Kohlenstoff- und Wasserstoffnachbarn besitzt
   - den Atomtyp *C101* besitzt
   - das Element Schwefel oder ein Halogen ist
2. Bei Ringatomen werden die Stellen durch eine Interaktionsstelle im Ringzentrum ersetzt, wenn der Ring maximal neun Atome besitzt oder aromatisch ist.
3. Benachbarte Stellen, die durch eine Bindung verbunden sind, werden durch eine hydrophobe Interaktionsstelle in der Mitte der Bindung ersetzt.
4. Isolierte Stellen verbleiben auf den Atomzentren.

Durch diese Vorgehensweise werden hydrophobe Merkmale ähnlich zu den gängiger Pharmakophormethoden etabliert (vgl. Kapitel 3.4.1).

Hydrophobe Interaktionsstellen der aktiven Bindetasche werden dadurch erhalten, dass die aktive Bindetasche mit Methylrepräsentanten sondiert wird. Ein Methylrepräsentant ist durch die Sphäre eines Kohlenstoffatoms mit einem Radius von 1,7 Å charakterisiert. Die Sonden werden auf die Gitterpunkte eines regulären, dreidimensionalen

Gitters mit einer Voxelgröße von 1,2 Å gelegt und mit der direkten Proteinumgebung im Abstand von 6,5 Å bewertet. Jedes Atom der Umgebung wird mit der Sonde durch ein klassisches (12, 6)-Lennard-Jones-Potential bewertet. Die Summe der Bewertungen wird am Gitterpunkt annotiert. Atome, die kein Kohlenstoff und kein Schwefel eines Methionins sind, tragen dabei ausschließlich zur Repulsion bei. Gitterpunkte mit einer positiven Gesamtbewertung werden verworfen, da ein dort platziertes hydrophobes Molekülatom zu Überlappungen mit dem Protein führen würde. Die bestbewerteten Gitterpunkte werden zu hydrophoben Interaktionsstellen konvertiert. Die Anzahl der Stellen ist durch einen Hydrophobizitätswert der Bindetasche limitiert. Er wird aus der Summe der Kohlenstoffatome und Methionin-Schwefel der Bindetasche gebildet.

Rotierbare Interaktionsstellen sind Stellen frei rotierbarer, hydrophiler Interaktionsatome. Rotierbare Interaktionsatome sind durch den Atomtyp eines potentiellen Wasserstoffbrückendonors oder -akzeptors mit tetrahedraler VSEPR-Geometrie charakterisiert und besitzen lediglich einen Schweratomnachbarn. Der Kegel, der die freie Rotation um die Hauptachse beschreibt, wird bei Donoren durch sechs Interaktionsstellen und bei Akzeptoren durch sechs Interaktionsrichtungen diskretisiert (vgl. Abbildung 6.3). Im Vergleich zu TrixX-BMI sind frei rotierbare Wasserstoffbrückendonoren und -akzeptoren somit vollständig und explizit auf Protein- und Ligandseite modelliert.

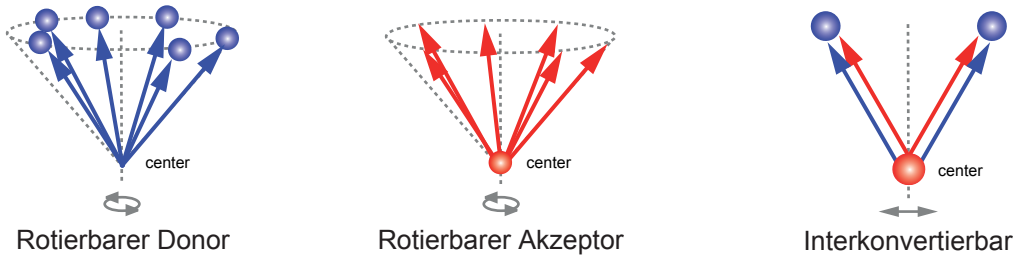


Abbildung 6.3: Frei rotierbare und interkonvertierbare Interaktionsstellen.

Interkonvertierbare Interaktionsstellen sind Stellen hydrophiler planarer sp^2- und tetrahedraler sp^3-Atome, die genau ein freies Wasserstoffatom und ein freies Elektronenpaar aufweisen. Die Ausrichtung des Wasserstoffs und des freien Elektronenpaars ist bei diesen Atomen nicht eindeutig. Sie können ihre Ausrichtung tauschen. Deshalb werden Donor- und Akzeptorstellen in beide Richtungen generiert (vgl. Abbildung 6.3).

6.3 Der RAISE-Deskriptor

In Analogie zum TrixX-BMI-Deskriptor spiegelt der RAISE-Deskriptor ein Drei-Punkt-Pharmakophor wider, das durch eine Zusammenstellung von drei potentiellen Interaktionsstellen eine Dreiecksbasis bildet. Er kodiert ebenso den Typ und die Seitenlängen der Dreiecksbasis, Interaktionsrichtungen und die Längen von 80 Bulk-Strahlen, die den Deskriptor mit einer sterischen Beschreibung der molekularen Form bzw. des aktiven Bindetaschenvolumens ausstatten. Abbildung 6.4 stellt beispielhaft einen RAISE-Molekülдeskriptor dar.

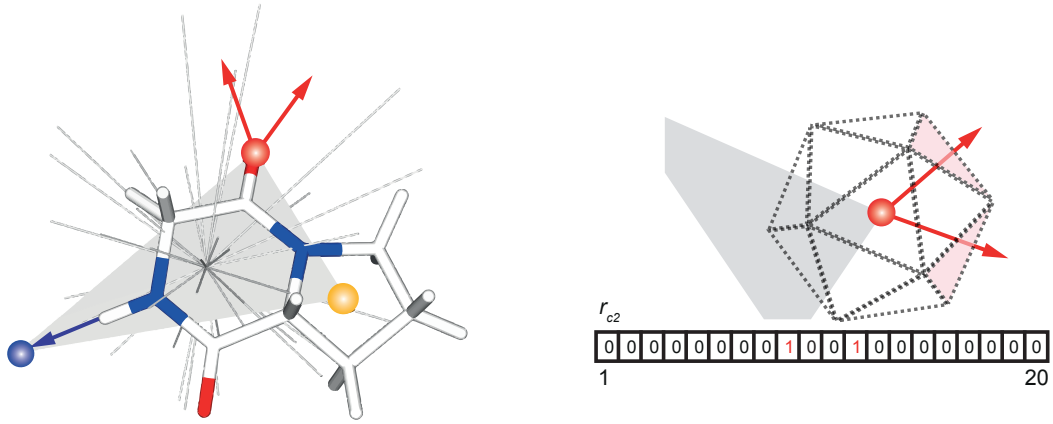


Abbildung 6.4: RAISE-Molekülдeskriptor und Kodierung von Interaktionsrichtungen.

Der RAISE-Deskriptor besitzt dieselben Eigenschaften wie sein Vorgänger, unterscheidet sich aber in der Kodierung der Deskriptoreigenschaften. Insbesondere wurde die Repräsentation von Interaktionsrichtungen abgewandelt, sodass nun mehr als drei Interaktionsrichtungen pro Deskriptor kodiert werden können.

Deskriptortyp: Der Deskriptortyp ist eine Zahl $t \in \{0, 1, \ldots, 8\}$. Sie kodiert die Zusammenstellung eines geordneten Tripletts von bis zu drei unterschiedlichen Typen potentieller Interaktionsstellen. Für die Ordnung der Interaktionsstellen gilt $D < A < H$, wobei D eine Interaktionsstelle vom Typ *Donor*, A eine Interaktionsstelle vom Typ *Akzeptor* und H eine Interaktionsstelle vom Typ *Hydrophob* bezeichnet. Damit ergeben sich prinzipiell zehn Deskriptortypen. cRAISE nutzt davon nur die neun Deskriptortypen, die in Tabelle 6.1 gelistet sind. Auf Deskriptoren vom Typ *HHH*, die einzig aus hydrophoben Interaktionsstellen resultieren, wird verzichtet. Dadurch werden Posen forciert, die zumindest eine hydrophile Interaktion etablieren.

Tabelle 6.1: Deskriptortypen und zugehörige, geordnete Tripletts.

Deskriptortypen				
0: DDD	2: DDH	4: DAH	6: AAA	8: AHH
1: DDA	3: DAA	5: DHH	7: AAH	

Seitenlängen: Für ein geordnetes Triplett (c_1, c_2, c_3) potentieller Interaktionsstellen werden die paarweisen euklidischen Distanzen $l_1(c_1, c_2)$, $l_2(c_2, c_3)$ und $l_3(c_1, c_3)$ im Deskriptor kodiert. Die Folge der Distanzen ergibt sich in erster Ordnung auf Basis des geordneten Tripletts. Bei Tripletts mit identischen Typen von Interaktionsstellen wird die Folge in zweiter Ordnung durch aufsteigende Distanzbeträge der Seiten definiert. Die Seitenlängen der Dreiecksbasis beschreiben die relative Anordnung der Interaktionsstellen im Raum. Dreiecke werden von cRAISE nur dann gebaut, wenn die Interaktionsstellen eines Tripletts paarweise mindestens eine Distanz von 1,0 Å und höchstens eine Distanz von 9,5 Å aufweisen.

Sterische Bulk-Beschreibung: Mit einem geordneten Triplett und geordneten Seitenlängen ist ein Bezugssystem geschaffen, mit dem sich ein lokales Koordinatensystem definieren lässt. Für die Erstellung der Bulk-Beschreibung wird ein mit 80 Flächen verfeinerter Ikosaeder an diesem Bezugssystem transformationsinvariant ausgerichtet. 80 Strahlen werden jeweils vom Ikosaederzentrum durch die Mitte einer Fläche gelegt und durch das Van-der-Waals-Volumen des Moleküls bzw. das innere Volumen der aktiven Bindetasche beschränkt, sodass ein Strahl mindestens eine Länge von 1,0 Å und höchstens von 7,1 Å einnimmt. Die Längen der beschränkten Strahlen werden im 80-dimensionalen Bulk-Vektor $(b_1, b_2, \ldots, b_{80})$ verzeichnet.

Interaktionsrichtungen: Der bereits transformationsinvariant ausgerichtete Ikosaeder wird zudem genutzt, um Interaktionsrichtungen in drei 20-dimensionalen Bit-Vektoren r_1, r_2 und r_3 zu kodieren (vgl. Abbildung 6.4). Der Ikosaeder wird auf jede Ecke der Dreiecksbasis gelegt. Eine Ikosaederoberfläche wird markiert und die Markierung im 20-dimensionalen Richtungsvektor $(r_1, r_2, \ldots, r_{20})$ der entsprechenden Ecke verzeichnet, wenn eine Interaktionsrichtung von der Ecke ausgehend auf das Oberflächendreieck zeigt.

Rekonstruktionsdaten: Zudem sind drei Identifizierer $m_{id} \in \mathbb{N}$, $c_{id} \in \mathbb{N}$, $d_{id} \in \mathbb{N}$ in Moleküldeskriptoren verzeichnet. Sie geben die Zugehörigkeit des Deskriptors zu einem bestimmten Molekül und einer bestimmten Konformation an und bestimmen seine Position innerhalb der von der Konformation abgeleiteten Deskriptorfolge. Die Information wird zur Rekonstruktion von Posen aus Deskriptortreffern genutzt.

6.4 Molekülpräparierung

Da cRAISE tatsächlich ein starres Protein-Ligand-Docking (vgl. Abschnitt 3.3.2) realisiert, muss zur Berücksichtigung von Molekülflexibilität die gegebene Bibliothek präpariert und mit Konformeren angereichert werden. Zur Konformergenerierung nutzt cRAISE den in Abschnitt 4.4 vorgestellten Algorithmus von Confect. Zur Anwendung im großangelegten virtuellen Screening, müssen jedoch zwei entscheidende Faktoren bei der Generierung berücksichtigt werden:

1. Die Anzahl erzeugter Konformationen muss limitiert sein, um ein unkontrolliertes Wachsen der präparierten Molekülbibliothek für unerwartet große und flexible Moleküle zu vermeiden und eine effiziente Screening-Anwendung zu gewährleisten.[337]
2. Leichte Abweichungen von den optimalen Torsionen müssen toleriert werden, da Optima intramolekulare Kollisionen verursachen können, zu deren Vermeidung das Molekül auf eine eher suboptimale Torsion ausweicht. Zudem kann das Zielprotein eine etwas ungünstigere Konformation des Moleküls forcieren, wenn dadurch energetisch günstige Interaktionen gebildet werden können.[229]

Um diesen konträren Positionen gleichermaßen zu genügen, integriert cRAISE eine Confect-Variante. Sie erzeugt für kleine und starre Moleküle auch etwas suboptimale Konformationen, um im Docking die Chance für eine Passung in die Bindetasche zu erhöhen. Für eher große und flexible Moleküle durchmustert sie den Torsionsraum dagegen granulär und selektiert möglichst diverse Repräsentanten des Konformationsraums. Hierfür wurden verfeinerte Abstufungen der Confect-Qualitätsstufen eingeführt, aus denen für jedes gegebene Molekül zunächst individuell die passende Stufe gewählt wird. Die Stufen erweitern (Auto Extend 1–5), reduzieren (Auto Reduce 1–3) bzw. erhalten (Normal) den durch Histogramm-Peaks definierten Torsionsraum ausgewählter Bindungen. Erweiterungen finden über die Hinzunahme von Winkelabweichungen des zweiten Toleranzniveaus statt. Reduzierungen vergröbern die Diskretisierung der Rotation einer Bindung durch sechs und gegebenenfalls durch drei gleichmäßig verteilte Torsionswinkel. Für welche Bindungen die Anzahl der Torsionen erhöht bzw. reduziert wird, hängt von der Anzahl der Peaks $|P|$ im Histogramm ihrer zugewiesenen Signatur ab. Tabelle 6.2 listet die Qualitätsstufen, die Torsionen und die Bedingungen für eine Erweiterung bzw. Reduzierung von Torsionen auf. Bindungen, für die $|P|$ unbeschränkt bleibt, werden gemäß der Histogramm-Peaks der Signatur eingestellt.

Für ein gegebenes Limit maximal zu generierender Konformationen wird molekülspezifisch die passende Qualitätsstufe gewählt, um den Raum möglichst gut abzutasten.

Tabelle 6.2: cRAISE-Qualitätsstufen: Bindungen mit Histogrammen mit $|P|$ Peaks p werden durch weitere Torsionen $p \pm \delta_2$ des zweiten Toleranzniveaus in den Qualitätsstufen Auto Extend 1–5 erweitert. Die Qualitätsstufen Auto Reduce 1–3 reduzieren eine Rotation frei rotierbarer ($|P| = 12$) und nahezu rotierbarer Bindungen ($6 \leq |P| < 12$) durch sechs bzw. drei gleichmäßig verteilte Torsionswinkel.

Qualitätsstufe	Durchmusterte Torsionen	Bedingung
Auto Extend 5	p, $p \pm \delta_2$	$\|P\| \leq 5$
Auto Extend 4	p, $p \pm \delta_2$	$\|P\| \leq 4$
Auto Extend 3	p, $p \pm \delta_2$	$\|P\| \leq 3$
Auto Extend 2	p, $p \pm \delta_2$	$\|P\| \leq 2$
Auto Extend 1	p, $p \pm \delta_2$	$\|P\| \leq 1$
Normal	p	—
Auto Reduce 1	$\frac{\pi}{3}, \frac{2\pi}{3}, \ldots, 2\pi$	$\|P\| = 12$
Auto Reduce 2	$\frac{\pi}{6}, \frac{\pi}{3}, \ldots, 2\pi$	$\|P\| = 12$
	$\frac{2\pi}{3}, \frac{4\pi}{3}, 2\pi$	$6 \leq \|P\| < 12$
Auto Reduce 3	$\frac{2\pi}{3}, \frac{4\pi}{3}, 2\pi$	$6 \leq \|P\| \leq 12$

Die Vorgehensweise zur Konformergenerierung ist in Algorithmus 6.1 beschrieben. Ausgehend von der feinsten zur granulärsten Qualitätsstufe werden die Stufen hierarchisch durchlaufen und bestimmt, ob die theoretisch mögliche Anzahl an Konformationen, die damit erzeugt werden würde, die Grenze von 50 000 Konformationen nicht übersteigt. Ist dies der Fall, werden Konformationen entsprechend dieser Stufe generiert. Andernfalls wird eine gröbere Qualitätsstufe gewählt, um den Konformationsraum weiter zu reduzieren. Die Anzahl der Konformationen lässt sich mit Gleichung 4.4 und 4.5 berechnen. Ist die optimale Qualitätsstufe gewählt, wird der Komponentenbaum entsprechend der Torsionen dieser Qualitätsstufe erweitert. Für wirkstoffähnliche Moleküle, die keine rotierbaren Bindungen besitzen, übersteigt die Anzahl theoretisch möglicher Konformationen die Grenze von 50 000 Konformationen in der Regel nicht. Größere Moleküle mit rotierbaren Bindungen, die diese Grenze überschreiten, werden reduziert. Zuletzt wird die Anzahl erzeugter Konformationen ohne Kollisionen bezüglich des TFD[329] mit einem k-Medoid-Algorithmus geclustert bis die gewünschte Anzahl k an repräsentativen Konformationen erreicht wurde. Das Clustering ist notwendig, da Kollisionen bei Wahl der Qualitätsstufe noch nicht absehbar sind. Es stellt sicher, dass Bereiche des Konformationsraum, die ungehinderte, kollisionsfreie Rotierungen flexibler Molekülteile darstellen und gegebenenfalls zuvor zu fein diskretisiert wurden, durch granuläre Repräsentanten beschrieben werden.

Algorithmus 6.1 : cRAISE-Konformergenerierung

Eingabe : Molekül m, Limit k
Ausgabe : Konformationen C
`ErzeugeKonformationen(`m, k`)`
$Q \leftarrow$ (Auto Extend 5, Auto Extend 4, ..., Auto Reduce 3)
$q \leftarrow$ Auto Extend 5
for $r \in Q$ **do**
  $n \leftarrow$ berechne Anzahl Konformationen von m in Stufe r
  if $n \leq 50\,000$ **then**
    $q \leftarrow r$
    break
$C \leftarrow$ erzeuge Confect-Konformationen von m in Qualitätsstufe q
$C \leftarrow k$-Medoid-TFD-Clustering(C)
return C

6.5 Rezeptorpräparierung

Für ein virtuelles Screening mit cRAISE müssen eine Proteinstruktur und ein Referenzligand bereitgestellt sein und ein aktiver Radius r angegeben werden. Mit dieser Information werden die Atome der Proteinstruktur wie folgt klassifiziert:

Bulk-Atome: Die gegebene Proteinstruktur ist nach Initialisierung in Form eines Naomi-Komplexes repräsentiert, der das Protein selbst und gegebenenfalls Kofaktoren, Metalle und Wassermoleküle enthält (vgl. Abschnitt 4.1.4). Enthalten die Moleküle des Komplexes bereits den Referenzliganden wird dieser zunächst aus dem Komplex entfernt. Ebenso werden Wassermoleküle entfernt wenn sie nicht zumindest drei Kontakte zum Protein etablieren. Die verbleibenden Wassermoleküle werden als essentiell betrachtet, die bei Ligandbindung nicht aus der Bindetasche verdrängt werden, formgebend und an der Bindung beteiligt sein können. Alle verbleibenden Schweratome werden als Bulk-relevante Atome klassifiziert. Für sie werden Wasserstoffkoordinaten gemäß der am Atomtyp annotierten VSEPR-Geometrie berechnet, um etwaige untypische Bindungslängen und Orientierung zu beheben und anschließend mit Protoss optimiert.

Aktive Atome: Aktive Atome sind eine echte Teilmenge der Bulk-Atome. Zu ihrer Identifizierung wird auf jedes Schweratomzentrum des Referenzliganden eine Sphäre mit aktivem Radius r gelegt. Die Bulk-Atome, deren Zentren sich zumindest innerhalb einer dieser Sphären befinden, bilden die Menge der aktiven Atome, d. h. die aktive Bindetasche des Proteins.

Zugängliche Atome: Zugängliche Atome sind eine echte Teilmenge der aktiven Atome. Zu ihrer Bestimmung wird über die Vereinigung der aktiven Atome eine Kugel mit einem Radius von 1,4 Å abgerollt. Die Kugel repräsentiert hierbei das Lösungsmittel. Die Abrolloberfläche (Connolly-Oberfläche[338]) besteht aus den konvexen Kontaktflächen der Abrollkugel mit der Van-der-Waals-Oberfläche des Proteins und den konkaven Flächen, die durch den zeitgleichen Kontakt der Abrollkugel mit zwei oder drei Atomen entstehen. Die Abrolloberfläche wird trianguliert, indem die Van-der-Waals-Oberflächen jedes Atoms mit einem zweimalig verfeinerten Ikosaeder diskretisiert werden. Aktive Atome, die einen Beitrag zur Abrolloberfläche leisten, d. h. die einen Oberflächenpunkt besitzen, werden als zugänglich charakterisiert.

Die verschiedenen Klassen der Atome sind im cRAISE-Rezeptor zusammengefasst. Sie werden zur Bestimmung potentieller Interaktionsstellen zur Berechnung von Anfragedeskriptoren und zur Beschränkung von Deskriptorstrahlen genutzt. Zudem dienen sie zur Detektion von Überlappungen und zur Vorberechnung von Bewertungsinformation:

Aktive konvexe Hülle: Für die aktiven Atome wird eine konvexe Hülle berechnet. Dafür wird zunächst ein mit 80 Strahlen verfeinerter Ikosaeder an jedem Atomzentrum ausgerichtet. Für die Schnittpunkte der Strahlen mit der Van-der-Waals-Oberfläche berechnet der in Algorithmus 6.2 beschriebene Quickhull Algorithmus von Barber et al.[339] einen minimalen Polyeder, der alle Punkte umhüllt. Der Algorithmus ist ein „teile-und-herrsche"-Verfahren, das die gegebene Punktwolke iterativ in eine äußere und innere Punktmenge durch die Facetten eines stetig wachsenden Polyeders unterteilt. Im dreidimensionalen Fall wird zu Beginn ein maximaler Tetraeder initialisiert, dessen Basisdreieck anhand der sechs Extrempunkte mit minimaler und maximaler x-, y- und z-Ausrichtung konstruiert wird. Die Tetraederspitze wird durch den zum Basisdreieck entferntesten Punkt gebildet. Initial wird jeder Punkt p einmalig einer Facette f des maximalen Tetraeders zugewiesen, wenn der Punkt für die Facette sichtbar ist:

Definition 6.5.1 (Sichtbarkeit eines Punktes)**.** *Ein Punkt p ist für die Facette f eines Polyeders sichtbar, wenn der Winkel zwischen den Einheitsvektoren, der nach außen gerichteten Facettennormalen $\overrightarrow{n_f}$ und dem Vektor $\overrightarrow{zp}$ zwischen Facettenzentrum z und Punkt p 90° nicht übersteigt. D. h. wenn für das Skalarprodukt beider Vektoren gilt:*

$$\overrightarrow{n_f} \cdot \overrightarrow{zp} > 0 \tag{6.1}$$

Der Polyeder wird in jedem Schritt mit dem entferntesten, äußeren Punkt erweitert, indem ein Polyeder konstruiert wird, der den bestehenden ergänzt. Der neue Polyeder unterteilt die äußere Punktwolke weiter und reduziert sie zugleich. Die Iteration wird

so lange fortgesetzt bis alle Punkte entweder Bestandteil des Polyeders sind oder in ihm enthalten sind. Das Resultat ist eine Menge von Facetten, die die triangulierte Oberfläche der konvexen Hülle repräsentiert. Sie wird in der Rezeptorstruktur registriert.

Algorithmus 6.2 : QuickHull-Algorithmus nach Barber[339]

Eingabe : Punkte $P \in \mathbb{R}^d$
Ausgabe : Facetten F der konvexen Hülle
`QuickHull(`P`)`
$F \leftarrow$ erzeuge maximalen Simplex aus $d+1$ Punkten
for *Facette* $f \in F$ **do**
  for *nicht hinzugefügten Punkt* $p \in P$ **do**
    if p *sichtbar für* f **then**
      füge p zu f's äußerer Punktmenge hinzu
while *Facette* $f \in F$ *mit nichtleerer äußerer Punktmenge* **do**
  $p \leftarrow$ wähle entferntesten Punkt aus f's äußerer Punktmenge
  $V \leftarrow$ identifiziere für p sichtbare Facetten
  $H \leftarrow$ bestimme äußere Kanten von V (Horizont)
  $G \leftarrow \emptyset$
  for *Horizontkante* $h \in H$ **do**
    $g \leftarrow$ neue Facette, die von Kante h zu Punkt p gebildet wird
    $G \leftarrow G \cup \{g\}$
  for *umschlossene Facette* $v \in V$ **do**
    $F \leftarrow F \setminus \{v\}$
    for *neue Facette* $g \in G$ **do**
      for *nicht hinzugefügten Punkt* p *von* v*'s äußerer Punktmenge* **do**
        if p *sichtbar für* g **then**
          füge p zu g's äußerer Punktmenge hinzu
  $F \leftarrow F \cup G$
return F

Überlappgitter: Um während der Posenprozessierung nicht wiederholt Proteinatome mit Posenatomen überprüfen zu müssen und eine rasche Überlappdetektion zu ermöglichen, werden Regionen für potentielle Überlappungen vorab identifiziert. Dafür wird ein dreidimensionales, kartesisches Gitter über die Proteinatome gelegt. Die Koordinaten der aktiven Atome mit minimaler und maximaler x-, y- und z-Auslenkung definieren hierbei die Ausdehnung des Gitters. Jeder einzelne Gitterpunkt wird mit einem hydrophilen und hydrophoben Atomrepräsentanten $P \in \{P_{\text{hphil}}, P_{\text{hphob}}\}$ sondiert. P_{hphil} ist eine Sphäre mit einem Radius von 1,5 Å, der typischerweise dem Van-der-Waals-Radius von Stick- und Sauerstoffatomen entspricht. Eine hydrophobe Sonde P_{hphob} ist durch

einen Radius von 1,7 Å charakterisiert, dem Van-der-Waals-Radius von Kohlenstoffatomen. Jede positionierte Sonde wird mit den Proteinatomen in Kontaktdistanz auf Überlappungen überprüft. Eine Überlappung wird festgestellt und am Gitterpunkt für den hydrophilen oder hydrophoben Atomtyp annotiert, sobald die Sonde das Zentrum eines Proteinatoms umschließt. Am Rand des Gitters werden nicht nur aktive Atome, sondern auch weiter entfernte Bulk-Atome ausgewertet, um vollständig die durch das Protein belegten Bereiche zu identifizieren. Die Voxelgröße des Gitters ist mit $a = 0,25$ Å derart gewählt, dass bei späteren Anfragen an einen Gitterpunkt Distanzabweichungen bis maximal $\frac{a}{2}\sqrt{3} = 0,22$ Å auftreten können. Um diese Abweichungen abzufangen und eine fälschliche Detektion von Überlappungen zu vermeiden, werden die van-der-Waals-Radien der Proteinatome um diesen Betrag reduziert, d. h. das Überlappgitter wird auf Basis eines weich modellierten Proteins berechnet. Das mit Überlappinformation annotierte Gitter wird in der cRAISE-Rezeptorstruktur vermerkt.

Bewertungsgitter: Die initiale Bewertungsphase von cRAISE beruht auf der Annahme, dass unterschiedliche Posenatome vom selben Typ, die wiederholt an dieselbe Stelle in der aktiven Bindetasche gelegt werden, auch stets identisch bewertet werden. Solche Ereignisse treten während eines virtuellen Screenings häufig auf. Um in solchen Fällen die selben Proteinatome nicht wiederholt auswerten zu müssen, werden Bewertungsbeiträge an jeder Stelle der Bindetasche einmalig vorberechnet und persistent zur Laufzeit vorgehalten. Dafür werden die Atomrepräsentanten herangezogen, die auch zur vorzeitigen Überlappdetektion genutzt werden. Die hydrophilen und hydrophoben Sonden werden mit den zugänglichen Oberflächenatomen der Bindetasche bewertet und individuelle Bewertungsbeiträge werden am Gitterpunkt vermerkt. Dies sind Lennard-Jones-ähnliche Atompaarpotentialwerte und auf die Sonden gerichtete Proteininteraktionsrichtungen. Das mit der Bewertungsinformation versehene Bewertungsgitter ist ebenfalls in der Rezeptorstruktur verwaltet. Die einzelnen Beiträge werden während der frühen Bewertungsphase abgefragt (vgl. Abschnitt 6.11), um Posen gemäß der in Abschnitt 6.10 vorgestellten cRAISE-Funktion zu bewerten.

6.6 Berechnung von Molekül- und Proteindeskriptoren

Auf Basis vorberechneter Konformere und der Rezeptorstruktur werden Molekül- bzw. Proteindeskriptoren wie folgt berechnet:

Identifizierung potentieller Interaktionsstellen: Interaktionsstellen werden wie in Abschnitt 6.2 beschrieben für die gegebenen Atome erstellt. Im Fall von Molekülen

werden sie für jedes Konformer erzeugt. Proteinseitig werden hydrophile Interaktionsstellen für zugängliche Atome erstellt. Hydrophile Interaktionsstellen werden eliminiert, wenn sie zu intramolekularen Interaktionen führen würden. Für das Protein werden Interaktionsstellen auch entfernt, wenn ihre Interaktionsrichtungen in andere Proteinatome hinein oder aus dem Volumen der aktiven Bindetasche heraus ragen. Typen und Richtungen der Interaktionsstellen des Proteins werden invertiert, sodass während der Screening-Phase ein Abgleich von Deskriptortypen und Interaktionsrichtungen stattfinden kann.

Berechnung der Dreiecksbasis: Über die Interaktionsstellen werden Tripletts enumeriert und Basisdreiecke gebildet. Sie werden entsprechend der Typen ihrer Interaktionsstellen und der Seitenlängen kanonisiert. Bei Deskriptoren, die nicht paarweise verschiedene Typen von Interaktionsstellen besitzen und gleichschenklige Dreiecke sind, führt eine strenge kanonische Ordnung jedoch dazu, dass während des Deskriptorabgleichs mögliche Ligandplatzierungen übersehen werden. Deshalb wird für solche Proteindeskriptoren die Ordnung aufgegeben und alle möglichen Tripletts aufgezählt. Ein Deskriptorabgleich produziert dadurch alle möglichen Ausrichtungen des Liganden in der Bindetasche. Zudem sichern konstitutionelle und geometrische Kriterien während der Aufzählung zu, dass sinnvolle Drei-Punkt-Pharmakophore erhalten werden:

- Ein Triplett enthält mindestens eine hydrophile Interaktionsstelle.
- Ein Triplett geht nicht aus Interaktionsstellen des selben Atoms bzw. aus Atomen hervor, die weniger als zwei Bindungen voneinander entfernt sind.
- Seitenlängen der Dreiecksbasis dürfen die Minimal- bzw. Maximaldistanz von 1,0 und 9,5 Å nicht unter- bzw. überschreiten.
- Um nicht zu spitzwinklige Dreiecke zu erhalten, müssen die inneren Winkel der Dreiecksbasis mindestens einen Wert von 0,15 rad besitzen.
- Proteindeskriptoren müssen zumindest zwei Seiten aufweisen, die nicht durch Proteinatomsphären geschnitten werden. Der Schnitt einer Seite wird toleriert, da ein Molekül die dadurch beschriebene Proteinausstülpung umschließen könnte.

Berechnung der Bulk-Beschreibung: Die Berechnung des Bulks nutzt einen verfeinerten Ikosaeder, bei dem die 20 Flächen weiter in vier gleichseitige Dreiecke unterteilt sind. Ausgehend vom Ikosaederzentrum durchstoßen Strahlen jeweils das Zentrum eines der 80 Oberflächendreiecke. Diese Strahlen werden an der kanonisierten Dreiecksbasis definiert ausgerichtet: Das Zentrum des verfeinerten Ikosaeders wird auf das Zentrum der Dreiecksbasis positioniert. Der Ikosaeder wird dann rotiert, sodass sein erster Strahl durch die erste Ecke der Basis führt. Eine weitere Rotation legt den zweiten Strahl

in die Dreiecksebene, sodass die erste Dreiecksseite geschnitten wird. Die so orientierten Strahlen werden beschränkt. Auf Molekülseite werden die Atomsphären mit einem Strahl geschnitten, die den Strahl in Segmente unterteilen. Ist der Strahl nicht segmentiert so erhält er die Minimallänge eines Bulk-Strahls. Zerfällt er in mehrere Segmente, so beschränkt der entfernteste Schnittpunkt den Strahl (Ausgangsdistanz). Übersteigt diese die Maximallänge, so erhält er die Maximallänge. Auf Proteinseite wird für jedes aktive Atom bestimmt, ob es einen Bulk-Strahl schneidet. Ist dies der Fall, so wird der bislang zum Zentrum nächste Schnittpunkt ermittelt (Eingangsdistanz). Wird ein Strahl nicht geschnitten, so erhält er seine Maximallänge.

Berechnung assoziierter Interaktionsrichtungen: Jede hydrophile Ecke einer Dreiecksbasis wird mit den Interaktionsrichtungen der zugehörigen Interaktionsstelle ausgestattet. Für jede Ecke der Deskriptorbasis wird das Zentrum des bereits am Dreieck transformationsinvariant ausgerichteten Ikosaeders weiter auf die Ecke transformiert. Ausgehend vom Ikosaederzentrum durchstoßen nun 20 Strahlen jeweils das Zentrum eines der 20 Oberflächendreiecke. Für jede Interaktionsrichtung wird die Ikosaederoberfläche markiert und ein entsprechendes Bit im Richtungsvektor gesetzt, für die der Winkel zwischen Interaktionsrichtung und Strahl minimal ist.

Rekonstruktionsdaten: Da die Moleküle bei Registrierung kanonisiert werden, ist sichergestellt, dass die Folge der generierten Konformationen und Deskriptoren stets dieselbe ist. Jedem Moleküldeskriptor kann deshalb zur ID des Moleküls m_{id}, zusätzlich die ID der Konformation c_{id} und die ID des Moleküldeskriptors d_{id} in diesen Folgen eindeutig zugewiesen werden.

Tabelle 6.3 stellt zusammenfassend die wesentlichen Unterschiede bei der Berechnung von Molekül- und Proteindeskriptoren gegenüber.

Tabelle 6.3: Unterschiede bei der Berechnung von Molekül- und Proteindeskriptoren.

	Molekül	Protein
Interaktionsstellen	entsprechender Typ	invertierter Typ
Deskriptorbasis	kanonisiert	ggf. enumeriert
Richtungen	entsprechende Richtungen	invertierte Richtungen
Bulk	Ausgangsdistanz	Eingangsdistanz
Rekonstruktionsdaten	m_{id}, c_{id}, d_{id}	keine

6.7 Molekül- und Konformerverwaltung

Die Moleküldatenbank ist ein integraler Bestandteil der persistent gespeicherten Molekülinformation, die – einmalig in der Präparierungsphase erstellt – für verschiedene Screening-Anwendungen verwendet werden kann. Abbildung 6.5 zeigt das Schema der Moleküldatenbank. Ist eine Molekülbibliothek in Form ein oder mehrerer Multi-MOL2 oder Multi-SDF-Dateien gegeben, so wird ihr Inhalt zunächst in der Moleküldatenbank registriert. Aus den Einträgen werden sukzessiv Naomi-Moleküle aufgebaut und kanonisiert, die in Abschnitt 4.5 aufgelisteten Moleküleigenschaften werden berechnet und zusammen mit der Valenzstruktur in Form des Mol-Strings als einzelne Moleküleinträge in der Datenbank gespeichert. Die gegebene Konformation wird als Instanz der Eingabe markiert und ihre kartesischen Koordinaten in Form eines BLOBs als erster Konformationseintrag verzeichnet. Die Markierung der Konformation gewährleistet, dass die Initialkonformation später nicht wieder verwendet wird. Wird während der Molekülregistrierung ein Duplikat erkannt, so wird dessen Konformation ebenso als Eingabeinstanz markiert und zum bereits registrierten Molekül hinzugefügt. Das Duplikat wird dann von cRAISE nicht weiter verwendet.

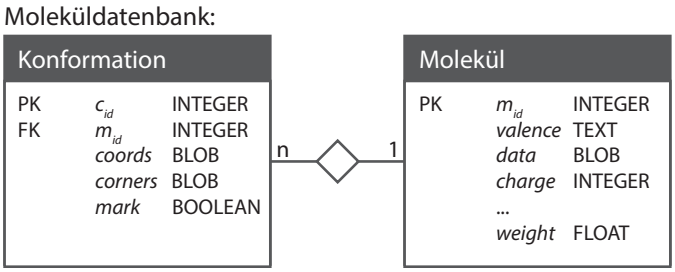


Abbildung 6.5: Schema der cRAISE-Moleküldatenbank

Für jedes registrierte Molekül wird die erste Eingabeinstanz genutzt, um Konformationen gemäß Abschnitt 6.4 zu berechnen. Jede erzeugte Konformation wird als generierte Instanz markiert und ihre Koordinaten als neuer Konformationseintrag des zugehörigen Moleküls gespeichert. Ist das Molekül starr, so wird die Initialkonformation beibehalten und die zugehörige Eingabeinstanz als generiert markiert. Für jede generierte Konformation eines Moleküls werden Moleküldeskriptoren gemäß Abschnitt 6.6 berechnet. Die Deskriptoren werden dem folgend beschriebenen Deskriptorindex hinzugefügt. Die Kanonisierung der Moleküle gewährleistet, dass sowohl die Folge der erzeugten Konformationen, als auch die Folge der erzeugten Deskriptoren eindeutig ist. Die definierte Folge der kartesischen Koordinaten aller Dreiecksbasen einer Konformation werden als komprimierter Daten-BLOB an ihren Konformationseintrag notiert.

6.8 Der Deskriptorindex

Der in cRAISE genutzte komprimierte Bitmap-Index ist ein FastBit-Index (vgl. Abschnitt 4.7). Er unterteilt die in ihm hinterlegten Deskriptoren vertikal, bezüglich einzelner Deskriptoreigenschaften. Für die kontinuierlichen Eigenschaften, d. h. für die Seitenlängen und die Längen der 80 Bulk-Strahlen, wird ein Binning-Schema verwendet. Die Bin-Grenzen sind dabei so definiert, dass einzelne Bins im ersten Fall einen Bereich von 0,1 Å und im zweiten Fall von 0,4 Å abdecken, sodass 85 Bits die Länge einer Seite bzw. 15 Bits die Länge eines Bulk-Strahls kodieren. Für Deskriptoreigenschaften, die diskrete Werte einnehmen, wird kein Binning-Schema verwendet. Dies betrifft die Kodierung der Interaktionsrichtungen, da sie mittels markierter Ikosaederflächen bereits granulär über 20 Bits repräsentiert sind. Über die Bit-kodierten Deskriptoreigenschaften wird eine Index-Struktur etabliert, wie sie schematisch in Abbildung 6.6 dargestellt ist. Für jede ihrer Dimensionen erstellt FastBit einen komprimierten Bitmap.

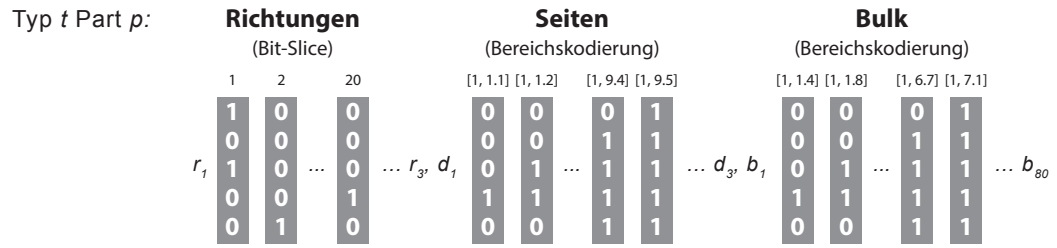


Abbildung 6.6: cRAISE-Subindex

Jedes Attribut des RAISE-Deskriptors wird während des Deskriptorabgleichs auf die Einhaltung einer Bedingung überprüft. Der sterische Bulk muss in einem bestimmten Bereich und die Seitenlänge in einem gewissen Intervall liegen. Der Typ wird auf Gleichheit und die Interaktionsrichtungen werden auf eine nicht-leere Schnittmenge überprüft. Damit diese unterschiedlichen Bedingungen effizient auf Basis von Bitmaps bewertet werden können, ist es notwendig die Bitmaps entsprechend zu kodieren (vgl. Abschnitt 4.7.3). Deshalb werden die Bitmaps der Seitenlängen und Bulk-Strahlen für Bereichsanfragen kodiert. Mit der Bereichskodierung kann FastBit auch effizient die Einhaltung eines Intervalls überprüfen, wie es für den Abgleich der Seitenlängen notwendig ist. Die 20 Bits von Interaktionsrichtungen werden bitweise vertikal unterteilt (Bit-Slice), sodass der nichtleere Schnitt direkt mittels einer Booleschen UND-Operation ausgewertet werden kann.

Die Granularität des Index ist über den Deskriptortyp $t \in \{0, 1, \ldots, 8\}$ erhöht. Er wird nicht explizit gespeichert, sondern unterteilt die Indexstruktur horizontal in Sub-

indizes. Jeder Subindex enthält somit ausschließlich Deskriptoren eines Typs. Dadurch muss bei einem Deskriptorabgleich der Typ nicht mehr konkret ausgewertet werden. Passende Deskriptortypen werden über eine Anfrage an den entsprechenden Subindex direkt erhalten. Während des Deskriptorabgleichs werden Bitmaps in den Hauptspeicher geladen und im FastBit-Cache vorgehalten. Um die Anzahl der Speicherseiten, die während dieses Vorgangs transferiert werden müssen, so gering wie möglich zu halten, wurde bei der Erstellung des Index darauf geachtet, dass ein Subindex vollständig im Speicher gehalten werden kann. Deshalb ist er darauf beschränkt maximal 6 500 000 Deskriptoreinträge zu enthalten. Wird dieses Limit überschritten, so wird ein neuer Subindex $p \in \mathbb{N}$ des entsprechenden Deskriptortyps t erstellt. FastBit ermöglicht es, den maximalen Speicherverbrauch zu beschränken, indem ein globales Cache-Limit angegeben wird. Mit dem Deskriptorlimit eines Subindizes ist gewährleistet, dass der FastBit-Cache nie mehr als 2 GB erfordert.

6.9 Der indexbasierte Deskriptorabgleich

Anfragedeskriptoren eines Proteins werden dazu genutzt, passende Moleküldeskriptoren aus dem Index zu extrahieren. Um dafür während einer Indexanfrage jeden Subindex nur einmalig in den FastBit-Cache laden zu müssen, werden die Anfragedeskriptoren zunächst entsprechend ihres Deskriptortyps sortiert und SQL-ähnliche Anfragen formuliert. Die Subindizes werden dann typweise prozessiert. Ist ein Subindex geladen, werden die Anfragen gestellt, die seinem Typ entsprechen. Algorithmus 6.3 skizziert die Vorgehensweise bei der Indexanfrage.

Algorithmus 6.3 : cRAISE-Indexanfrage

Eingabe : Anfragedeskriptoren Q, Index I, Toleranzen Δ_l, Δ_b, Δ_r
Ausgabe : Treffer T
`QueryIndex(`Q, I, T, Δ_l, Δ_b, Δ_r`)`
for $t \in \{0, \ldots 8\}$ **do**
  $I(t) \leftarrow$ Subindizes vom Typ t
  $Q(t) \leftarrow$ Formuliere Anfragen vom Typ t
  for $p \in I(t)$ **do**
    for $q \in Q(t)$ **do**
      $T(q,p) \leftarrow \mathit{Treffer}(q, p, \Delta_l, \Delta_b, \Delta_r)$
      speichere $T(q,p)$
return T

Um zu entscheiden, ob ein Molekül in die aktive Bindetasche passt, werden Deskriptoreigenschaften miteinander verglichen. Nach Invertierung der Anfragedeskripto-

ren wird ein Deskriptortreffer bei gleichen Typen der Dreiecksbasen, ähnlich orientierten Interaktionsrichtungen, ähnlichen Seitenlängen und einer Inklusion aller 80 Ligandstrahlen durch ihre entsprechenden Bindetaschengegenstücke erkannt. Dieser Abgleich wird über eine Anfrage an den Index realisiert. Sei ein Anfragedeskriptor q gegeben, dann ist ein im Index I hinterlegter Moleküldeskriptor d genau dann ein Treffer, wenn folgende Bedingung erfüllt ist:

$$\begin{aligned} \mathit{Treffer}(q, \Delta_l, \Delta_b, \Delta_r) = \{d \in I \mid\ & t(q) = t(d) \\ & \wedge \bigwedge_{i=1}^{3} (((r_i(q) \cup \Delta_r) \cap r_i(d)) \neq \emptyset) \\ & \wedge \bigwedge_{j=1}^{3} (l_j(q) - \Delta_l \leq l_j(d) \leq l_j(q) + \Delta_l) \\ & \wedge \bigwedge_{k=1}^{80} (b_k(q) + \Delta_b \geq b_k(d))\} \end{aligned} \tag{6.2}$$

Hierbei bezeichnen Δ_l und Δ_b Toleranzen, die beim Abgleich von Seitenlängen und Bulk-Strahlen auf die entsprechende Deskriptoreigenschaft des Anfragedeskriptors addiert bzw. subtrahiert werden. Für cRAISE betragen sie respektive 1,0 Å und 0,5 Å. Die Toleranz Δ_r ist ein 20-dimensionaler Bitvektor, der für Anfragedeskriptoren ein Bit im Richtungsvektor setzt, sobald der Winkel zwischen einer Interaktionsrichtung und einem Ikosaederstrahl $\alpha \leq \frac{\pi}{4}$ ist.

Die Treffer, die aus einer Anfrage resultieren, werden in einer Treffertabelle hinterlegt. Sie registriert für jeden Treffer die ID des Moleküls m_{id}, die ID der Konformation c_{id}, die ID des Moleküldeskriptors d_{id} in der Deskriptorfolge der Konformation und die ID des Anfragedeskriptors q_{id}. Das Modell zur Speicherung der Treffer ist in Abbildung 6.7 dargestellt. Mit diesen Daten ist es möglich, vollständig Naomi-initialisierte

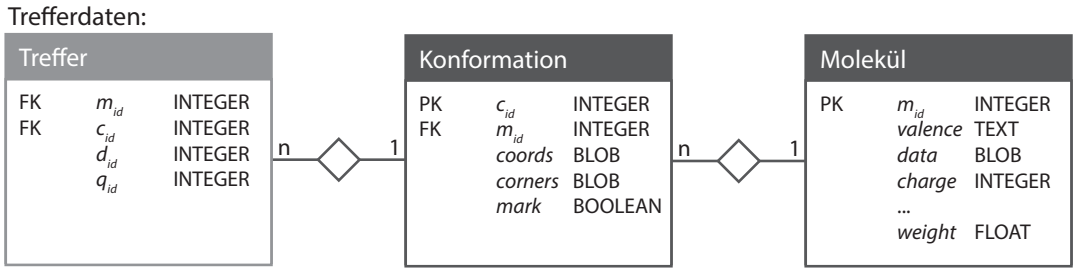


Abbildung 6.7: cRAISE-Treffertabelle, die aus einer Indexanfrage resultiert und zur Posengenerierung genutzt wird.

Posen zu erzeugen, d. h. eine dreidimensionale Struktur eines Moleküls in einer gewissen Lage und Orientierung zum Protein:

Molekülinitialisierung: m_{id} verweist auf einen Eintrag in der Molekül datenbank, der die kanonisierten Molekülzeichenkette zum Aufbau eines validen Naomi-Moleküls enthält (vgl. Abschnitt 4.5.2). Anhand dieser Information wird eine interne Molekülrepräsentation, d. h. der Molekülgraph mit ausgezeichneten Atom- und Bindungstypen aufgebaut.

Konformerinitialisierung: c_{id} verweist innerhalb eines Moleküleintrags auf eine Konformation, d. h. eine Folge von Atomkoordinaten. Sie können sukzessive auf das kanonisierte Molekül angewendet werden, um eine valide dreidimensionale Struktur des Moleküls zu erhalten.

Transformation: Für jede Konformation ist in der Konformationstabelle eine Folge der Dreieckskoordinaten komprimiert gespeichert. d_{id} verweist für eine Konformation auf ein bestimmtes Triplett in dieser Folge. q_{id} verweist auf ein bestimmtes Triplett von Dreieckskoordinaten innerhalb der Folge von Anfragedeskriptoren, die zur Laufzeit jederzeit zugänglich sind. Mit dem ausgezeichneten Dreieckspaar (d_{id}, q_{id}) wird eine affine Transformation berechnet, die das Moleküldreieck auf das Proteindreieck positioniert. Die Transformation wird auf die Konformation angewandt, um das Molekül in die aktive Bindetasche zu platzieren.

6.10 Die cRAISE-Bewertungsfunktion

cRAISE schätzt die Passung einer Pose durch eine empirische Bewertungsfunktion ab. Sie berücksichtigt enthalpische Beiträge von Wasserstoffbrücken, Metallinteraktionen, lipophile Kontakte und den Entropieverlust, der durch das Einfrieren von Ligandtorsionen während der Bindung resultiert:

$$\begin{aligned}\Delta G_{\mathrm{cRAISE}} = \Delta G_0 &+ \Delta G_{\mathrm{hb}} \sum_{\text{Donor, Akzeptor}} f_{\mathrm{hb}}(r, \alpha, \beta, *) \\ &+ \Delta G_{\mathrm{met}} \sum_{\text{Metall, Akzeptor}} f_{\mathrm{met}}(r, \alpha, \beta, *) \\ &+ \Delta G_{\mathrm{lipo}} \sum_{\text{Hphob, Hphob}} f_{\mathrm{lipo}}(r) \\ &+ \Delta G_{\mathrm{mis}} \sum_{\text{Hphil, Hphob}} f_{\mathrm{mis}}(r) \\ &+ \Delta G_{\mathrm{rot}} N_{\mathrm{rot}}\end{aligned} \tag{6.3}$$

Die Funktion ist eine Böhm-Bewertungsfunktion (vgl. Abschnitt 3.3.5), die anstatt stückweise linearer Bestrafungsfunktionen, Lennard-Jones-ähnliche Atompaarpotentiale $f_{\mathrm{hb}}(r, \alpha, \beta, *)$, $f_{\mathrm{met}}(r, \alpha, \beta, *)$, $f_{\mathrm{lipo}}(r)$ und $f_{\mathrm{mis}}(r)$ nutzt, um gute Atomkontakte zu honorieren und zu nahe zu bestrafen. N_{rot} zählt die Anzahl rotierbarer Ligandbindungen, die während der Komplexbildung immobilisiert werden. ΔG_0, ΔG_{hb}, ΔG_{met}, ΔG_{lipo} und ΔG_{rot} sind justierbare Parameter, die auf 78 Protein-Ligand-Kristallkomplexen des Iridium-Highly-Trustworthy-Datensatzes[340] einmalig kalibriert wurden. Für die publizierten K_i- bzw. K_d-Werte wurden hierfür nach Gleichung 2.3 ΔG-Werte berechnet und mittels einer multiplen linearen Regression die Parameter ermittelt, die sie möglichst gut erklären. ΔG_{mis} wurde bei der Regressionsanalyse ausgeschlossen, da unvorteilhafte Beiträge innerhalb von Kristallstrukturen kaum beobachtet werden können.

Die in die Bewertungsfunktion integrierten Lennard-Jones-ähnlichen (R, A)-Potentiale $LJ(r)$ hängen von der intermolekularen Distanz r eines Protein-Ligand-Atompaars ab. Sie beschreiben mittels ihres positiven Terms die Repulsion, die durch die Überlappung der Atomorbitale bei zu kurzer Distanz forciert wird. Ihr negativer Term beschreibt die Attraktion, die durch eine intermolekulare Interaktion zwischen den Atomen bei adäquater Distanz herbeigeführt wird:

$$LJ(r) = \epsilon \left[\left(\frac{r_m}{r} \right)^R - 2 \left(\frac{r_m}{r} \right)^A \right] \tag{6.4}$$

$r_m = r_{vdW}(i) + r_{vdW}(j) + t$ ist die Distanz, bei der ein Potential sein Minimum $-\epsilon$ erreicht. Sie entspricht der Summe der Van-der-Waals-Radien der Atome i und j (bzw. des ionischen und des Van-der-Waals-Radius bei Metall-Akzeptor-Kontakten) und ist durch die Toleranz t auf eine ideale Atompaardistanz für die jeweilige Interaktionsart justiert. Da die Atompaarpotentiale lediglich bis zu einer gewissen Cutoff-Distanz r_c betrachtet werden, sind alle Atompaarpotentiale $LJ(r)$ zudem korrigiert. Um einen Sprung des Potentials zu vermeiden, wird es leicht nach oben verschoben, sodass bei einer Distanz von r_c das Potential einen Wert von null annimmt:

$$\Delta G \cdot f(r) = \begin{cases} \frac{1}{-\epsilon}(LJ(r) - LJ(r_c)) & \text{für } r \leq r_c \\ 0 & \text{für } r > r_c \end{cases} \tag{6.5}$$

Zudem gilt $\Delta G = -\epsilon$, sodass ein ΔG-Parameter die Tiefe eines Potentialtopfes bestimmt, der bei optimaler Distanz der Interaktionsatome eingenommen wird.

In der cRAISE-Bewertungsfunktion sind für f_{hb}, f_{met}, f_{mis} weiche (4, 2)-Potentiale gewählt, um eine Überlappung der Van-der-Waals-Volumen zu gewähren, wie es in dieser Art von Interaktionen beobachtet werden kann. Das hydrophobe (12, 3)-Potential ist dagegen hart und bewertet die Passung der Pose genauer. Sein recht flacher Potentialtopf des attraktiven Teils ermöglicht es, kürzere Kontakte zwischen aliphatischen

Atomen und weitreichende aromatische Interaktionen zu registrieren. Abbildung 6.8 stellt die unterschiedlichen Atompaarpotentiale dar. Sie berücksichtigen die Attraktion bzw. die Repulsion zwischen Donoren und Akzeptoren, Metallen und Akzeptoren, hydrophoben Atompaaren bzw. unpassenden hydrophoben und hydrophilen Atomen. Tabelle 6.4 fasst alle Parameter der cRAISE-Bewertungsfunktion zusammen.

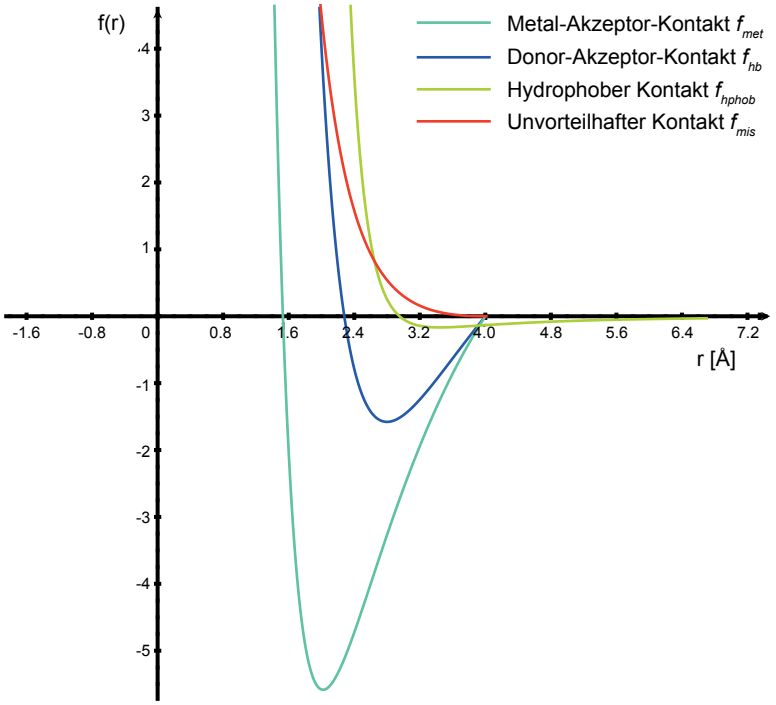


Abbildung 6.8: Atompaarpotentiale der cRAISE-Bewertungsfunktion.

Tabelle 6.4: Parameter der cRAISE-Bewertungsfunktion.

	ΔG	t	R	A	r_c
hb	-6.07	-0.20	4.0	2.0	4.0
met	-10.6	-0.20	4.0	2.0	4.0
lipo	-0.14	-0.21	12.0	3.0	6.5
mis	-0.50	0.6	4.0	2.0	3.6
rot	1.30				
0	-14.09				

Die Atompaarpotentiale bewerten die Kontaktdistanz. Bei gerichteten, hydrophilen Interaktionen wird zudem die Abweichung der etablierten von der idealen Interaktionsgeometrie bewertet. Hierbei können drei Fälle auftreten, die unterschiedliche Frei-

heitsgrade an Abweichungen definieren. Bezeichne D das Zentrum eines Donor- und A das Zentrum eines Akzeptorschweratoms, dann ist bei einer etablierten Interaktion die Hauptinteraktionsrichtung durch die direkte Verbindung DA gegeben. Seien zudem d und a assoziierte Interaktionsrichtungen von D und A, dann sind $\alpha = \measuredangle(d, DA)$ und $\beta = \measuredangle(a, AD)$ die Winkelabweichungen der Interaktionsrichtungen von der Hauptinteraktionsrichtung.

- Besitzen beide Interaktionsatome eine trigonal-planare VSEPR-Geometrie, so spannen sie jeweils eine Referenzebene auf, mit den Normalen n_D und n_A. Die Idealgeometrie tritt ein, wenn n_D und n_A mit $\gamma = \measuredangle(n_D, DA)$ und $\delta = \measuredangle(n_A, AD)$ senkrecht und die Interaktionsrichtungen gleich bzw. entgegengesetzt zu DA orientiert sind, d. h. wenn gilt $\sin\gamma = 1$, $\sin\delta = 1$, $\cos\alpha = 1$ und $\cos\beta = 1$.
- Sei D trigonal-planar und A tetrahedral oder umgekehrt, so reduzieren sich die gemessenen Freiheitsgrade, da eine Atomgeometrie keine Referenzebene besitzt. Die Interaktionsgeometrie ist ideal wenn gilt $\sin\gamma = 1$, $\cos\alpha = 1$ und $\cos\beta = 1$.
- Besitzen beide Interaktionsatome eine tetrahedrale Atomgeometrie oder sind linear, so können lediglich die Abweichungen der Interaktionsrichtungen von der Hauptinteraktionsrichtung gemessen werden. Die Idealgeometrie tritt ein wenn gilt $\cos\alpha = 1$ und $\cos\beta = 1$.

Das geometrische Mittel $G(\cos\alpha, \cos\beta, *)$ der Abweichungen dient als Faktor, um die Tiefe des Potentialtopfes weiter zu reduzieren und eine Suboptimalität der Interaktion bei ungünstiger Orientierung der Interaktionsatome zu verdeutlichen:

$$f(r, \alpha, \beta, *) = \begin{cases} G(\cos\alpha, \cos\beta, *) \cdot f(r) & \text{für } f(r) \leq 0 \\ f(r) & \text{für } f(r) > 0 \end{cases} \tag{6.6}$$

Bei repulsiven Bewertungen skaliert der Richtungsfaktor den Beitrag nicht. Für frei rotierbare Wasserstoffe bzw. Elektronenpaare (Schweratome mit tetrahedraler VSEPR-Geometrie und einem Schweratomnachbarn) kann optional, vor der eigentlichen Bewertung, die assoziierte Interaktionsrichtung entlang des durch die Atomgeometrie beschriebenen Kegels in Richtung des Gegenatoms optimal ausgerichtet werden.

Um für eine Pose eine Bewertung nach Gleichung 6.3 zu erhalten, wird jedes Posenschweratom mit jedem Schweratom seiner Proteinumgebung bewertet. Die Einzelbewertungen ergeben in Summe dann die Gesamtbewertung der Pose. Bei hydrophilen Interaktionsatomen wird für jede der assoziierten Donor- und/oder Akzeptorinteraktionsrichtungen die bestmögliche Interaktionsgeometrie mit der direkten Umgebung ermittelt und deren Bewertung mit dem Atompaarpotential verrechnet. Die Anzahl der

aufsummierten, gewichteten Paarpotentiale ist hierbei limitiert. Die Anzahl der Summanden wird durch das Minimum von a) der Anzahl der Wasserstoffatome und/oder freier Elektronenpaare des Posenatoms und b) der Anzahl der Wasserstoffatome und/oder freier Elektronenpaare der nahen Proteinumgebung bestimmt.

6.11 Die Bewertungshierarchie

Wird ein Docking-Werkzeug zum großangelegten VS genutzt, produziert dies eine enorme Anzahl von Posen. Die direkte Anwendung einer aufwendigen Bewertungsfunktion auf alle Posen würde massiv den Durchsatz eines VS-Prozess verringern. Die folgend vorgestellte Bewertungshierarchie verfolgt deshalb konsequent die Strategie einer möglichst frühen Ausmusterung wenig versprechender Posen. Sie identifiziert und verwirft Posen mit Proteinüberlappungen und spärlichen Proteinkontakten, sodass die Passung lediglich für bereits vielversprechende Posen bewertet werden muss. Die Bewertungshierarchie lässt sich in die *frühe* und die *späte Bewertungsphase* gliedern. Beide werden direkt nach der Initialisierung einzelner Posen sukzessiv durchgeführt. Zur Vermeidung von Mehrfachinitialisierungen, werden die Deskriptortreffer sortiert prozessiert. Dadurch wird die Molekül- und die Konformerinitialisierung für jedes getroffene Molekül und Konformation nur einmalig durchlaufen. Unterschiedliche Transformationen werden dann auf die einmalig initialisierten Konformationen angewendet. Algorithmus 6.4 stellt die Vorgehensweise vor und zeigt die Integration der frühen und späten Bewertungsphase innerhalb dieses Prozesses.

6.11.1 Frühe Bewertung

Die frühe Bewertungsphase gliedert sich in drei Bestandteile. Sie identifizieren und verwerfen gegebenenfalls Posen mit starken Überlappungen zum Protein oder mit spärlichen Proteinkontakten und selektieren eine Menge potentiell guter Posen, die in die späte Bewertungsphase eingehen. Jeder dieser Schritte greift hierbei auf die in Abschnitt 6.5 vorgestellte Bewertungsinformation zurück, die – einmalig während der Proteinpräparierung vorberechnet – zur Laufzeit stets verfügbar bleibt:

Enthaltensein im aktiven Volumen: Posen mit spärlichen Proteinkontakten sind mutmaßlich nicht dazu in der Lage, mit Proteinatomen zu interagieren und würden deshalb in späteren Phasen schlechter als andere Posen bewertet. Daher können sie bereits vorzeitig identifiziert und von der aufwendigeren Bewertungsphase ausgeschlossen werden. Sie zeichnen sich dadurch aus, dass sich die Mehrheit ihrer Atome nach Anwendung der Transformation größtenteils außerhalb des aktiven Volumens befindet.

Algorithmus 6.4 : cRAISE-Posenprozessierung

```
Eingabe : Treffer T, Rezeptor R
Ausgabe : H Molekülrangfolge
BewerteTreffer(T, R)
H ← ∅
M ← eindeutige Moleküle aus T
for i ∈ M do
    m ← initialisiere Molekül i
    T(m) ← Treffer des Moleküls nach Konformationen sortiert
    c ← initialisiere erste Konformation
    s ← ∅
    for j ∈ T(m) do
        if nächste Konformation then
            c ← initialisiere nächste Konformation
        t ← bestimme Transformation für Treffer j
        p ← initialisiere Pose durch Transformation t
        S_früh ← FrüheBewertung(p, R)
        if S_früh then
            s ← s ∪ {(S_früh, p)}
            s ← extrahiere die bislang 1 000 besten Bewertungen
    H ← H ∪ SpäteBewertung(s, R)
return H
```

Zu ihrer Identifikation wird für jede Schweratomkoordinate c einer Pose überprüft, ob sie außerhalb der vorberechneten aktiven konvexen Hülle liegt. Dies ist der Fall, wenn nach Definition 6.5.1 gilt, dass c für eine Facette f der triangulierten Hüllenoberfläche sichtbar ist. Erfüllt mindestens die Hälfte der Schweratome dieses Kriterium, so darf die Pose die Bewertungshierarchie nicht weiter passieren.

Identifizierung von Überlappungen: Da einzelne RAISE-Deskriptoren die Eigenschaften eines Moleküls nur lokal beschreiben, kann der Deskriptorabgleich die sterische Passung einer Pose in der aktiven Bindetasche nicht global für das gesamte Molekül garantieren. Einzelne Molekülbereiche können noch massive Überlappungen mit dem Protein an Stellen aufweisen, die nicht durch den Deskriptor abgedeckt sind. Da solche Posen keine validen Docking-Lösungen darstellen, können sie vorzeitig identifiziert und von der weiteren Posenprozessierung ausgeschlossen werden. Die gitterbasierte Proteinrepräsentation ermöglicht eine effiziente Überlappdetektion zwischen Posen und Protein. Während der Posenprozessierung wird für jedes Posenatom die atomtypspezifische Überlappinformation des nächst gelegenen Gitterpunkts abgefragt. Eine Überlappung des Posenatoms mit dem Protein wird festgestellt, wenn zuvor bereits die Sonde des

entsprechenden Typs eine Überlappung aufwies. Sobald ein Atom einer Pose eine Überlappung zeigt, wird die Pose vollständig verworfen und nicht in der Bewertungshierarchie weitergereicht. Dieser Test erlaubt immer noch leichte, intermolekulare Überlappungen. Sie werden von cRAISE an dieser Stelle toleriert und in der späten Bewertungsphase durch die Bewertungsfunktion registriert.

Gitterbasierte Bewertung: Die vorberechnete Bewertungsinformation wird während der Posenprozessierung direkt und ohne erneute Berechnung genutzt, um die Gesamtbewertung einer Pose zu kompilieren. Für jedes Posenatom wird auf die atomtypspezifischen Paarpotentiale am nächstgelegenen Gitterpunkt zugegriffen und die dort annotierten entgegengesetzten Interaktionsrichtungen werden bewertet. Ausschließlich die vorgemerkten Proteininteraktionsrichtungen werden mit den Interaktionsrichtungen des Posenatoms auf Abweichungen zur idealen Interaktionsgeometrie evaluiert. Frei rotierbare Interaktionsrichtungen werden zuvor optimal orientiert. Die Bewertungen der Interaktionsgeometrien skalieren die am Gitterpunkt annotierten Einzelbeiträge der Wasserstoffbrücken- und Metallinteraktionsatompaarpotentiale, die zusammen mit den hydrophoben Beiträgen und den Bestrafungsbeiträgen unpassender Atomkontakte über alle Posenatome aufsummiert werden. Zusammen ergeben sie eine gitterbasierte Bewertung der Pose entsprechend der cRAISE-Bewertungsfunktion nach Gleichung 6.3.

6.11.2 Späte Bewertung

Die späte Bewertungsphase fängt die Bewertungsdiskrepanzen ab, die in der frühen Bewertungsphase durch die gitterbasierte Abfrage der Potentialwerte entstehen können und stellt die finalen Hitliste des virtuellen Screening-Laufes zusammen.

Clustern von Posen: Nach der frühen Bewertungsphase liegt für jedes Molekül eine Selektion der 1 000 besten Posen vor, die gemäß der gitterbasierten cRAISE-Bewertungsfunktion sortiert sind. Da sie ähnlich bewertet werden, folgen in dieser Auswahl ähnlich orientierte Posen aufeinander. Damit die Docking-Lösungen möglichst diverse, repräsentative Bindungsmodi widerspiegeln, kann ein einzelner Repräsentant aus dieser Folge gewählt und sehr ähnliche Posen vorzeitig ausgeschlossen werden. Hierzu werden aufeinanderfolgende Posen mit einem RMSD von weniger als 0,5 Å eliminiert und nicht an den nächsten Bewertungsschritt weitergereicht.

Wasserstoffbrückenoptimierung: Die frühe Bewertung orientiert frei rotierbare Wasserstoff- und Elektronenpaarrichtungen optimal vor der Bewertung einer Interaktionsgeometrie. Sie betrachtet dabei allerdings nicht die weiter entfernten Abhängigkeiten im

Wasserstoffbrückennetzwerk und überschätzt dadurch gelegentlich den Beitrag einer gerichteten Interaktion. Vor der letzten Bewertungsphase erfolgt daher eine mit PROTOSS durchgeführte Wasserstoffbrückennetzwerkoptimierung, die nun auf einer eingeschränkten Menge von Protein-Posen-Komplexen erfolgen kann. Im grundlegenden Screening-Ablauf von cRAISE orientiert PROTOSS lediglich Wasserstoffkoordinaten von Posen, um eine objektivere Bewertung von Interaktionsgeometrien zu gewährleisten.

Atombasierte Bewertung: Der letzte Schritt der Posenprozessierung bewertet die maximal 1 000 Posen eines Moleküls erneut. Im Gegensatz zur gitterbasierten Bewertung nutzt die atombasierte Bewertung keine vorberechnete Bewertungsinformation, sondern bewertet die Passung direkt auf Atomkoordinaten gemäß der cRAISE-Bewertungsfunktion 6.3. cRAISE optimiert hierbei nicht die Interaktionsrichtungen, sondern verlässt sich zur Bewertung der Interaktionsgeometrien auf die von PROTOSS ermittelten Wasserstoffkoordinaten. Nach dieser Bewertung wird die beste Pose jedes getroffenen Moleküls bestimmt und zusammen mit der Molekülinformation und den transformierten Koordinaten in einer Lösungsdatenbank gespeichert. Die Bewertung wird letztendlich in die finalen Hitliste des virtuellen Screening-Laufes aufgenommen, die aus der Lösungsdatenbank jederzeit extrahiert werden kann. Die Hitliste wird gemäß der Bewertung sortiert und als Lösung des virtuellen Screenings präsentiert.

6.12 Parallelisierung

Bei der Anwendung von cRAISE ist die Molekülbibliothek in Form ein oder mehrerer Multi-SDF- und/oder Multi-Mol2-Dateien gegeben, die insgesamt N Moleküleinträge enthalten. Sowohl in der Präparierungs- als auch in der Screening-Phase muss jeder dieser Einträge bzw. die davon abgeleitete Information individuell verarbeitet werden. Die Berechnungen sind unabhängig voneinander. Deshalb ist es möglich, die notwendigen Rechenaufgaben gleichmäßig auf einen Rechnerverbund zu verteilen. Ist ein Rechnerverbund von n Knoten mit jeweils m Recheneinheiten gegeben, dann teilen sich die Recheneinheiten eines Knotens den peripheren Sekundär- und den Hauptspeicher. In solch einer Rechenumgebung können einzelne Rechenaufgaben der Präparierungs- und Screening-Phase von einzelnen Recheneinheiten wie folgt übernommen werden:

Präparierungsphase: cRAISE ermöglicht es, einen Bereich von Moleküleinträgen anzugeben und nur diesen Teil der Bibliothek zu verarbeiten (vgl. Anhang C). Dadurch ist es möglich, stückweise über die gesamte Eingabe zu iterieren. Im Rechnerverbund kann die Eingabebibliothek so in $n \cdot m$ disjunkte Partitionen unterteilt und unter Nutzung der Sun Grid Engine (SGE 6.2u5) jeweils einer Recheneinheit zugewiesen werden.

Jede Recheneinheit führt die komplette Präparierungsphase auf dem ausgewählten Eingabebereich durch. Eine verteilte Präparierungsphase erzeugt dadurch insgesamt $n \cdot m$ Partitionen, die jeweils eine Moleküldatenbank mit assoziierten Index aus Deskriptoren von $N/(n \cdot m)$ registrierten Molekülen umfasst. Die erzeugten Partitionen können auf einem für jede Recheneinheit zugänglichen, zentralen Speichermedium verwaltet oder auf die restlichen Knoten verteilt werden, um die Netzwerklast in späteren Screening-Phasen zu minimieren.

Screening-Phase: Zur Durchführung der Screening-Phase ist eine Moleküldatenbank mit assoziiertem Index notwendig. Nach einer verteilten Präparierungsphase liegen sie in den $n \cdot m$ Partitionen vor und können wiederum an $n \cdot m$ Recheneinheiten verteilt werden. Eine Recheneinheit führt dann eine vollständige Screening-Phase mit der ihr zugewiesenen Partition aus. Das Resultat ist eine Lösungsdatenbank, die die Koordinaten der besten Pose von maximal $N/(n \cdot m)$ Molekülen zusammen mit ihren Bewertungseinträgen enthält. Um ein globales Screening-Resultat zu erhalten, müssen die $n \cdot m$ Resultate vereint werden. cRAISE bietet hierfür die Möglichkeit die Einträge mehrerer angesammelter Lösungsdatenbanken zu einer globalen Lösungsdatenbank zu vereinen. Von einer Lösungsdatenbank kann dann eine Hitliste abgeleitet werden, die alle Lösungseinträge sortiert darstellt.

6.13 Molekülprofile

Die bislang beschriebenen Komponenten finden sich im Abschnitt 5.1 vorgestellten, grundlegenden Ablauf eines virtuellen Screening mit cRAISE wieder. Zur Modellierung von Randbedingungen, die durch ein Molekülprofil gegeben sind, wurden einzelne Komponenten der Screening-Phase modifiziert.

6.13.1 Spezifikation von Molekülprofilen

Die Filterung von Molekülen während der Screening-Phase anhand eines gegebenen Molekülprofils ist auf Basis der Funktionalitäten von Mona umgesetzt (vgl. Abschnitt 4.5). Während der Registrierung der Moleküle in der Moleküldatenbank (vgl. Abschnitt 6.7) werden *per se* konstitutionelle und topologische Moleküleigenschaften berechnet und in der Datenbank gespeichert. Die registrierten Moleküleigenschaften werden im grundlegenden Screening-Ablauf nicht weiter genutzt. Sie finden jedoch bei einem durch den Anwender gegebenen Molekülprofil ihre Anwendung. Tabelle 4.6 listet alle von Mona und somit auch von cRAISE unterstützten molekularen Eigenschaften auf, die bei der Molekülregistrierung berücksichtigt werden. Sie können klassifiziert werden nach

Eigenschaften, für die die Moleküldatenbank optimierte Bereichs- oder Existenzanfragen unterstützt. Ein Molekülprofil kann durch eine logische Kombination verschiedener Bereichs- und Existenzbedingungen auf den molekularen Eigenschaften definiert werden. Im Anhang A.3 ist exemplarisch ein Molekülprofil definiert.

6.13.2 Molekülprofilgeleitetes virtuelles Screening

Ist ein Molekülprofil durch den Anwender gegeben, so wird vor dem eigentlichen Deskriptorabgleich aus der Profildefinition zunächst eine Filterkette etabliert. Die Filterkette wird durch eine an die Moleküldatenbank gestellte SQL-Anfrage realisiert. Sie überprüft die gegebenen Existenz und/oder Bereichsbedingungen auf den registrierten Moleküleigenschaften und extrahiert die IDs der Moleküle m_{id}, die die gegebenen Bedingungen erfüllen. Die durch die Anfrage an die Moleküldatenbank extrahierten IDs werden in der Filtermenge F zusammengefasst. Die Zusammenhänge, der genutzten Daten während einer Molekülfilterung sind in Abbildung 6.9 dargestellt.

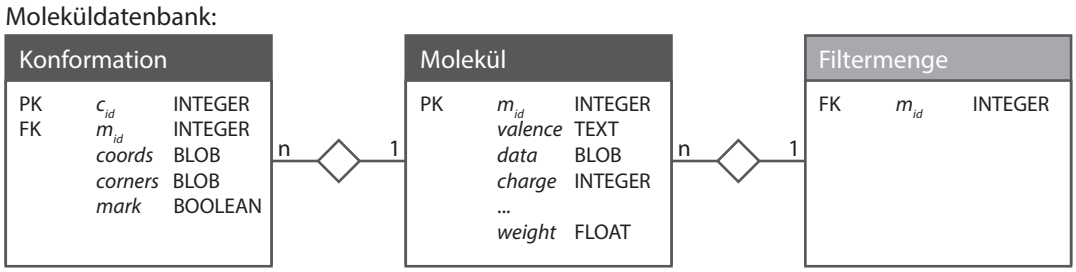


Abbildung 6.9: Erweiterung der cRAISE-Moleküldatenbank durch eine Filtermenge.

Die Filtermenge F nutzt cRAISE, um die gewöhnlichen Anfragen an den Deskriptorindex nach Gleichung 6.2 durch weitere Bedingungen zu ergänzen. Seien ein Anfragedeskriptor q und eine Filtermenge F gegeben, dann ist ein im Index I hinterlegter Moleküldeskriptor d genau dann ein Treffer, wenn folgende Bedingung erfüllt ist:

$$\begin{aligned} \mathit{Treffer}(q, F) = \{d \in I \mid d \in \mathit{Treffer}(q, \Delta_l, \Delta_b, \Delta_r) \\ \wedge\, m_{id}(d) \in F\} \end{aligned} \tag{6.7}$$

Durch diese Ergänzung wird eine eingeschränkte Auswahl an Treffern erhalten. Sie besteht ausschließlich aus Treffern mit Indexdeskriptoren, die in der Präparierungsphase aus Molekülen abgeleitet wurden, die das gegebene Molekülprofil erfüllen. Die Treffer eines gewöhnlichen Deskriptorabgleichs, die dem Molekülprofil widersprechen, werden nicht aus dem Index extrahiert. Für sie werden somit auch keine Posen produziert, die während der Posenprozessierung in der Bewertungshierarchie weitergereicht und

als Lösung des virtuellen Screenings präsentiert werden müssen. Die Anwendung von Molekülprofilen ermöglicht dadurch einen Ausschluss von Molekülmengen aus dem virtuellen Screening-Prozess, ohne dafür erneut eine aufwendige Präparierungsphase auf einer zuvor eingeschränkten Molekülbibliothek durchführen zu müssen.

6.14 cRAISE-Pharmakophorhypothesen

Die Randbedingungen, die durch eine Pharmakophorhypothese gegeben sind, modifizieren ebenso die Screening-Phase. Zur Anwendung muss die Hypothese gemäß des cRAISE-Pharmakophorformats definiert sein.

6.14.1 Spezifikation eines Pharmakophormodells

cRAISE unterstützt die Spezifikation pharmakophorartiger Inklusions- und Exklusionsmerkmale, die Bedingungen bezüglich des Vorhandenseins, der Lage, des Typs und der Orientierung bei der Platzierung von Ligandatomen stellen. Ein Inklusionsmerkmal ist eine Bedingung, die eine Region in der aktiven Bindetasche definiert, an der ein Ligandatom liegen muss. Exklusionsmerkmale definieren für die Platzierung verbotene Bereiche in der aktiven Bindetasche. Generell ist ein Pharmakophormerkmal eine Sphäre, die durch Zentrum und Radius definiert ist. Sowohl Inklusions- als auch Exklusionmerkmale besitzen einen Typ $T \in \{Donor, Akzeptor, Hydrophob, Hydrophil, Any\}$. Inklusionsmerkmale vom Typ *Donor*, *Akzeptor* oder *Hydrophil* sind mit einer Richtung ausgestattet, um die Orientierung von Protonen bzw. freier Elektronenpaare eines Ligandatoms zu beschränken. Die anderen Inklusionsmerkmale und im Speziellen Exklusionsmerkmale sind ungerichtet. Tabelle 6.5 fasst alle Arten möglicher Pharmakophormerkmale zusammen und beschreibt, welche Bedingungen sie während der Posengenerierung fordern. Eine Pharmakophorhypothese kann durch eine willkürliche Zusammenstellung von Inklusions- und Exklusionsmerkmalen spezifiziert werden. Zudem kann eine *Zahl essentieller Inklusionsmerkmale* N_e angegeben werden. Sie gibt an, wie viele Inklusionsmerkmale zugleich durch ein platziertes Molekül erfüllt werden müssen. Wird N_e durch den Anwender mit eins angegeben oder nicht weiter spezifiziert, so werden die enthaltenen Merkmale uneingeschränkt mit einem logischen ODER verknüpft interpretiert. D. h. eine Pose muss lediglich ein Merkmal erfüllen, um der gesamten Pharmakophordefinition zu genügen. Eine UND-Verknüpfung aller Inklusionsmerkmale kann erreicht werden, indem N_e auf die Anzahl der in der Pharmakophordefinition enthaltenen Merkmale gesetzt wird. Für ein N_e größer eins aber kleiner als die Anzahl, der in der Definition enthaltenen Merkmale, wird die Disjunktion jeder möglichen N_e-Konjunktion

Tabelle 6.5: Interpretation von Pharmakophormerkmalen bei der Posengenerierung.

Merkmalstyp	Interpretation
Donorinklusion[a]	Platziere ein H-Brückendonor- oder Kationzentrum mit entsprechender Protonausrichtung
Akzeptorinklusion[a]	Platziere ein H-Brückenakzeptor- oder Anionzentrum mit entsprechender Elektronenpaarausrichtung
Hydrophobe Inklusion[b]	Platziere eine hydrophobe Gruppe
Hydrophile Inklusion[a]	Platziere ein H-Brückendonor-, H-Brückenakzeptor-, Kation- oder Anionzentrum mit entsprechender Proton- bzw. Elektronenpaarausrichtung
Any-Inklusion[b]	Platziere irgendein Atomzentrum
Donorexklusion[b]	Platziere kein H-Brückendonor- und kein Kationzentrum
Akzeptorexklusion[b]	Platziere kein H-Brückenakzeptor- und kein Anionzentrum
Hydrophobe Exklusion[b]	Platziere kein hydrophobes Atomzentrum
Hydrophile Exklusion[b]	Platziere kein H-Brückendonor-, H-Brückenakzeptor-, Kation- und kein Anionzentrum
Any-Exklusion[b]	Platziere kein Atomzentrum

[a] Gerichtetes Pharmakophormerkmal
[b] Ungerichtetes Pharmakophormerkmal

der gegebenen Merkmale ausgewertet. Im Anhang A.2 ist exemplarisch eine cRAISE-Pharmakophordefinition bereitgestellt, die alle unterstützten Merkmalstypen umfasst.

6.14.2 Pharmakophorgeleitetes virtuelles Screening

Ist eine Pharmakophorhypothese gemäß des cRAISE-Formats der Screening-Phase zur Verfügung gestellt, so werden gegebene Inklusionsmerkmale an zwei Stellen verarbeitet: Um pharmakophorerfüllende Moleküle aus der Bibliothek zu extrahieren, werden

1. vor dem Deskriptorabgleich Anfragedeskriptoren ausgefiltert und
2. nach dem Deskriptorabgleich Posen verworfen

wenn sie der gestellten Hypothese widersprechen. Abbildung 6.10 stellt die Verarbeitung einer Pharmakophorhypothese während der Anfragendeskriptorberechnung und der Posenfilterung dar. Bei der Anfragengenerierung wird die Einhaltung der Hypothese lokal überprüft. Nach Invertierung der Interaktionsstellen werden ausschließlich Deskriptorbasisdreiecke erstellt, die zumindest eine Ecke in einer *Donor*-, *Akzeptor*-, *Hydrophilen* oder *Any*-Inklusionssphäre enthalten, den entsprechenden Typ besitzen und gleichermaßen orientiert sind. Da RAISE-Deskriptoren Moleküle nur lokal abdecken, könnten

Falsch-Negativ-Vorhersagen auftreten, wenn zu diesem Zeitpunkt die Erfüllung mehrerer Inklusionen forciert werden würde. Hydrophobe Merkmale beschränken grundsätzlich nie Anfragedeskriptoren. Deshalb beeinflussen Hypothesen mit ausschließlich hydrophoben Merkmalen die Anfragengenerierung nicht. Sie werden nach dem Deskriptorabgleich ausgewertet. Hierfür wird während der Posenprozessierung die Hypothese

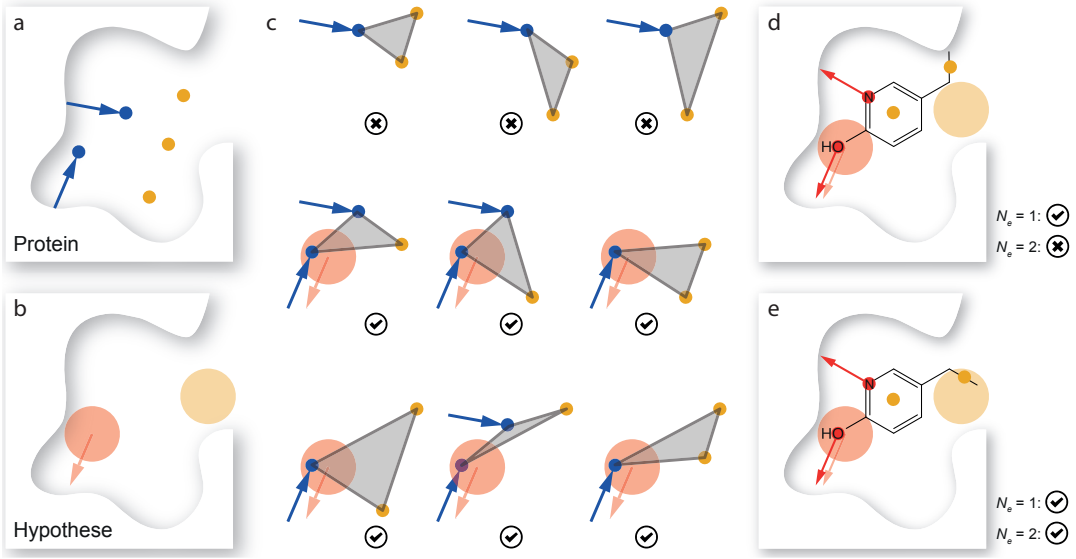


Abbildung 6.10: Verarbeitung von Pharmakophormerkmalen: a) Schematische Darstellung von Interaktionsstellen eines Proteins. b) Hypothese, die ein Akzeptoratom bzw. eine hydrophobe Gruppe an einer bestimmten Stellen der aktiven Bindetasche fordert. c) Anfragedeskriptoren, die von den Interaktionsstellen abgeleitet werden können. Drei von ihnen werden durch die Hypothese verworfen. d) Die Pose erfüllt ein einzelnes Merkmal und die gesamte Hypothese wenn $N_e = 1$ spezifiziert wurde. e) Die Pose erfüllt zwei Merkmale und die gesamte Hypothese wenn $N_e = 1$ oder $N_e = 2$ spezifiziert wurde.

erneut überprüft. Für jede erzeugte Pose werden alle Inklusionsmerkmale getestet. Findet sich ein Posenatom entsprechenden Typs in der Sphäre des Merkmals wieder und ist gegebenenfalls ähnlich orientiert, so wird das Merkmal erfüllt. Erfüllt die Pose zumindest N_e Merkmale, dann erfüllt sie die Hypothese auch global. Der Pharmakophortest wird in der frühen Bewertungsphase (vgl. Abschnitt 6.11) im Anschluss zum Überlapp- und Aktiven-Volumen-Test durchgeführt. Bei Erfolg wird die Pose in der Bewertungshierarchie weitergereicht.

Exklusionmerkmale werden in der Rezeptorpräparierung (vgl. Abschnitt 6.5) verarbeitet. Bei der Vorberechnung des Überlappgitters werden Exklusionssphären hierbei

quasi als Proteinatomsphären betrachtet. Gitterpunkte, die in die Sphären hineinragen werden als belegte Bereiche markiert. Im Gegensatz zu ordinären Proteinatomen werden Überlappungen jedoch nur für bestimmte Atomtypen vermerkt. So schließt ein Exklusionsmerkmal vom Typ *Hydrophob* nur die Platzierung hydrophober Ligandatome aus. Akzeptor- und Donorexklusionen schließen analog dazu die Platzierung eines bestimmten Atomtyps aus. Hydrophile Exklusionen verbieten die Platzierung zweier Atomklassen. Sowohl Donor- als auch Akzeptoratome dürfen in diesen Bereichen nicht zu liegen kommen. Exklusionen vom Typ *Any* verbieten jegliche Platzierung in den dadurch definierten Bereich. Nach der Annotation der Exklusionmerkmale am Überlappgitter wird die Erfüllung von Exkluxionsmerkmalen dann mittels eines typisierten Überlapptest in der frühen Bewertungsphase der Bewertungshierarchie durchgeführt.

6.15 Modellierung makro-/molekularer Zustände

Zur Modellierung molekularer und makromolekularer Zustände wurde der grundlegende Ablauf der Präparierungs- und Screening-Phase modifiziert. Im Kern wurde das in Abschnitt 6.1 vorgestellte Interaktionsmodell erweitert, sodass Interaktionen vor einem expliziten Zustandswechsel von Protein und/oder Molekül bei der Ligandbindung registriert werden können. Auf Basis des erweiterten Interaktionsmodells wurde die Berechnung von Protein- und Moleküldeskriptoren (vgl. Abschnitt 6.6) angepasst. Der in Abschnitt 6.9 beschriebene Deskriptorabgleich selbst blieb von den Anpassungen unberührt. Auf *Multizustandsdeskriptoren* angewendet, wählt er jedoch passende, lokale Zustände der Bindungspartner. Die Zustandsselektion auf Deskriptorebene geschieht indirekt. D. h. cRAISE erkennt welche Atome prinzipiell ihren Zustand ändern könnten, um eine Interaktion zu etablieren, die Atome müssen jedoch nicht bereits diesen Zustand angenommen haben. Die explizite Zustandsänderung wird während der Posenprozessierung durchgeführt, nachdem die Lage des Liganden in der aktiven Bindetasche bekannt ist. Modifizierungen wurden auch bei der Molekülregistrierung (vgl. Abschnitt 6.7) vorgenommen, um normalisierte Konformere für Moleküle mit unterschiedlichen Zuständen zu garantieren.

6.15.1 Erweiterung des Interaktionsmodells

Protomere, d. h. Moleküle, die ihren Protonierungszustand oder ihre tautomere Form ändern können, erhalten oder verlieren Wasserstoffatome bzw. sie werden innerhalb des Moleküls an eine andere Stelle transferiert. Dadurch wechseln hydrophile Atome ihre Eigenschaft. Potentielle Wasserstoffbrückendonoren werden zu Akzeptoren und potentielle

Wasserstoffbrückenakzeptoren werden zu Donoren. Solche *Multizustandsatome* sind im VSC-Modell (vgl. Abschnitt 4.2.1) charakterisiert. Es handelt sich hierbei um Atome, die in Multizustandszonen liegen und nach Definition 4.2.3 alternative Valenzzustände besitzen. Nach dem Naomi-Modell ändert ein Zustandswechsel nie die Geometrie eines Multizustandsatoms jedoch dessen Klassifizierung als potentieller Wasserstoffbrückendonor bzw. -akzeptor. Um dieses Verhalten in cRAISE zu modellieren, erhalten Multizustandsatome in Multizustandszonen sowohl Donor- als auch Akzeptorinteraktionsstellen mit assoziierten Interaktionsrichtung(en) derselben Atomgeometrie (*Multizustandsinteraktionsstellen*). Dadurch wird es möglich, Wasserstoffbrücken indirekt zu identifizieren, auch wenn Interaktionsatome noch nicht explizit den komplementären Zustand zur Etablierung der Interaktion angenommen haben (vgl. Abbildung 6.11).

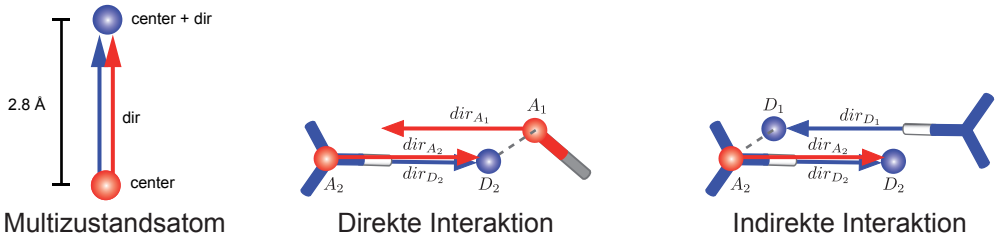


Abbildung 6.11: Multizustandsinteraktionsmodell: Multizustandsatome erhalten Donor- und Akzeptorinteraktionsstellen mit identischen Interaktionsrichtungen. Dadurch können gerichtete Interaktionen direkt (komplementäre Atomtypen) und indirekt (inkompatible Donor-Donor- oder Akzeptor-Akzeptor-Paaren) erkannt werden.

Zur Berechnung von Multizustandsinteraktionsstellen werden Multizustandszonen im Molekül und Protein wie in Abschnitt 4.2.2 deduziert und Valenzzustandsfolgen für diese Komponenten aufgezählt. Um lediglich wahrscheinliche Zustandsänderungen zu berücksichtigen, werden Valenzzustandsfolgen mit einem $S_{VSC} < 93$ (vgl. Gleichung 4.1) verworfen. Positionen innerhalb der verbleibenden Valenzzustandsfolgen, die unterschiedliche Valenzzustände aufweisen, werden als Multizustandsatome identifiziert. Generell sind dies Heteroatome funktioneller Gruppen, die leicht de-/protoniert werden können und Heteroatome in zusammenhängenden Komponenten konjugierter Systeme, die nicht eindeutig durch eine einzelne lokalisierte Struktur beschrieben werden können. Sie werden mit Multizustandsinteraktionsstellen versehen. Proteinseitig geschieht dies für die Heteroatome von Histidin und gegebenenfalls für die von Kofaktoren. Atome in Multizustandszonen, die nur einen Valenzzustand besitzen und hydrophile Atome außerhalb von Multizustandszonen sind zustandsinvariant. Sie erhalten die herkömmlichen Interaktionsstellen entsprechend ihres Atomtyps. Abbildung 6.12 zeigt exemplarisch alle Interaktionsstellen von Pyrimethamin.

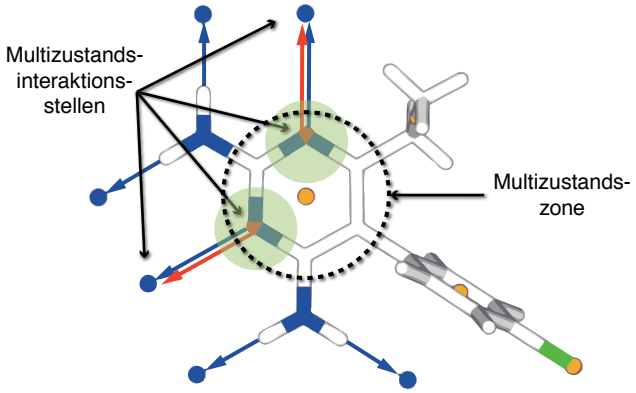


Abbildung 6.12: Multizustandsinteraktionsstellen von Pyrimethamin: Multzustandsatome in Multizustandszonen (grün hinterlegt) werden mit Multizustandinteraktionsstellen ausgestattet. Sie erhalten ein Paar von Donor- (blau) und Akzeptorinteraktionsstellen (rot). Hydrophile Atome zustandsinvarianter Zonen erhalten statische Interaktionsstellen, d. h. entweder Donor- oder Akzeptorinteraktionsstellen.

6.15.2 Berechnung von Multizustandsdeskriptoren

Der cRAISE-Multizustandsansatz nutzt den RAISE-Deskriptor in seiner herkömmlichen Form, leitet die Deskriptoren jedoch von Multizustandsinteraktionsstellen ab. Für die Deskriptorberechnung bildet jedes Triplett von Interaktionsstellen die Ecken eines Basisdreiecks. Multizustandsinteraktionsstellen können jedoch nicht willkürlich zu Basisdreiecken kombiniert werden, da die Kombination nicht notwendigerweise einen sinnvollen, globalen Zustand eines Moleküls widerspiegeln muss. Besitzt ein Molekül mehrere Multizustandszonen, so können nach dem VSC-Modell die Zustände der Zonen unabhängig zu plausiblen, globalen Zuständen eines Moleküls rekombiniert werden. Multizustandsinteraktionsstellen aus unterschiedlichen Zonen können daher ebenso unabhängig zu validen Basisdreiecken kombiniert werden. Nicht sinnvolle Deskriptoren werden nur aus einer Kombination von Interaktionsstellen derselben Multizustandszone erhalten. Da die hier annotierten Valenzzustandsfolgen jeweils einen plausiblen Zustand der Zone reflektieren, können zudem nicht plausible Deskriptoren nur aus Interaktionsstellen von Atomen mit Valenzzuständen unterschiedlicher Folgen hervorgehen. Zur Enumerierung sinnvoller Basisdreiecke werden die Tripletts daher auf ihre Plausibilität überprüft. Für einen validen Multizustandsdeskriptor müssen folgende Bedingungen erfüllt sein:

- Ein Triplett resultiert nicht aus Multizustandsinteraktionsstellen desselben Atoms.

- Ein Triplett resultiert nicht aus einer Kombination von Interaktionsstellen, deren Valenzzustände aus unterschiedlichen Valenzzustandsfolgen einer Zone stammen.

Abbildung 6.13 zeigt exemplarisch Multizustandsdeskriptoren von Pyrimethamin.

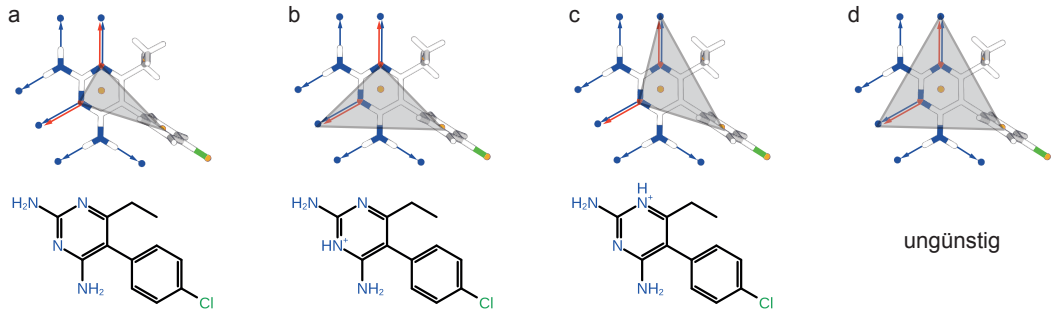


Abbildung 6.13: Multizustandsdeskriptoren von Pyrimethamin: a) Ein Deskriptor, der den deprotonierten Zustand der Pyrimidingruppe widerspiegelt. b) Ein Deskriptor, der dem N1-Tautomer entspricht c) Ein Deskriptor, der dem N3-Tautomer entspricht. d) Ein Deskriptor, der einen nicht sinnvollen protonierten Zustand von Pyrimidin widerspiegelt. Er wird während der Deskriptorberechnung ausgeschlossen.

6.15.3 Molekularer und makromolekularer Grundzustand

Im cRAISE-Multizustandsansatz werden alle Moleküle der gegebenen Bibliothek vor ihrer Registrierung in der Moleküldatenbank und der Deskriptorberechnung jeweils in einen einheitlichen Grundzustand überführt. Die Methode zur Erzeugung des Grundzustands folgt dem VSC-Modell.[313] Sie generiert eine kanonische Repräsentation des Moleküls, die invariant bezüglich des Protonierungszustand und der tautomeren Form des Eingabemoleküls ist. Dazu begünstigt sie deprotonierte Säuregruppen, protonierte Basen und erhält andere funktionelle Gruppen neutral, um den mutmaßlich stabilsten Molekülzustand zu etablieren. Delokalisierte Systeme werden durch eine einzelne Resonanzform dargestellt, um identische Valenzstrukturrepräsentationen für unterschiedliche Zustände des Moleküls zu erhalten. Der zugewiesene Grundzustand ermöglicht es, mit diversen Zuständen überrepräsentierte Moleküle in der Bibliothek als Duplikate zu identifizieren und sie während der Registrierung als verschiedene Eingabeinstanzen eines Moleküls zu markieren (vgl. Abschnitt 6.7). So wird die Einmaligkeit der Moleküle in der Moleküldatenbank garantiert und eine wiederholte Verarbeitung identischer Moleküle während des Screenings vermieden. Zudem sichert der Grundzustand bezüglich des gegebenen Eingabezustands unvoreingenommene Docking-Berechnungen zu. Zwar ist die

Erzeugung von Multizustandsinteraktionsstellen und -deskriptoren unabhängig vom gegebenen Molekülzustand, speziell die Konformergenerierung und die dafür notwendige Molekülkanonisierung sind es jedoch nicht. Die Bestimmung der zentralen Molekülkomponente und die Identifizierung von Torsionssignaturen flexibler Bindungen (vgl. Abschnitt 4.4) sind bei der Konformergenerierung vom Molekülzustand abhängig. Mit der Überführung der Moleküle in den Grundzustand gehen diese Abhängigkeiten verloren und die Konformere eines Protomers werden durch die Konformere des Grundzustands normalisiert.

Vor der Berechnung von Anfragedeskriptoren wird die aktive Bindetasche ebenso in einen Grundzustand überführt. Um diesen Rezeptorgrundzustand zu erzeugen, wird der stabilste Zustand jeder Aminosäure zugewiesen. Säure- und Basenreste erhalten ihren geladenen Zustand. Im Fall von Histidin ist der stabilste Zustand allerdings nicht eindeutig. Die neutralen $N^{\delta 1}$- und $N^{\epsilon 2}$-Tautomere und der protonierte Zustand des Imidazols sind in Lösung nahezu gleich wahrscheinlich. Bleibt die direkte Umgebung der Aminosäure im Protein unberücksichtigt, so kann nicht eindeutig entschieden werden welcher Zustand tatsächlich der stabilste ist. Daher werden die Grundzustände von Histidin mit Hilfe von PROTOSS bestimmt. Es betrachtet die Aminosäurezustände im Wasserstoffbrückennetzwerk und justiert sie bezüglich der Bildung möglichst optimaler, proteininterner Wasserstoffbrücken. Der Referenzligand bleibt von der Optimierung ausgeschlossen, um einen komplementären Rezeptorzustand, der einen Abdruck des gegebenen Liganden bezüglich seiner ausgerichteten Wasserstoffe und freier Elektronenpaare bildet, zu vermeiden.

6.15.4 Deskriptorabgleich mit Multizustandsdeskriptoren

CRAISE realisiert ein Protein-Ligand-Docking über den in Abschnitt 6.9 beschriebenen Abgleich von Protein- und Moleküldeskriptoren. Dieser Deskriptorabgleich kann ohne jegliche Anpassung auf den Abgleich von Multizustandsdeskriptoren angewendet werden. Der Abgleich wählt dadurch indirekt den passenden Zustand der Bindungspartner. Abbildung 6.14 veranschaulicht diese Zustandsselektion auf Deskriptorebene. Da jeder Proteindeskriptor einen bestimmten Rezeptorzustand repräsentiert und lediglich Liganddeskriptoren komplementären Typs extrahiert, deutet ein Paar passender Deskriptoren eine mögliche Rezeptor-Ligand-Zustandskonstellation an. Bezüglich des Multizustandsansatzes hat das indexbasierte Screening zudem einen entscheidenden Vorteil: Die Indexierung separiert Deskriptoren unterschiedlicher Tautomere und Protonierungszustände voneinander, da sich die Moleküldeskriptoren unterschiedlicher Molekülzustände entscheidend im Deskriptortyp und maßgeblich auch in den Seitenlängen unterscheiden.

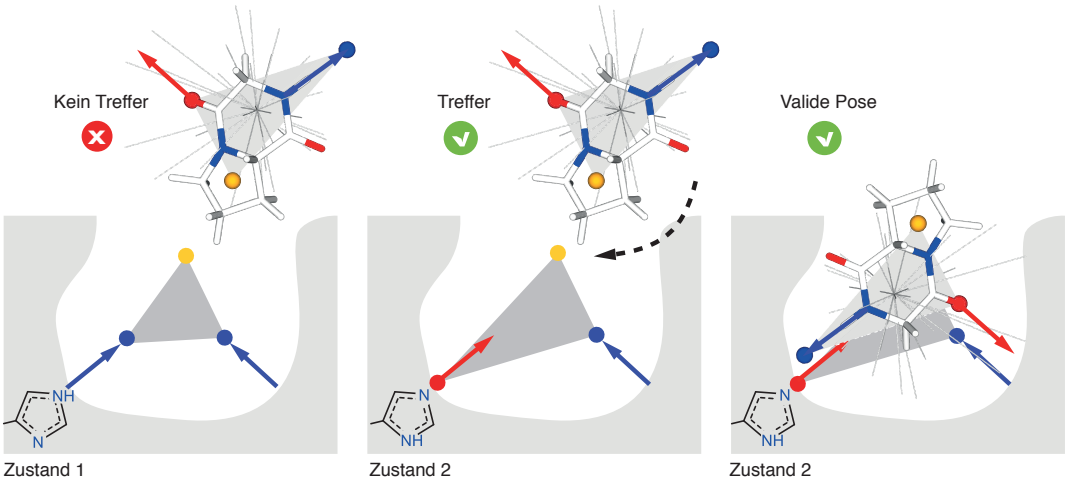


Abbildung 6.14: Implizite Zustandsselektion auf Deskriptorebene: Rezeptorzustand 1 ist unpassend, da der Deskriptortyp nicht vom komplementären Typ des Liganddeskriptors ist. Rezeptorzustand 2 ist komplementär. Er erzeugt eine valide Pose.

Dadurch ruft das indexbasierte Nachschlagen ausschließlich günstige Moleküldeskriptorzustände ab und schließt ungünstige Deskriptorzustände grundsätzlich aus. Dadurch entfällt die Auswertung jeder möglichen Rezeptor-Ligand-Zustandskonstellation.

6.15.5 Bewertung von Posen

Da die Zustandsselektion implizit auf Deskriptorebene geschieht, können Posen und Rezeptor nach dem Deskriptorabgleich tatsächlich noch unpassende Atomtypen aufweisen. Um aber eine objektive Bewertung der Posen zu ermöglichen, müssen Wasserstoffkoordinaten explizit berechnet werden. Diese Aufgabe übernimmt Protoss. Es passt die Zustände des Proteins und des Moleküls an und richtet frei rotierbare Wasserstoffkoordinaten aus, um so ein möglichst optimales Wasserstoffbrückennetzwerk für jeden generierten Komplex zu erzeugen. Um den Durchsatz während der Bewertungsphase zu gewährleisten, werden Protoss lediglich vielversprechende Posen angeboten. Hierfür wurde im cRAISE-Multizustandsansatz die Bewertungsfunktion der frühen Bewertungsphase angepasst. Da zu diesem Zeitpunkt Interaktionsatome desselben Typs noch aufeinander zeigen können, wird bei der Ermittlung der Wasserstoffbrücken- und Metallinteraktionsbeiträge die Überprüfung auf Komplementarität für Multizustandsatome vernachlässigt. Hierfür werden zur Vorberechnung der Bewertungsinformation, Wasserstoffbrückenpotentiale von Rezeptormultizustandsatomen als solche gekennzeichnet. Sie tragen bei Abfrage zur Bewertung von Donor- und Akzeptorposenatomen bei. Zudem

werden Multizustandsatome der Moleküle auch in der Posenprozessierung ermittelt. Solche Posenatome erhalten die Bewertung mit rezeptorseitigen Akzeptoren, Donoren und mit Metallen. Die Geometrie, einer Multizustandsinteraktion ist unabhängig vom tatsächlich eingenommen Atomtyp. Sie ändert sich auch bei einem Zustandswechsel nicht. Der Richtungsfaktor kann deshalb mit den idealen, typunabhängigen Potentialwerten verrechnet werden. Mit dieser Bewertungsstrategie werden Posen mit Multizustandsatomen oder Posen nahe zu Rezeptormultizustandsatomen willentlich überschätzt. Sie werden dadurch in der Bewertungshierarchie weitergereicht und der späten Bewertung präsentiert. Dort werden die Wasserstoffe dann explizit adaptiert und die Posen erneut mit der cRAISE-Bewertungsfunktion atomtypabhängig, nach Gleichung 6.3 bewertet. Die beste Pose jedes Moleküls geht in die finale Hitliste des virtuellen Screenings ein.

7 Bewertungsmaße, Daten und Experimente

Dieses Kapitel skizziert die Experimente, die zur Leistungsbewertung von cRAISE durchgeführt wurden. Die Experimente bewerten die Methode bezüglich ihrer Leistung, den aktiven Bindemodus vorherzusagen, aktive Moleküle im virtuellen Screening anzureichern und die aufzuwendenden Ressourcen. Zudem dienen sie dazu, den Effekt geleiteter Vorhersagen und den Einfluss von Zuständen bei der Anwendung von cRAISE zu untersuchen. Neben der Beschreibung der Experimente sind die genutzten Maße, kritischen Kennzahlen und die verwendeten Daten in diesem Kapitel zusammenfasst.

7.1 Vorhersage des aktiven Bindungsmodus

7.1.1 Bewertungsstrategie

Um zu bewerten, ob der korrekte Bindungsmodus eines Liganden vorhergesagt werden kann, wird vorausgesetzt, dass eine experimentell bestimmte Kristallstruktur des Protein-Ligand-Komplexes die dreidimensionale Struktur des Liganden zeigt. Wird der Ligand aus diesem Komplex extrahiert und erneut in die Proteinstruktur gedockt, dann soll eine Docking-Methode die Lage des nativen Liganden (*Kristallligand*) reproduzieren. Diese Art von Experiment wird als *Redocking* oder *natives Docking* bezeichnet.

7.1.2 Maße

Root Mean Square Deviation (RMSD): Der RMSD bewertet wie gut der vorhergesagte dem realen Bindungsmodus entspricht. Er vergleicht die kartesischen Koordinaten p der Atome einer generierten Pose mit den entsprechenden Koordinaten c in der Kristallstruktur und bestimmt die Quadratwurzel aus deren mittlerer quadratischen Abweichung:

$$\mathrm{RMSD}(c,p) = \sqrt{\frac{\sum_{i=1}^{n}(c_i - p_i)^2}{n}} \tag{7.1}$$

Standardmäßig werden nur Abweichungen der n Schweratomkoordinaten berücksichtigt, da Kristallstrukturen Wasserstoffkoordinaten zumeist unaufgelöst lassen. Zudem ist für Moleküle mit symmetrischen Gruppen keine eindeutige Zuordnung korrespondierender Schweratome möglich. Deshalb wird dort der RMSD bezüglich Symmetrien korrigiert, d. h. der RMSD wird auf Basis der bestmöglichen Zuordnung topologisch identischer Atome berechnet.[341] cRAISE begegnet dem Docking-Problem durch die Hintereinanderausführung verschiedener Komponenten, die jeweils ein Teilproblem lösen – die Erzeugung von Konformeren, die Platzierung der Konformere und final die Priorisierung der Posen durch die Bewertungsfunktion. Um die Leistung einzelner Komponenten zu bewerten, kann der RMSD nach jeder Phase auf einer anderen Basis berechnet werden:

Qualität der Konformergenerierung: Erzeugte Konformationen werden bestmöglich auf den Kristallliganden superpositioniert. Der RMSD identifiziert die überlagerte Konformation mit minimaler Distanz ($\mathrm{RMSD}_{\mathrm{min\,Conf}}$) zum Kristallligand.

Qualität der Posengenerierung: Bei der Bewertung von Posen entfällt der Überlagerungsschritt. Der RMSD wird fix für jede erzeugte Platzierung berechnet und die beste Pose mit minimalem RMSD ($\mathrm{RMSD}_{\mathrm{min\,Pose}}$) identifiziert.

Qualität der Posenbewertung: Die Bewertungsfunktion wird durch den RMSD der Pose auf dem ersten Rang evaluiert ($\mathrm{RMSD}_{\mathrm{Top}}$). Unter der Annahme, dass der Kristallligand die beste Bewertung erhält und nahe Posen ähnlich gut bewertet werden, ist dort im Idealfall die bestplatzierte Pose zu finden. Die RMSD-minimale Pose, die innerhalb der vorderen X Ränge gefunden wird ($\mathrm{RMSD}_{\leq \mathrm{Top}X}$), ist zudem Indiz für die Etablierung einer adäquaten Rangfolge. Da die Posenbewertung den letzten Schritt des Docking-Prozess darstellt, bestimmt sie letztendlich die Leistung der Methode den nativen Bindungsmodus vorherzusagen.

Erfolgsraten: Für einen gegebenen Datensatz mit n Kristallkomplexen ist die *Erfolgsrate* der Konformergenerierung, der Platzierung oder der Posenbewertung durch den prozentualen Anteil aller erfolgreicher Vorhersagen definiert. Tabelle 7.1 listet RMSD-Kennzahlen, die hierbei über Erfolg und Fehlschlag entscheiden.[342]

Tabelle 7.1: Kritische RMSD-Werte [Å], die über Erfolg und Misserfolg der Konformergenerierung und des Redocking (Posengenerierung und -bewertung) entscheiden

Interpretation	Konformergenerierung	Redocking
Erfolg	$[0, 1]$	$[0, 2]$
Partieller Erfolg	$(1, 1.5]$	$(2, 3]$
Fehlschlag	$(1.5, \infty)$	$(3, \infty)$

7.1.3 Daten

Das Redocking wurde auf einem Datensatz durchgeführt, der von den Organisatoren des Docking Symposiums 2011 der American Chemical Society (ACS) gestellt wurde. Im Folgenden sei dieser Datensatz als $\text{Astex}_{\text{ACS}}$ bezeichnet. Er stellt eine überarbeitete Version des *Astex Diverse Sets*[343] dar, einem Datensatz, der Kristallstrukturen von 85 diversen Zielproteinen enthält. Der $\text{Astex}_{\text{ACS}}$ umfasst Kristallstrukturen dieser Zielproteine, allerdings mit verfeinerten Schweratomkoordinaten von Protein und Ligand. Zudem sind den Komplexen ausgerichtete Wasserstoffe hinzugefügt. Für monomere Proteine ist jeweils ein, für multimere Strukturen sind alle Kristallliganden als Referenz zur Definition von aktiven Bindetaschen bereitgestellt (*Referenzliganden*). In allen Docking-Berechnungen dieser Arbeit sind Bindetaschen *per se* durch Proteinatome charakterisiert, deren Zentren näher als 6,5 Å zu einem Atomzentrum des Referenzliganden liegen.[1] Für den Datensatz werden so insgesamt 146 gut aufgelöste Bindetaschen erhalten. Zusammen mit den zusätzlich zur Verfügung gestellten, nicht-kristallinen und entspannten Ligandstrukturen (*Startliganden*) bilden sie den Ausgangspunkt für die durchgeführten Docking-Berechnungen. Geringfügige Korrekturen, die an den Daten gemacht wurden, sind im Anhang A.1 zusammengefasst.

7.1.4 Experimente

Mit dem $\text{Astex}_{\text{ACS}}$ wurden Redocking-Läufe unter Verwendung der cRAISE-Docking-Software (vgl. Anhang C) durchgeführt. Bei gegebener Proteinstruktur, Referenz- und Startligand, erzeugt sie zunächst Konformere für den Startligand, die die Basis für die Deskriptorberechnung und die Erstellung eines temporären Index bilden. Anhand der Proteinstruktur und des Referenzliganden definiert cRAISE die Bindetasche, von der dann Anfragedeskriptoren abgeleitet und mit den Indexdeskriptoren abgeglichen werden. Jeder erhaltene Deskriptortreffer resultiert in einer Pose, die die Bewertungshierarchie durchläuft. Das Ergebnis eines solchen Docking-Laufes ist eine Liste von Posen, die gemäß der Bewertungsfunktion sortiert ist. Zur Auswertung wurden innerhalb eines Docking-Laufes $\text{RMSD}_{\text{min Conf}}$, $\text{RMSD}_{\text{min Pose}}$, $\text{RMSD}_{\text{Top-}}$, $\text{RMSD}_{\leq\text{Top5}}$, $\text{RMSD}_{\leq\text{Top20}}$ und $\text{RMSD}_{\leq\text{Top32}}$ berechnet. Für die Konformergenerierung, die Posengenerierung und die Posenbewertung wurden Erfolgsraten auf Grundlage des kompletten Datensatzes ($n = 146$) und auf Basis des qualitativ besten Resultats multimerer Zielproteine ($n = 85$) bestimmt. Da sich bei Multimeren die Qualität der Bindetaschen

[1] Weicht der sogenannte *Active Site Radius* in vereinzelten Fällen vom typischen Wert von 6,5 Å ab, so wird dies explizit erwähnt.

unterscheiden kann, kann ein Vergleich der Werte Aufschluss über die Präzision der Methode bei variierenden Eingaben liefern. Minimal-, Maximal-, Median-, Durchschnittswerte, Standardabweichungen und Konfidenzintervalle sollten zudem eine Aussage über die zu erwartenden Bindungsmodusvorhersagen bei Anwendung auf fremden Daten liefern. Die Editoren des Datensatzes berichteten bereits über Probleme und Fehler in den bereitgestellten Daten. Dennoch motivierten sie die Teilnehmer des Symposiums, Docking-Fehlschläge zu diskutieren, um die Vorhersageleistung in Abhängigkeit von äußeren Einflüssen zu veranschaulichen. Diese Aufgabe wurde in dieser Arbeit ebenso durchgeführt.

7.2 Anreicherung im virtuellen Screening

7.2.1 Bewertungsstrategie

Sind in einer Bibliothek von N Molekülen A bekannte bioaktive und D inaktive Moleküle (auch als *Decoys* bezeichnet) enthalten, dann soll ein Screening beide Submengen separieren. Die bioaktiven Moleküle sollen *angereichert* werden, d. h. möglichst zu Beginn der generierten Hitliste erscheinen. Die Anreicherung soll strukturspezifisch geschehen, d. h. Aktive sollen für ein bestimmtes Protein identifiziert jedoch nicht als Aktive eines anderen Proteins präsentiert werden. Im Folgenden werden Maße vorgestellt, um die Anreicherungsleistung einer Screening-Methode zu quantifizieren. Unter der Voraussetzung, dass die Mengen von Aktiven und Inaktiven bereits bekannt sind, können sie anhand der Hitliste direkt berechnet werden. Die Maße wurden in dieser Arbeit genutzt, da sie die gängigsten sind und einen Vergleich zu etablierten Methoden erlauben. Darüber hinaus existieren einige weniger etablierte Maße, die in dieser Arbeit jedoch nicht verwendet wurden, da ihr Nutzen bei der Bewertung der Anreicherung noch immer Gegenstand aktueller Diskussionen ist.[337]

7.2.2 Maße

Analyse der Receiver Operating Characteristic Kurve (ROC): Die ROC-Kurve visualisiert die Separierungsleistung einer Screening-Methode. Die Ordinate trägt die *Richtig-Positiv-Rate* (TPR oder Sensitivität) auf. Sie gibt das Verhältnis tatsächlich identifizierter Aktiver (TP) zur Gesamtzahl aller Aktiven ($TP + FN = A$) wieder. Die Abszisse trägt die *Falsch-Positiv-Rate (FPR)* auf. Sie gibt das Verhältnis verworfener Inaktiver (TN) zur Gesamtzahl aller Inaktiven ($TN + FP = D$) an:

$$TPR = \text{Sensitivität} = \frac{TP}{TP + FN} \tag{7.2}$$

$$FPR = 1 - \text{Spezifität} = 1 - \frac{TN}{TN + FP} \tag{7.3}$$

Die Gesamtfläche unter der ROC-Kurve oder *Area under the curve (AUC)* dient als Maß der Separierungsleistung. Sie kann anhand einer Hitliste unter Verwendung der Trapezregel angenähert werden:[344]

$$AUC = \frac{1}{D}\sum_{i=1}^{D} TPR_i = 1 - \frac{1}{A}\sum_{i=1}^{A} FPR_i \tag{7.4}$$

TPR_i bezeichnet die Richtig-Positiv-Rate bei Decoy i (die Anzahl Aktiver, die besser bewertet wurden als Decoy i, im Verhältnis zur Anzahl A aller Aktiven). Alternativ kann der AUC auch über FPR_i berechnet werden (die Falsch-Positiv-Rate beim Aktiven i). Bei optimaler Separierung, d. h. alle Aktiven werden zuerst gefunden, erreicht die ROC-Kurve mit einem AUC von 1 ihr Optimum. Ein Wert von 0,5 repräsentiert eine zufällige Auswahl von Aktiven und Inaktiven. Die Kurve verläuft dann entlang der Hauptdiagonalen. AUC-Werte zwischen 0,5 und 0 geben an, dass Inaktive besser als Aktive bewertet werden, ein Umstand, den es im Screening zu vermeiden gilt. Der AUC ist äußerst intuitiv, da er die Wahrscheinlichkeit widerspiegelt mit der Aktive gefunden werden. Allerdings kann er nicht zwischen früher und später Anreicherung unterscheiden. Häufig wird die ROC-Kurve deshalb punktuell im vorderen Bereich ausgewertet, indem die TPR bei einer FPR von 1%, 2% und 5% ($\text{ROC}_{\text{x\%}}$) bestimmt wird.

Zielstrukturspezifische Anreicherung: Seien eine Zielstruktur t mit zugehörigen Mengen von Aktiven A_t und Inaktiven D_t und eine Anti-Zielstruktur nt gegeben. Falls eine strukturbasierte Methode Aktive alleinig auf Basis von Ligandinformation identifiziert und die durch das Protein gegebene Information missachtet, dann werden die Aktiven A_t sowohl beim Screening mit t als auch beim Screening mit nt angereichert. Zur Widerlegung dieser Nullhypothese kann der AUC für das Screening mit t und mit nt ($tAUC$ und $ntAUC$) bestimmt werden. Die Differenz

$$\Delta AUC = tAUC - ntAUC \tag{7.5}$$

gibt dann die Leistung zur zielstrukturspezifischen Unterscheidung von Aktiven an. Im Idealfall liefert das Screening einer Anti-Zielstruktur gegen die Bibliothek einer Zielstruktur einen AUC von 0,5 und zeigt damit an, dass die Anti-Zielstruktur nicht in der Lage ist zwischen Aktiven und Inaktiven der eigentlichen Zielstruktur zu unterscheiden. Wenn $tAUC$ und $ntAUC$ optimal sind (1,0 und 0,5), dann deutet ein positiver ΔAUC-Wert von 0,5 zielstrukturspezifische Anreicherung und geringere Werte eine eher unspezifische Anreicherung an. Ein Wert nahe 0 kann allerdings auch andeuten, dass beide

Zielstrukturen verwandt sind und die Aktiven tatsächlich an beide Proteine binden. Negative Werte zeigen an, dass Decoys der Originalzielstruktur durch die Anti-Zielstruktur angereichert wurden, was mit einem idealen Datensatz prinzipiell nicht geschehen sollte.

7.2.3 Daten

Anreicherungsstudien wurden auf dem DUD_{ACS}-Datensatz durchgeführt, der ebenso von den Organisatoren des ACS Docking Symposiums 2011 gestellt wurde. Er umfasst die 40 medizinal-chemisch relevanten Zielstrukturen des DUD-Datensatzes[286], die den Proteinfamilien der Serinproteasen, Kinasen, Metalloenzyme, Kernrezeptoren, Folatenzymen und anderer Enzyme zuzuordnen sind. Jede Zielstruktur ist mit Protein, Referenzligand und ausgezeichneten Mengen von aktiven und inaktiven Molekülen ausgestattet. Die Strukturen wurden von den Organisatoren analog zum $Astex_{ACS}$ präpariert, mit Ausnahme, dass die Proteinstrukturen mit entscheidenden Kristallwassern versehen wurden. Die Organisatoren des Symposiums boten außerdem semi-zufällig gewählte Anti-Zielstrukturen. Für jede der 40 Zielstrukturen wurde hierfür eine andere, zumeist aus der selben Proteinfamilie ausgewählt.

7.2.4 Experimente

Mit dem DUD_{ACS} wurden 40 individuelle Screening-Läufe unter Verwendung der cRAISE-Screening-Software (vgl. Anhang C) durchgeführt. Vorbereitend wurden 40 Indizes erzeugt, die Deskriptoren der aktiven und inaktiven Konformere umfassten. Die Screening-Phase definierte die Bindetasche anhand der gegebenen Proteinstruktur und des Referenzliganden und leitete Anfragedeskriptoren ab. Die Anfragedeskriptoren wurden mit den Indexdeskriptoren abgeglichen und die erhaltenen Posen durch die Bewertungshierarchie bewertet. Die Bewertung jeder Top-Pose wurde in der Hitliste registriert und final gemäß der Bewertungsfunktion sortiert. Die Verteilung der Aktiven in den resultierenden Hitlisten wurde analysiert und ROC-basierte Anreicherungswerte bestimmt. Um prospektiv eine Aussage über die zu erwartende Anreicherungsleistung auf unbekannten Bibliotheken zu geben, wurden Minimal-, Maximal-, Median-, Durchschnittswerte, Standardabweichungen und Konfidenzintervalle berechnet. Die Resultate dienten für einen Vergleich mit veröffentlichten Anreicherungswerten gängiger, strukturbasierter Screening-Methoden, die ebenso auf dem Datensatz bewertet wurden.[345, 346, 347, 348, 349, 350, 351, 352] Die semi-zufällig gewählten Paare von Zielstrukturen dienten zur Untersuchung der Nullhypothese. Hierfür wurden zusätzlich Screening-Läufe mit den definierten Anti-Zielstrukturen durchgeführt. Die Resultate dienten zur Diskussion, ob die Anreicherung zielstrukturspezifisch geschieht.

7.3 Laufzeit und Selektivität

7.3.1 Bewertungsstrategie

Bei der Analyse des Laufzeitverhaltens galt es, die laufzeitbestimmenden Schritte der Screening-Phase ausfindig zu machen und Faktoren zu identifizieren, die diese beeinflussen. Dafür wurden Laufzeiten gemessen und verschiedenen Kenngrößen gegenübergestellt. Im Speziellen wurde die Abhängigkeit der Screening-Zeit vom Umfang der gescreenten Bibliothek (Anzahl der Moleküle, der Konformationen und der Indexdeskriptoren), vom Umfang der Anfrage (Anzahl der Anfragedeskriptoren) und vom Umfang der Screening-Resultate (Anzahl der Deskriptortreffer) untersucht.

7.3.2 Maße

Laufzeiten: Sei N die Größe der Bibliothek (eindeutige Moleküle), dann kann die Laufzeit eines flexiblen Dockings durch $t_f = t_{total}/N$ abgeschätzt werden. Sei M die Anzahl der Konformationen in der Bibliothek, dann schätzt $t_s = t_{total}/M$ die Laufzeit eines starren Dockings ab. Die Gesamtlaufzeit t_{total} bezeichne die Summe der Laufzeiten (wall-clock time) aller Prozesse, die für ein Screening notwendig sind. Wird ein Index partitioniert und auf lokale Festplatten eines Clusters verteilt, dann entspricht t_{total} der Summe der Screening-Zeiten aller Partitionen, da cRAISE vollständig und unabhängig eine Partition in einen Prozess bearbeitet. Die Summe ist abhängig von der Größe der Bibliothek, jedoch unabhängig von der Anzahl der Pakete und der verfügbaren Prozessorkerne. Sei K die Anzahl frei verfügbarer Kerne auf einem Cluster, dann kann die parallele Laufzeit durch $t_p = t_{total}/K$ abgeschätzt werden. Sie reflektiert die bestmögliche Laufzeit in einer parallelen Umgebung mit K frei verfügbaren Kernen.

Selektivität σ: Enthält ein Index n Deskriptoren, dann kann ein einzelner Anfragedeskriptor prinzipiell alle Indexdeskriptoren extrahieren. Wenn m Anfragedeskriptoren von einem Zielprotein abgeleitet werden, kann dies im schlimmsten Fall zu $T = n \cdot m$ Deskriptortreffern führen. Auch wenn dieser Fall nie eintritt, so gibt die Selektivität σ an wie häufig der gesamte Index durch die Anfrage tatsächlich extrahiert wird:

$$\sigma = \frac{T}{n} \tag{7.6}$$

7.3.3 Daten

Großangelegte Screening-Studien wurden auf Submengen der ZINC-Datenbank[277, 276] durchgeführt. Der *ZINC Clean Leads*-Datensatz[353] umfasste 4 230 832 leitstrukturähnliche Moleküle. In seiner ursprünglichen Form ist er durch ein Präparierungspro-

tokoll der Herausgeber mit Stereoisomeren, Protonierungszuständen und Tautomeren angereichert.[276] Da cRAISE Letztere als Duplikate identifiziert, wurden zufällig ein, zwei und drei Millionen eindeutiger Moleküle aus dieser Menge gewählt. Diese Selektionen seien folgend als $\text{ZINC}_{\text{CL1M}}$-, $\text{ZINC}_{\text{CL2M}}$- und $\text{ZINC}_{\text{CL3M}}$-Bibliothek bezeichnet.

7.3.4 Experimente

Zur Laufzeitbewertung wurde jeweils ein Index mit zugehöriger Moleküldatenbank für die $\text{ZINC}_{\text{CL1M}}$-, $\text{ZINC}_{\text{CL2M}}$- und $\text{ZINC}_{\text{CL3M}}$-Bibliotheken präpariert. Partitionen (à 2500 Moleküle, 6,5 GB) wurden vor einem Screening auf den lokalen Festplatten eines High-Performance-Computing-(HPC)-Clusters verteilt. Der heterogene Cluster bestand aus Knoten dreier Intel-Xeon-CPU-Generationen aus den Jahren 2007 (E5410), 2010 (E5630, E5640) und 2012 (E5-2680). Jeder Kern prozessierte jeweils einen Teil des Index und bekam maximal 8 GB Arbeitsspeicher zur Verfügung gestellt. Aus dem DUD_{ACS} wurden repräsentative Zielstrukturen gewählt und deren Anfragedeskriptoren mit den $\text{ZINC}_{\text{CL1M}}$-, $\text{ZINC}_{\text{CL2M}}$- und $\text{ZINC}_{\text{CL3M}}$-Indizes abgeglichen. Für jeden Index wurde die Anzahl enthaltener Deskriptoren, Konformere und Moleküle registriert. Bei jedem durchgeführten Screening wurde zudem die Anzahl der Anfragedeskriptoren, extrahierte Deskriptortreffer, bewertete Posen und σ bestimmt. Zusätzlich zu den Zeiten t_s und t_f, die für ein starres und flexibles Docking aufzuwenden sind, zur Gesamtlaufzeit t_{total} und zur parallelen Laufzeit t_p mit $K = 200$, wurden Zeiten auch nach der Ausführung einzelner Prozesskomponenten gemessen und ihr prozentualer Anteil zur Gesamtlaufzeit berechnet.

7.4 Effekt geleiteter Vorhersagen

7.4.1 Bewertungsstrategie

Ein Pharmakophormodell kann einen positiven Effekt auf das Resultat einer damit geleiteten Vorhersage haben. Idealerweise lenkt es das Docking in die richtige Richtung und verstärkt die Anreicherung im Screening. Außerdem kann es den Suchraum derart einschränken, sodass ein günstiger Laufzeiteffekt merkbar wird. Um diese erwarteten Effekte zu quantifizieren, wurden pharmakophorgeleitete Redocking-, Anreicherungs- und Laufzeitexperimente durchgeführt und die Resultate mit den Ergebnissen der bisher beschriebenen Experimente verglichen. Zudem wurde die Anwendbarkeit von Molekülprofilen bewertet. Der molekülprofilgeleitete Ansatz ist nur dann methodisch vertretbar, wenn eine Molekülfilterung während des Screenings, im Vergleich zu einer vor-

oder nachbereitenden Filterung, keinen Laufzeitnachteil mit sich bringt. Dass diese Anforderung durch cRAISE tatsächlich eingehalten wird, wurde exemplarisch durch die Analyse eines molekülprofilgeleiteten Screening demonstriert.

7.4.2 Maße

Signifikanz der Vergleiche: Der Unterschied zwischen geleiteter und nicht geleiteter Bindungsmodusvorhersage wurde über RMSD-Differenzen auf verschiedenen Rängen der Listen bewerteter Posen ermittelt. Für einen Datensatz der Größe n, kann mit Hilfe von Paardifferenzentests (*paired t-test*) die Signifikanz dieser Vergleiche bestimmt werden. Unter der Annahme, dass die gepaarten Differenzen unabhängig und gleichermaßen normalverteilt sind, ist die Wahrscheinlichkeit $p \ll 1$ genau dann, wenn die Differenzen zwischen geleiteter und nicht geleiteter Vorhersage nicht zufällig sind. Ein geringer Wert deutet auf einen signifikanten Unterschied hin und ist somit Indiz für einen Effekt.

Effektstärke: *Cohen's d* ist ein substantielles Maß zum Messen der Effektstärke. Es betrachtet den Mittelwertunterschied $\overline{x_1} - \overline{x_2}$ zweier Messreihen der Größe n im Verhältnis zu deren geschätzten Varianzen s_1^2 und s_2^2:

$$d = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(s_1^2 + s_2^2)/2}}, \quad s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ji} - \overline{x_i})^2 \tag{7.7}$$

Nach Cohen[354] ist ein Wert von 0,2 gleichbedeutend mit einem kleinen Effekt. Ein mittlerer Effekt ist bei einem Wert von 0,5 und ein starker Effekt bei 0,8 angezeigt.

7.4.3 Daten

Definition von Pharmakophorhypothesen: Anwenderdefinierte Pharmakophormodelle führen einen Bias in Docking-Berechnungen ein. Obwohl dies bei der Anwendung gewollt ist, hat es für die Evaluierung einer pharmakophorgeleiteten Methode einen entscheidenden Nachteil — sie ermöglichen nicht ohne Weiteres die Reproduzierbarkeit und Objektivität der Experimente. Aus diesem Grund wurden anhand eines gegebenen Protein-Ligand-Komplexes Modelle automatisiert nach folgendem Schema abgeleitet:

1. Inklusionsmerkmale werden initial an Stellen innerhalb der aktiven Bindetasche platziert, die gemäß des cRAISE-Interaktionsmodells, komplementär zu Proteininteraktionsstellen (vgl. Abschnitt 6.1) sind. Dadurch werden Merkmale nicht auf Ligandatome, sondern relativ zu Proteinatomkoordinaten platziert.
2. Den zugehörigen Toleranzsphären wird ein für strukturbasierte Modelle typischer Radius von 1,7 Å zugewiesen.

3. Diese Inklusionen werden auf Merkmale eingeschränkt, die mutmaßlich die Ligandbindung propagieren. Dies sind Merkmale, die gemäß des cRAISE-Interaktionsmodells komplementär zu den Interaktionsstellen des gegebenen Liganden sind.
4. Von jeder so erhaltenen Menge an Inklusionen werden n verschiedene Modelle erzeugt, die sich lediglich darin unterscheiden, wie viele der n Inklusionen tatsächlich während der Platzierung durch eine Pose erfüllt werden müssen. Dazu wird die Anzahl essentieller Inklusionsmerkmale N_e jeweils auf ein $N_e \in \{1, \ldots, n\}$ gesetzt.

Eine Relaxation von N_e hat zwei praktischen Anwendungen: Zum einen kann ein strukturbasiertes Modell zu Merkmalskonstellationen führen, die zur vollständigen Erfüllung gespannte oder unrealistische Posen erfordern würden. Zum anderen ist es möglich, dass im Screening unterschiedliche Klassen von Liganden unterschiedliche Merkmalskonstellationen benötigen. Entspannte Modelle begegnen beiden Situationen und gleichen lediglich Submengen von Merkmalen mit Mindestgröße N_e ab.

Definition eines beispielhaften Molekülprofils: Exemplarisch wurde ein einfaches Screening-Szenario konstruiert: Die $\text{ZINC}_{\text{CL3M}}$-Bibliothek beinhaltet drei Millionen leitstrukturähnliche Moleküle (vgl. Abschnitt 2.5.2). Mit Mona wurde ein Molekülprofil erstellt, das die Bibliothek weiter, auf Moleküle mit einem Molekulargewicht von 300, maximal fünf rotierbaren Bindungen und einem logP von höchstens 3,5 beschränkt. Das Profil ist in Anhang A.3 bereitgestellt. Es wurde dazu genutzt, um seine Auswirkung auf die Laufzeit beim Screening gegen den DUD_{ACS}-Estrogenrezeptoragonisten zu bestimmen, wenn die Bibliothek zugleich gefiltert wird.

7.4.4 Experimente

Es wurden Pharmakophormodelle für die 146 Komplexe des $\text{Astex}_{\text{ACS}}$ und die 40 Komplexe des DUD_{ACS} automatisiert erstellt. Die Redocking- und Anreicherungsexperimente 7.1 – 7.3 wurden wiederholt, wobei nun die Modelle den Docking- bzw. Screening-Prozess lenkten. Die erhaltenen Ergebnisse wurden mit den Ergebnissen nicht geleiteter Vorhersagen verglichen und der Einfluss der Modelle auf die Bindungsmodusvorhersage, die Anreicherung im Screening und die Laufzeit untersucht. Hierfür wurden Paardifferenzentests durchgeführt und Cohen's d ermittelt, um die Signifikanz der Vergleiche zu bestimmen und den hervorgerufenen Effekt zu quantifizieren. Außerdem wurden die in Abschnitt 7.3 beschriebenen Laufzeitexperimente unter Verwendung des beispielhaften Molekülprofils durchgeführt und dessen Auswirkung auf die Laufzeit betrachtet.

7.5 Auswirkungen von Zustandsänderungen

7.5.1 Bewertungsstrategie

Das Ziel des Dockings molekularer und makromolekularer Zustände ist es, konsistente Bindungsmodusvorhersagen für jeglichen Eingabezustand von Protein und Ligand zu gewährleisten. Zudem soll der cRAISE-Multizustandsansatz hierbei vergleichbare Resultate zu denen eines naiven Ensemble-Dockings (vgl. Abschnitt 3.8) liefern und zum umfangreichen Screening anwendbar bleiben. Um diese Ziele zu verifizieren, wurden Experimente analog zu 7.1 – 7.3 durchgeführt, allerdings auf einer anderen Datengrundlage. Zudem wurde anhand ausgewählter Literaturbeispiele gezeigt, dass es nicht ausreichend ist, lediglich den wahrscheinlichsten Grundzustand von Ligand und Rezeptor im Docking und Screening anzunehmen. Rezeptorseitig wird dies von gängigen strukturbasierten Methoden allerdings propagiert, indem nur ein wahrscheinlicher Zustand präpariert und die Zustandsenumeration völlig vernachlässigt wird. Die Auswirkungen für das Screening wurden deshalb demonstrativ, anhand eines Rezeptor-Ligand-Systems untersucht und veranschaulicht.

7.5.2 Maße

Bevorzugter Zustand: Seien zwei unterschiedliche Zustände S_1 und S_2 eines Komplexes gegeben, deren Bindungspotential gemäß einer Bewertungsfunktion X durch $\Delta G_X(S_1)$ und $\Delta G_X(S_2)$ eingeschätzt wird, so gibt

$$\Delta\Delta G_X = \Delta G_X(S_1) - \Delta G_X(S_2) \tag{7.8}$$

den günstigeren der beiden Zustände an. Ist der $\Delta\Delta G_X$ positiv, so wird der Zustand S_2 bevorzugt. Ist $\Delta\Delta G_X$ negativ, dann wird S_1 favorisiert.

7.5.3 Daten

Relevante $\mathbf{Astex_{ACS}}$-Komplexe: Der $\text{Astex}_{\text{ACS}}$ umfasst Kristallstrukturen von 85 Zielproteinen. Nicht alle erfordern Zustandsänderungen während der Bindungsmodusvorhersage, da weder die aktive Bindetasche noch der gebundene Ligand ihren Zustand ändern könnten. Die durchgeführten Redocking-Experimente berücksichtigen daher nur 56 *relevante* Komplexe, die nach folgenden Kriterien ausgewählt wurden:

- Der Ligand, die aktive Bindetasche oder beide besitzen Multizustandsatome.
- Hat nur der Rezeptor solche Atome, muss zumindest eines wasserzugänglich sein.

Für diese Komplexe wurden die wahrscheinlichsten Zustände des Rezeptors und Startliganden gemäß Abschnitt 6.15.1 aufgezählt und individuell in Molekülдateien festgehalten.

Rechenintensive DUD-E-Komplexe: Innerhalb von Screening-Experimenten sollte der Aufwand für eine simultane Durchmusterung von Rezeptor- und Ligandzuständen ermittelt werden, die jedoch nur dann erfolgt wenn

- wasserzugängliche Multizustandsatome rezeptor- und ligandseitig vorzufinden sind.

Da lediglich vier DUD_{ACS}-Komplexe dieses Kriterium erfüllen, wurde der DUD-E-Datensatz[355] herangezogen (eine Neuauflage des DUD-Datensatzes mit 102 Zielproteinen). Innerhalb des DUD-E konnten 14 solcher Komplexe identifiziert werden. Die entsprechenden aktiven und inaktiven Molekülmengen, die im DUD-E gegeben sind, wurden für die Laufzeitstudien genutzt. Sie können als besonders rechenintensiv betrachtet werden, da die möglichen Rezeptor-Ligand-Zustandskombinationen den Suchraum eines Einzel-Dockings entscheidend vergrößern.

Ricin A in Komplex mit Neopterin (PDB `1br5`): Zur Veranschaulichung wie die Multizustandsmethode eine notwendige, aber initial nicht offensichtliche Zustandsänderung des Liganden bei der Bindung löst, wurde ein in der Literatur diskutiertes Beispiel herangezogen.[239, 289] Im Fall von Ricin A im Komplex mit Neopterin (PDB: `1br5`[356]) nimmt die Pterin-Gruppe des Liganden in Wasser gelöst das stabilere 3H-Tautomer ein. Ist es jedoch im Komplex gebunden, so wird das eigentlich weniger wahrscheinliche 1H-Tautomer eingenommen.

HID110- und HIE110-Ensemble der Aldosereduktase: Für die Aldosereduktase (ALDR) existieren äußerst hoch aufgelöste Kristallstrukturen, die Wasserstoffpositionen im Komplex offenbaren. Die Strukturen zeigen, dass eine Änderung des Histidinzustands eine entscheidende Rolle bei der Bindung unterschiedlicher Inhibitorklassen spielen kann: Die meisten ALDR-Inhibitoren gehören entweder zu den Carboxylaten oder Spirohydantoinen, für die IDD594 und Fidarestat Repräsentanten sind. Bei physiologischem pH sind Spirohydantione neutral, wohingegen die Carboxylate im geladenen Zustand verbleiben.[1][357, 358] Beide Inhibitorklassen ähneln sich stark bzgl. der Wasserstoffbrücken, die sie mit drei Schlüsselresiduen (Tyr48, His110 und Trp111) bei der Bindung etablieren. Die Kristallstruktur der ALDR im Komplex mit IDD594 (PDB: `1us0`) zeigt, dass das Carboxylat an das $N^{\epsilon 2}$-H-Tautomer von His110 bindet.[2][359] Wird stattdessen die neutrale Hydantoin-Gruppe gebunden, so muss ein anderer Zustand

[1] Dies erklärt die höhere Wirksamkeit der Spirohydantione *in vivo*, da ihnen ihr neutraler Zustand bei physiologischem pH erlaubt, Membranen einfacher zu durchqueren als Inhibitoren der Säureklasse.

[2] Dieser His110-Zustand wird durch ein am $N^{\delta 1}$-Atom nah gelegenes Wassermolekül induziert.

von His110 eingenommen werden (PDB: `1pwm`).[1][360] Aufgrund dieser Beobachtungen wurde ein System zur Studie der Anreicherung bei variierendem Rezeptorzustand etabliert: Aus der PDB wurden insgesamt 50 ALDR-Strukturen extrahiert, die anhand des gebundenen Inhibitors in IDD594-ähnliche (43 Komplexe) und Fidarestat-ähnliche (7 Komplexe) klassifiziert wurden. Im Anhang A.4 ist die Klassifizierung vorzufinden. Den Komplexen wurden Wasserstoffe hinzugefügt, das Wasserstoffbrückennetzwerk mit PROTOSS optimiert und die Komplexe anhand ihrer Bindetaschenatome überlagert. Für beide Komplexklassen wurden je zwei Varianten erstellt – das *HID110-* und das *HIE110-Ensemble* – deren überlagerte Strukturen sich bis auf das entsprechende Tautomer von His110 nicht weiter unterscheiden.

7.5.4 Experimente

Für jede mögliche Kombination von Rezeptor- und Ligandeingabezuständen der relevanten $\text{Astex}_{\text{ACS}}$-Komplexe wurde der Bindungsmodus statisch (d. h. unter Erhaltung des Eingabezustands) und mit dem CRAISE-Multizustandsansatz (d. h. unter Adaption der Zustände) vorhergesagt. Die Ergebnisse der statischen Redocking-Läufe dienten der Quantifizierung der Unterschiede bei variierenden Eingabezuständen. Die Ergebnisse der Multizustandsvorhersagen verifizierten die Unabhängigkeit der Vorhersagen vom Eingabezustand. Zudem wurden die statischen Redocking-Resultate für jeden Komplex entsprechend der naiven Vorgehensweise eines Ensemble-Dockings (vgl. Abschnitt 3.8) vereinigt. Das Ergebnis wurde mit dem Resultat des Multizustandsansatz verglichen, um festzustellen ob beide zum selben Ergebnis führen. Die als mutmaßlich rechenintensiv klassifizierten DUD-E-Mengen wurden zur Analyse der Laufzeit genutzt. Hierbei wurden der statische, der naive Ensemble- und der Multizustandsansatz bezüglich der in Abschnitt 7.3 vorgestellten Maße verglichen und mit der Anzahl der Rezeptorzustände, der Ligandzustände und deren Kombinationen gegenübergestellt. Das System von Ricin A im Komplex mit Neopterin diente zur Studie einer notwendigen Ligandzustandsänderung während des Docking-Prozesses. Dafür wurden einzelne statische Docking und Multizustandsvorhersagen mit unterschiedlichen Eingabezuständen durchgeführt. Das ALDR-System diente der Untersuchung von Zustandsänderungen auf Rezeptorseite. Die HYDE-Bewertungsfunktion[347, 361] wurde verwendet, um die ALDR-Ensembles unter den folgenden Bedingungen zu bewerten:

[1] Das $N^{\epsilon 2}$-Tautomer verhindert die Etablierung einer Wasserstoffbrücke. Deshalb wurde für die Bindung von Fidarestat erklärt, dass der Bindungsvorgang durch die Abgabe des Wassermoleküls und der Aufnahme eines Chlorid-Ions begleitet wird. Das Ion forciert die Zustandsänderung von His110, welches das Wasserstoffatom des Stickstoffatoms an der 1'-Position des Hydantoins akzeptiert und so zu einer elektrostatischen Interaktion zwischen nun geladenem Histidin und geladenen Inhibitor führt.

1. Die 50 Kristallkomplexe wurden individuell im HID110- und im HIE110-Zustand bewertet.
2. Eine einzelne Proteinstruktur (PDB: `1su0`) wurde im HID110- und HIE110-Zustand mit jedem der 50 zuvor überlagerten Liganden bewertet.
3. Die 50 Inhibitoren wurden gegen `1su0` jeweils im HID110- und HIE110-Zustand gedockt. Die erhaltenen Top-Posen wurden bewertet.
4. Die 129 IDD594- und 30 Fidarestat-ähnlichen ALDR-Aktiven des DUD-E wurden gegen den HID110- und HIE110-Zustand von `1su0` gedockt. Die erhaltenen Top-Posen wurden bewertet.

Für alle korrespondierenden HID110- und HIE110-Bewertungen wurde $\Delta\Delta G_{\mathrm{Hyde}}$ bestimmt, um den bevorzugten Rezeptorzustand im jeweiligen Kontext zu ermitteln.

8 Resultate und Diskussion

Dieses Kapitel fasst die Resultate der in Kapitel 7 beschriebenen Experimente zusammen und diskutiert die Leistung von cRAISE bezüglich der dort vorgestellten Maße und Daten. In den ersten Abschnitten werden die Ergebnisse der Bindungsmodusvorhersagen, der Anreicherungs- und Laufzeitexperimente vorgestellt. Die Resultate bilden die Grundlage für den darauffolgenden Vergleich zum pharmakophor- und molekülprofilgeleiteten cRAISE-Ansatz. Abschließend wird der cRAISE-Multizustandsansatz dem naiven Ensembleansatz gegenübergestellt und die Notwendigkeit diskutiert, Zustände ligand- und rezeptorseitig im Docking- und Screening-Prozess zu integrieren.

8.1 Bindungsmodusvorhersagen

8.1.1 Konformergenerierung

Als erste Instanz in der Docking-Maschinerie ist die Konformergenerierung ein entscheidender Faktor für eine erfolgreiche Bindungsmodusvorhersage mit cRAISE. Um erfolgreich Koordinaten von Liganden zu reproduzieren, muss zumindest ein Konformer existieren das die Konformation des aktiven Bindemodus widerspiegelt. Um ein effizientes Screening zu gewährleisten, sollte zudem die Anzahl erzeugter Konformere möglichst gering sein. Da mit den ursprünglichen Confect-Qualitätsstufen (vgl. Abschnitt 4.4) die Anzahl erzeugter Konformere nicht reduziert und zugleich deren Qualität erhalten werden konnte, wurden verfeinerte Qualitätsstufen eingeführt. Aus ihnen wählt cRAISE individuell die optimale Qualitätsstufe für jedes Molekül (vgl. Abschnitt 6.4). Tabelle 8.1 vergleicht die Erfolgsraten der Konformergenerierung auf dem $\text{Astex}_{\text{ACS}}$ mit dieser automatisch ermittelten cRAISE-Qualitätsstufe (QS AUTO) und mit verschiedenen Confect-Qualitätsstufen (QS 1, QS 2, QS 3). Die Anzahl generierter Konformationen wurde jeweils auf 100, 250, 500 und 1 000 limitiert. Die cRAISE-Methode zeigt im Vergleich zu

Tabelle 8.1: Erfolgsraten der Konformergenerierung auf dem $\text{Astex}_{\text{ACS}}$ mit den Qualitätsstufen QS 1, QS 2, QS 3 und QS AUTO. Die Anzahl generierter Konformere ist auf 100, 250, 500 und 1000 beschränkt. QS AUTO (orange) wird standardmäßig von cRAISE verwendet. Alternativ können die grau hinterlegten Stufen eingestellt werden.

		r_{min}-erfolgreich reproduzierte Kristallliganden (%)					
QS	Limit	≤0.5 Å	**≤1.0 Å**	≤1.5 Å	≤2.0 Å	≤2.5 Å	≤3.0 Å
1	100	45.0	**74.8**	96.0	96.7	98.7	100.0
	250	46.4	**78.1**	96.0	97.4	98.7	100.0
	500	47.0	**79.5**	97.4	98.7	98.7	100.0
	1000	49.7	**84.1**	98.7	98.7	99.3	100.0
2	100	40.4	**71.5**	93.4	96.0	98.7	100.0
	250	46.4	**74.2**	93.4	96.0	99.3	100.0
	500	53.6	**77.5**	95.4	97.4	99.3	100.0
	1000	55.6	**80.1**	96.7	98.7	99.3	100.0
3	100	39.1	**71.5**	92.1	94.0	98.7	100.0
	250	43.7	**72.8**	94.0	96.0	98.7	100.0
	500	47.7	**74.2**	94.0	96.7	99.3	100.0
	1000	53.6	**76.2**	95.4	98.0	99.3	100.0
AUTO	100	49.0	**78.8**	98.7	99.3	100.0	100.0
	250	55.0	**85.4**	99.3	99.3	100.0	100.0
	500	59.6	**88.1**	99.3	99.3	100.0	100.0
	1000	61.6	**88.7**	99.3	99.3	100.0	100.0

den herkömmlichen Confect-Qualitätsstufen für jedes gegebene Limit durchweg bessere Erfolgsraten. Auch die übrigen Confect-Qualitätsstufen (vgl. Tabelle 4.5) erzielen keine Erfolgsraten, die mit der automatisch ermittelten Qualitätsstufe QS AUTO vergleichbar wären. In all den hier vorgestellten Docking- und Screening-Berechnungen ist das Limit erzeugter Konformere stets auf 250 gesetzt (orange hinterlegt). Damit werden für rund 85% des Datensatzes erfolgreich eine Konformation mit einem $\text{RMSD}_{\text{min Conf}}$ von weniger als 1 Å erzeugt. Bezüglich des relaxierten Kriteriums von 1,5 Å (vgl. Tabelle 7.1) kann ein partieller Erfolg sogar für 99% der Liganden verzeichnet werden. Da die Generierung von Ringkonformationen in Confect unabhängig und nicht über die Qualitätsstufen steuerbar ist, sind hohe RMSD-Werte vor allem bei solchen Molekülen zu beobachten, die Ringsysteme enthalten. Deshalb kann für `1yqy`, `1mzc`, `1s19`, `1t46`, `1q1g`, `1kzk`, selbst mit einem Limit von 1 000, keine bessere Konformation erzeugt werden. Die qualitative Verbesserung erreicht cRAISE durch die Strategie der dynamischen Auswahl einer fein justierten Qualitätsstufe (AUTO EXTEND 5 – AU-

TO REDUCE 3) in Abhängigkeit von Größe und Flexibilität eines Moleküls. Die Strategie erlaubt es, den Konformationsraum unter Einhaltung einer vorgegebenen Schranke abzutasten und das gegebene Limit dabei bestmöglich auszuschöpfen. Abbildung 8.1 veranschaulicht dies und vergleicht die Anzahl tatsächlich erzeugter Konformationen bei QS 2 und QS AUTO. Im Gegensatz zur CONFECT-Qualitätsstufe nutzt die cRAISE-Qualitätsstufe das gegebene Limit von 250 Konformationen zumeist voll aus und erreicht so bessere $\mathrm{RMSD}_{\mathrm{min\,Conf}}$-Werte. Bei relativ starren Molekülen, deren Bindungen Torsionssignaturen mit wenigen Peaks besitzen, sind beide Methoden vergleichbar. In solchen Fällen stellen sowohl CONFECT als auch cRAISE zusätzlich zu den Peaks Torsionen des zweiten Toleranzniveaus ein.

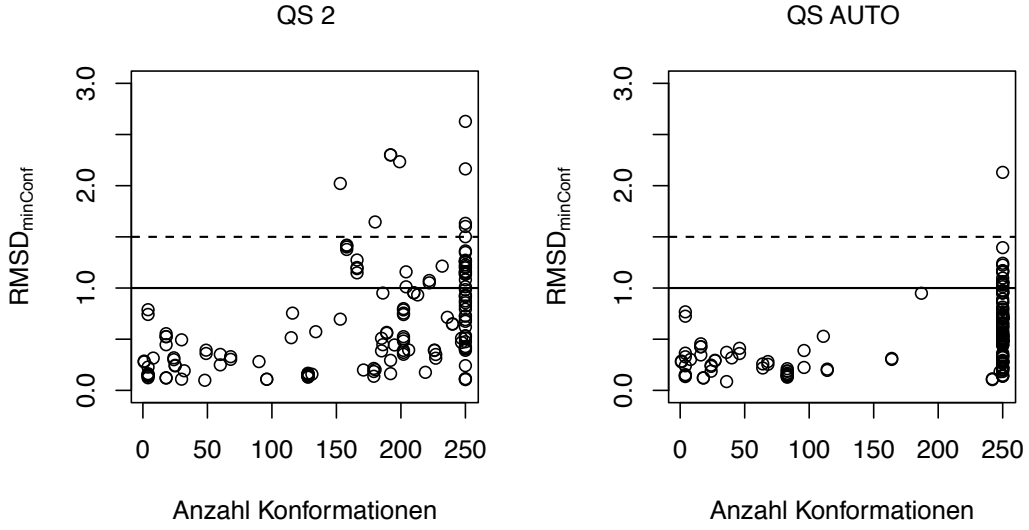


Abbildung 8.1: Anzahl erzeugter Konformationen gegen $\mathrm{RMSD}_{\mathrm{min\,Conf}}$ bei QS 2 und QS AUTO und maximal 250 erzeugten Konformationen mit $\mathrm{Astex}_{\mathrm{ACS}}$-Startliganden.

8.1.2 Vorhersage des aktiven Bindungsmodus

Die Fähigkeit den nativ gebundenen Liganden zu reproduzieren wurde, wie in Abschnitt 7.1 beschrieben, auf dem $\mathrm{Astex}_{\mathrm{ACS}}$ evaluiert. Tabelle 8.2 zeigt die durchschnittlichen, mittleren, minimalen und maximalen RMSD-Werte, die die Konformergenerierung ($\mathrm{RMSD}_{\mathrm{min\,Conf}}$), die Posengenerierung ($\mathrm{RMSD}_{\mathrm{min\,Pose}}$) und die Bewertungshierarchie ($\mathrm{RMSD}_{\mathrm{Top}}$) beim Docking in alle 146 Bindetaschen erreichte. Durchschnittlich produziert cRAISE Konformere, die um 0,5 Å, und Platzierungen, die um 0,9 Å vom Kristallliganden abweichen. Die Platzierung mittels Deskriptorabgleich lässt sich somit erfolgreich mit cRAISE-Konformeren realisieren. Sie führt lediglich zu einer weiteren

Tabelle 8.2: $\text{RMSD}_{\text{min Conf}}$, $\text{RMSD}_{\text{min Pose}}$ und RMSD_{Top} auf dem $\text{Astex}_{\text{ACS}}$ ($n = 146$). Fehlerbereiche entsprechen 95%-Konfidenzintervallen.

	$\text{RMSD}_{\text{min Conf}}$	$\text{RMSD}_{\text{min Pose}}$	RMSD_{Top}
Mean	0.52 (±0.05)	0.93 (±0.07)	2.06 (±0.27)
SD	0.33	0.44	1.69
Median	0.47	0.92	1.50
Min	0.09	0.25	0.25
Max	2.13	2.91	10.53

Diskrepanz von 0,4 Å, die aber bei Nutzung einer diskreten Platzierungsmethode toleriert werden muss. Die Bewertungshierarchie erzielt durchschnittlich Vorhersagen mit einer Abweichung von 2,0 Å auf dem ersten Rang. Der RMSD_{Top}-Median deutet an, dass die Mehrheit der Liganden erfolgreich gedockt werden kann. Gelegentlich wird ein partieller Erfolg und seltener ein Docking-Fehlschlag registriert. Ob die Fehlschläge auf qualitativ schlechte Konformationen bzw. Platzierungen zurückzuführen sind, ist in Abbildung 8.2 dargestellt. Sie zeigt die Abhängigkeit der Posengenerierung von der Güte der Konformationen und die Abhängigkeit der Vorhersageleistung auf dem ersten Rang von der Güte der zuvor erzeugten Posen. Während mit den erfolgreich generierten Konformationen generell gute Posen erzeugt werden, ist die Bewertungsfunktion nicht immer in der Lage, diese auch auf dem ersten Rang zu präsentieren. Häufig sind dafür äußere Umstände verantwortlich, wie die folgende Diskussion repräsentativer Docking-Fehlschläge zeigt.

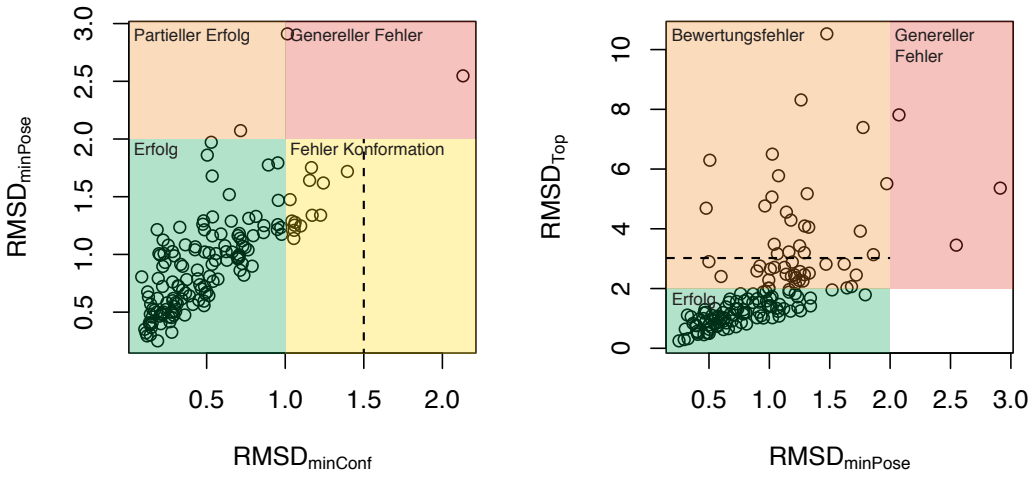


Abbildung 8.2: Abhängigkeit von $\text{RMSD}_{\text{min Conf}}$, $\text{RMSD}_{\text{min Pose}}$ zu RMSD_{Top} auf dem $\text{Astex}_{\text{ACS}}$ ($n = 146$). Gestrichelte Linien deuten partiellen Erfolg an.

8.1.3 Diskussion von Docking-Fehlschlägen

cRAISE reagiert sensitiv auf strukturelle Unstimmigkeiten wie unvollständige Bindetaschen, falsch orientierte Seitenketten, Kristallpackungseffekte, ungünstige Protein- und Ligandzustände und auf das Fehlen essentieller Wassermoleküle in der Bindetasche.

Unvollständige Bindetaschen beeinflussen die Platzierung und Bewertung von Liganden. Dies kann anhand der multimeren Strukturen `1tz8` und `1u1c` verdeutlicht werden (vgl. Abbildung 8.3). `1tz8` ist im Datensatz mit fünf Bindetaschen vertreten. Die Erste wird vollständig durch 122 Schweratome zweier Proteinketten beschrieben. Dagegen weisen die anderen vier Taschen nur rund die Hälfte der Atome auf. Sie stellen unvollständige, symmetrische Bestandteile der Tasche aus unterschiedlichen Kristalluntereinheiten dar. Ausschließlich die erste Bindetasche beschreibt die Kavität hinreichend genau, leitet plausible Interaktionsstellen ab und erzielt gute Top-Posen. Den Übrigen fehlen restriktive Atome, sodass dort primär partiell gedockte Posen generiert werden. Ähnliche Effekte sind auch für Bindetaschen aus Kristallstrukturen mit schlechter Elektronendichte zu beobachten. In den sechs Taschen von `1u1c` weist die Schleife 225–238 wenig oder keine Dichte nahe der gebundenen Liganden auf, sodass die Schleife unterschiedlich gut aufgelöst ist. Daher reichen viele Posen in den unaufgelösten Bereich hinein. Die vollständige Tasche erzielt dagegen ausschließlich gute Posen. Generell führen unvollständige Bindetaschen zu Top-Posen mit hoher RMSD-Varianz, da die Kavität unzureichend restriktiv ist, Interaktionsstellen nicht erzeugt werden und Bewertungsbeiträge unberücksichtigt bleiben. Posen nehmen ungehindert den freien Raum ein, werden nicht adäquat in der Tasche verankert und falsch eingeschätzt.

Alternative Seitenkettenorientierungen: cRAISE prozessiert für jede Aminosäure die erste in der Kristallstruktur vorgefundene Seitenkettenorientierung A. Ist diese unpassend, so nimmt sie eine Region in der Bindetasche ein, die es verbietet, Ligandatome dorthin zu platzieren oder sie richtet Interaktionsstellen ungünstig aus. Suboptimale Seitenkettenausrichtungen können dann anders orientierte Posen und/oder andere Bewertungsbeiträge verursachen. Für die meisten Strukturen ist Orientierung A bereits optimal. Ausschließlich bei `1vcj` ist Glu275 A ungünstig, da deren Carboxylatgruppe die Platzierung und Bewertung des Liganden beeinflusst (vgl. Abbildung 8.4). Die Orientierung weist bereits mit der aliphatischen Gruppe des Kristallliganden leichte Überlappungen auf. Dadurch können nahe der hydrophilen Seitenkette rezeptorseitig keine hydrophoben Interaktionsstellen erzeugt werden und über Deskriptortreffer keine aliphatischen Ligandgruppen in den Bereich verankert werden. Wird nahe zum Glutamat

dennoch eine Pose mit aliphatischem Rest platziert (zufällig, über alternative Deskriptortreffer), so wird der hydrophob-hydrophile Kontakt schlechter bewertet. Deshalb sind mit Seitenkette A die Top-32 von deplatzierten Posen dominiert. Mit Orientierung B werden dagegen ausschließlich gute Posen auf den vorderen Rängen erzeugt.

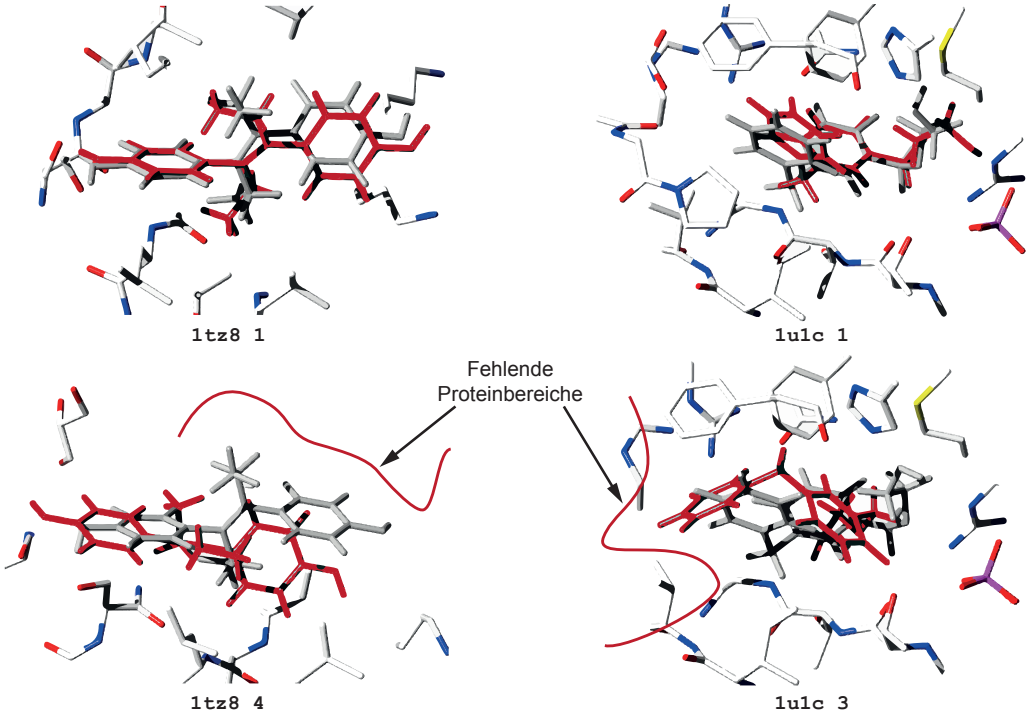


Abbildung 8.3: Einfluss unvollständiger Bindetaschen. Vollständige Taschen (oben) reproduzieren die Referenz (grau, Top-Posen rot), unvollständige nicht (unten).

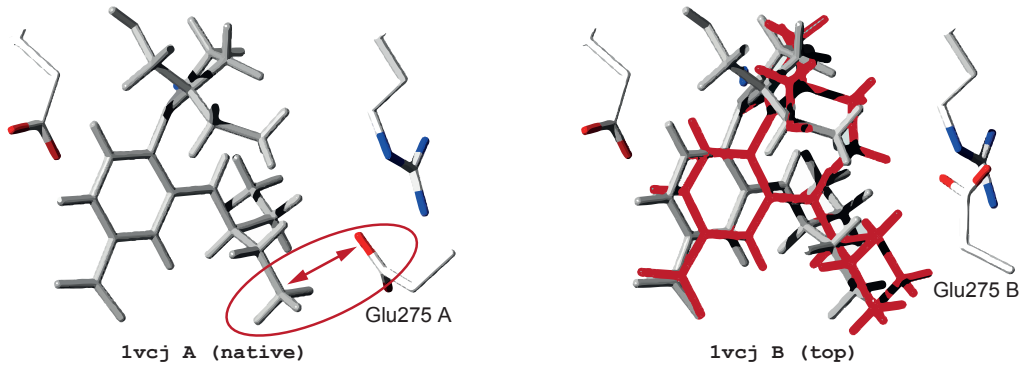


Abbildung 8.4: Einfluss alternativer Seitenketten bei `1vcj`. Orientierung Glu275 A reproduziert nicht den nativen Liganden (grau). Glu275 B führt zum Erfolg (rot).

Kristallpackungseffekte: Die Struktur von `1ia1` enthält einen offensichtlichen Fehler (vgl. Abbildung 8.5). Die Adeningruppe des NADPH-Kofaktors besteht fälschlicherweise nur aus Stickstoffen. Dadurch wird der Bestandteil des Rezeptors mit hydrophilen Interaktionsstellen überhäuft und in der Bindetasche weniger hydrophobe Bereiche identifiziert. Auch wenn die Struktur durch die Autoren des Datensatzes als unproblematisch eingestuft wurde, ist sie dennoch ein Beispiel für einen Docking-Fehlschlag aufgrund äußerer Umstände. Ähnliches ist auch bei Strukturen mit Kristallpackungseffekten zu beobachten. In solchen Fällen erstellt cRAISE Interaktionsstellen in Bereichen, an welchen eigentlich keine sein sollten. Dadurch entsteht ein Bias der Posen in Richtung der invaliden Interaktionsstellen und eine Überbewertung der Posen.

Alternative Ligandkonformationen: Kristallstrukturen können zuweilen uneindeutige oder alternative Ligandkonformationen enthalten. Wenn eine RMSD-basierte Qualitätsbewertung auf solche Referenzkoordinaten zurückgreift, kann dies zu einer Missinterpretation des Resultats und einem falsch detektierten Docking-Fehlschlag führen. Für drei Strukturen boten die Editoren des Datensatzes alternative Referenzligandkoordinaten an, unter anderem auch für beide Proteinbindetaschen von `1sg0`. cRAISE identifiziert für beide erfolgreich eine Pose mit eine RMSD von 1,0 Å bzw. 1,1 Å auf dem ersten Rang. Allerdings werden beträchtliche RMSD-Unterschiede innerhalb der oberen Ränge beobachtet. Zwei repräsentative Posen, die in den Top-32 gefunden werden, sind in Abbildung 8.6 dargestellt. Die erste ist nah zur initial gegebenen, die andere zur alternativ gegebenen Ligandkonformation. Beide Vorhersagen sind als richtig zu interpretieren, da sie mutmaßlich alternative Bindungsmodi des Liganden darstellen.

Suboptimale Zustände: Für eine erfolgreiche Reproduktion des Bindemodus ist es notwendig, dass Protein und Startligand bereits im optimalen Zustand zur Verfügung gestellt sind. Anderenfalls können typgleiche Interaktionsstellen von Ligand und Rezeptor während der Platzierung und Bewertung ungünstig aufeinandertreffen (vgl. Abbildung 8.7). Da im Zuge dieser Auswertung zunächst auf Protomerfreiheitsgrade verzichtet wurde, wurde deshalb für `1j3j` ein suboptimales Ergebnis erreicht. Die Top-Pose erreicht in Tasche 2 nur einen RMSD von 2,7 Å und die obersten Ränge weisen eine hohe RMSD-Varianz auf. Wird hingegen der optimale Eingabezustand verwendet, so können durchweg gute Posen in den Top-32 erzeugt und auf dem ersten Rang eine Pose mit einem RMSD von 0,6 Å präsentiert werden. Bei elf weiteren Strukturen sind die gegebenen Zustände auf Rezeptor- oder Ligandseite ebenfalls suboptimal. Dies führt im Gegensatz zum extremen Beispiel von `1j3j`, für das die betroffenen Interaktionsstellen wichtig zur Verankerung des Liganden in der aktiven Bindetasche sind, meist nur zu geringfügigen Posenabweichungen, aber dennoch zu fehlerhaften Bewertungen.

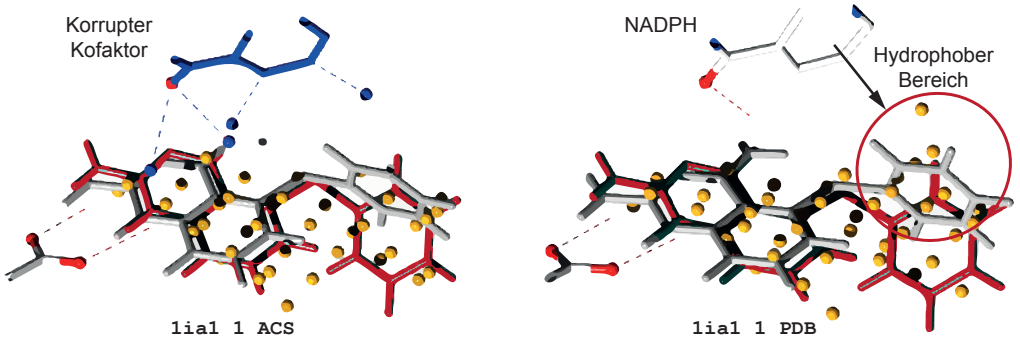


Abbildung 8.5: Einfluss korrupter Proteinstrukturen. NADPH besteht im Datensatz nur aus Stickstoffen (blau) und wird mit falschen Interaktionsstellen überhäuft. Gute Posen (rot) werden über alternative Deskriptortreffer produziert, aber falsch bewertet.

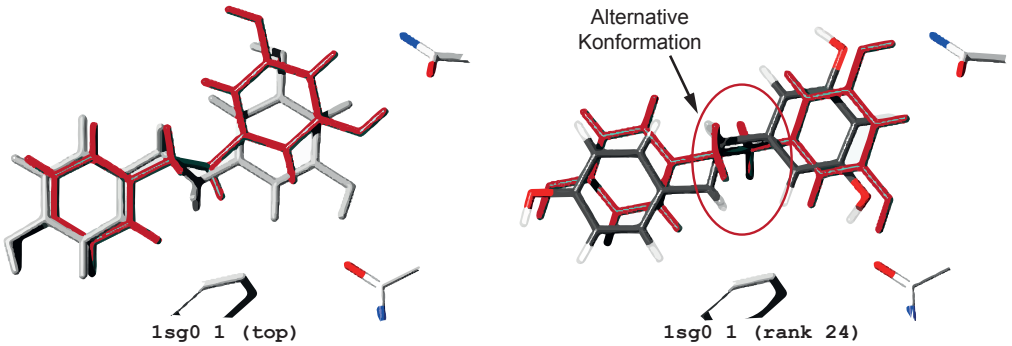


Abbildung 8.6: Einfluss alternativer Ligandkonformationen (grau und schwarz). In `1sg0` werden sie in den Top-32 identifiziert (rot), führen aber zu RMSD-Diskrepanzen.

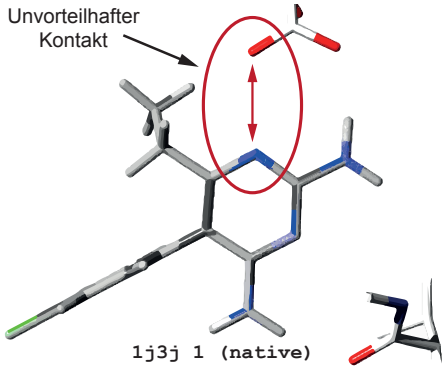


Abbildung 8.7: Einfluss suboptimaler Zustände. Der gegebene Ligandzustand von `1j3j` führt zum ungünstigen Kontakt zweier Akzeptoren, kann über die Carboxylatgruppe nicht gut verankert werden und resultiert im Docking-Fehlschlag.

Fehlen essentieller Wassermoleküle: Sind Wassermoleküle in der Proteinstruktur gegeben, dann fügt sie cRAISE zur internen Proteinrepräsentation hinzu. Sie können dann mit ihren Interaktionen eine Bindung von Pose zu Protein überbrücken oder Subtaschen der Bindetasche belegen, sodass die Platzierung in diesen Bereichen verhindert wird. Fehlen Wassermoleküle, so tendieren flexible, hydrophile Teile des Liganden dazu, mit wasserfreien Interaktionsstellen des Proteins zu interagieren. Abbildung 8.8 illustriert dies am Beispiel von `1hvy`. Die Top-32-Posen in den vier Bindetaschen zeigen ohne Wasser große RMSD-Varianzen und ein Docking-Erfolg auf dem ersten Rang ist nicht immer garantiert ($RMSD_{Top}$: 1,7 Å, 2,0 Å, 2,2 Å und 2,8 Å). Grund hierfür ist das Fehlen dreier Wassermoleküle, die in der PDB-Struktur eigentlich eine Subtasche befüllen. Im Rezeptor mitaufgenommen wird auf dem ersten Rang stets erfolgreich der native Bindemodus reproduziert ($RMSD_{Top}$: 1,4 Å, 1,4 Å, 1,6 Å und 1,7 Å). Die Vorhersagen von `1gm8`, `1gpk`, `1l2s`, `1l7f`, `1sq5`, `1sqn`, `1w2g`, `1x8x`, `1xm6`, `1xoq`, `1y6b`, `1ygc`, `1yqy` und `1yv3` sind ebenso durch fehlende Wassermoleküle beeinträchtigt. Im Datensatz sind sie *per se* nicht präsent.

Platzierungsfehler: Für `1yqy` ist es nicht möglich ein Pose unter 2 Å in den Top-32 zu identifizieren. Mit einem $RMSD_{min\,Pose}$ von 2,6 Å schlägt die Platzierungsroutine fehl, da bereits die Konformergenerierung nicht in der Lage ist, die bioaktive Konformation zu erzeugen. Die beste Konformation zeigt selbst bei einer optimalen Überlagerung auf die Referenz einen $RMSD_{min\,Conf}$ von mehr als 2 Å. Unter dieser Voraussetzung hat cRAISE keine Chance, eine bessere Pose zu produzieren. Für den Fehlschlag ist eine zentrale Bindung des Kristallliganden verantwortlich, die eine ungewöhnliche Torsion aufweist und dem Torsionshistogramm von Confect widerspricht. Selbst unter Hinzunahme zusätzlicher Toleranzwinkel weichen die erzeugten Konformationen noch deutlich ab. Dennoch etabliert die vorhergesagte Top-Pose zum Referenzligand vergleichbare Interaktionen (vgl. Abbildung 8.9). Sie ist lediglich in sich und um 180° verdreht.

Bewertungsfehler: Für `1y6b` in Abbildung 8.10 kann keine gute Top-Pose vorhergesagt werden. Zwar wird mit einem $RMSD_{min\,Pose}$ von 1,6 Å eine gute Pose generiert, die Bewertungsfunktion schätzt sie jedoch schlechter als andere ein. Dafür ursächlich sind ihre suboptimalen Interaktionsgeometrien, die dem Interaktionsmodell von cRAISE widersprechen. Nachdem die Pose sterisch und hinsichtlich der Interaktionsgeometrien optimiert und erneut mit der cRAISE-Funktion bewertet wird, kann sie auf dem ersten Rang identifiziert werden. Da die Optimierungsprozedur jedoch die Laufzeit des Dockings erhöht und für groß angelegte Screening-Läufe ungeeignet ist, wird sie standardmäßig nicht durchgeführt. Sie steht aber optional zur Verfügung.

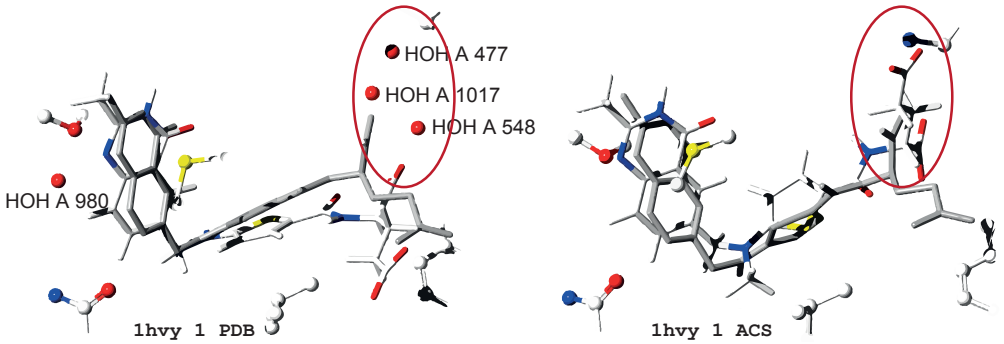


Abbildung 8.8: Einfluss von Wasser. Über HOH A 980 interagiert der Ligand in `1hvy` PDB mit dem Protein. Andere Moleküle füllen eine Subtasche, die Posen (Atomfarben) am Eindringen hindern. Ohne Wasser werden Abweichungen zur Referenz erhalten.

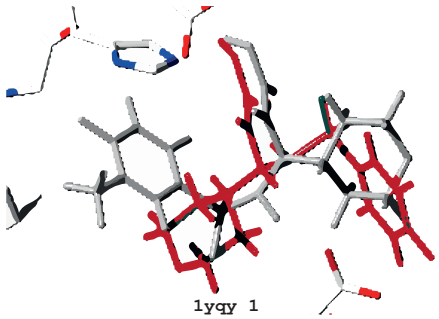


Abbildung 8.9: Platzierungsfehler. Die native Konformation von `1yqy` (grau) mit untypischer Torsion wird nicht vorhergesagt, dennoch aber eine plausible Top-Pose (rot).

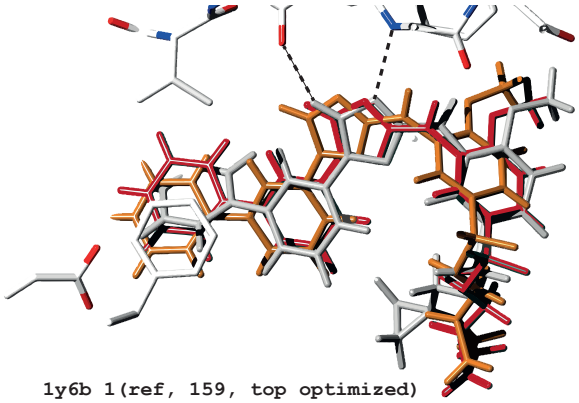


Abbildung 8.10: Bewertungsfehler. Die $\mathrm{RMSD}_{\mathrm{min\,Pose}}$-Pose wird mit 1,6 Å (orange) nur auf Rang 159 gefunden, da sie nicht die Interaktionen des Kristallliganden (grau) etabliert. Dieselbe, optimierte Pose (rot) wird auf Rang 1 identifiziert.

8.1.4 Erfolgsraten beim Redocking

Redocking-Läufe wurden auf den von den Organisatoren des ACS Docking-Symposiums 2011 gegebenen Rohdaten ($\text{Astex}_{\text{given}}$, 151 Bindetaschen) durchgeführt. Ebenso wurden die Strukturen – wie von den Organisatoren vorgeschlagen – präpariert und die Berechnungen zusätzlich auf zwei revidierten Datensätzen durchgeführt. Für den $\text{Astex}_{\text{ACS}}$ wurden hierfür offensichtliche Fehler in den Daten behoben (Korrektur falscher Seitenkettenorientierungen, korrupter Kofaktor und ungünstige Protein- und Ligandzustände). Zudem wurden die Strukturen, die zu schwere Fehler hatten (fehlende Bestandteile der aktiven Bindetasche) und daher nicht mehr vernünftig präpariert werden konnten, von der Auswertung ausgeschlossen. Der $\text{Astex}_{\text{h2o}}$ enthält außerdem explizit hinzugefügte, essentielle Wassermoleküle. Im Anhang A.1 sind alle Interventionen am ursprünglich gegebenen Datensatz vermerkt, um die zwei revidierten Datensätze zu erhalten. Tabelle 8.3 listet die Erfolgsraten, die mit den Daten erreicht wurden.

Tabelle 8.3: Redocking-Erfolgsraten (%) für den $\text{Astex}_{\text{h2o}}$, $\text{Astex}_{\text{ACS}}$ und $\text{Astex}_{\text{given}}$.

	RMSD [Å] ≤								
Rank	1.0	**2.0**	3.0	1.0	**2.0**	3.0	1.0	**2.0**	3.0
	$\text{Astex}_{\text{h2o}}$ $n = 85$			$\text{Astex}_{\text{ACS}}$ $n = 85$			$\text{Astex}_{\text{given}}$ $n = 85$		
1	33	**75**	86	27	**71**	84	26	**65**	79
5	41	**89**	94	37	**85**	95	34	**81**	93
20	52	**92**	99	47	**87**	97	45	**85**	95
32	54	**93**	99	49	**91**	98	47	**87**	97
all	64	**98**	100	61	**97**	100	58	**94**	99
	$\text{Astex}_{\text{h2o}}$ $n = 146$			$\text{Astex}_{\text{ACS}}$ $n = 146$			$\text{Astex}_{\text{given}}$ $n = 151$		
1	27	**69**	84	23	**65**	83	23	**60**	79
5	40	**85**	97	36	**80**	96	33	**77**	94
20	51	**91**	99	45	**84**	98	41	**83**	97
32	51	**91**	99	47	**88**	99	43	**86**	97
all	60	**99**	100	58	**98**	100	56	**96**	99

Im direkten Vergleich können mit revidierten Strukturen wesentlich bessere Vorhersagen erzielt werden. Das Ausräumen offensichtlicher Fehler und die Aufbereitung der Strukturen mit entscheidenden Wassermolekülen bringt eine Verbesserung von bis zu 10%. Zwischen den Vorhersagen in der besten Bindetasche ($n = 85$) und allen Bindetaschen ist allerdings, selbst mit den revidierten Daten, noch eine deutliche Diskrepanz festzustellen. Zwar kann die cRAISE-Bewertungsfunktion auch in den Bindetaschen multimerer Proteine gute Posen auf den vorderen 20 Rängen etablieren, zwischen den

Taschen existieren jedoch noch immer strukturelle Unterschiede, was vor allem auf die Bewertung der Top-Posen Einfluss hat. Die Rezeptoren definieren unterschiedliche wasserzugängliche Atome, die die Grundlage zur Ermittlung der Bewertungsbeiträge bilden.

8.2 Virtuelles Screening

8.2.1 Generelle Anreicherungsleistung auf dem DUD_{ACS}

Die Anreicherungsleistung von cRAISE wurde, wie in Abschnitt 7.2 beschrieben, auf dem DUD_{ACS}-Datensatz evaluiert. Tabelle 8.4 zeigt die durchschnittliche Anreicherungsleistung über alle 40 Zielstrukturen. Die durchschnittlichen $ROC_{1\%}$-, $ROC_{2\%}$-, $ROC_{5\%}$- und AUC-Werte liegen mit Werten von 9%, 14%, 22% und 65% deutlich über den Wahrscheinlichkeiten einer zufälligen Auswahl, die respektive nur 1%, 2%, 5% und 50% betragen würden. Damit ist cRAISE generell dazu in der Lage, Aktive sowohl früh als auch global anzureichern. Im Gegensatz zum Redocking wurde

Tabelle 8.4: DUD_{ACS} Anreicherung. Fehlerbereiche in 95%-Konfidenzintervallen.

	$ROC_{1\%}$	$ROC_{2\%}$	$ROC_{5\%}$	AUC	Q1-AUC
Mean	0.09 (±0.04)	0.14 (±0.06)	0.22 (±0.07)	0.65 (±0.05)	0.70 (±0.06)
SD	0.13	0.19	0.22	0.15	0.14
Median	0.05	0.09	0.142	0.61	0.66
Min	0.00	0.00	0.00	0.28	0.49
Max	0.52	0.70	0.82	0.95	0.95

bei den Anreicherungsexperimenten allerdings auf eine weitere Präparierung der Proteinstrukturen verzichtet, um die Vergleichbarkeit der Resultate mit etablierten Methoden zu erhalten. Für die ROC-Metriken ist deren Leistung auf dem Datensatz publiziert[345, 347, 348, 349, 350, 352] und in Tabelle 8.5 zusammengefasst. Im di-

Tabelle 8.5: Anreicherungsleistung (AUC) gängiger Methoden auf dem DUD_{ACS}.

	ICM	LeadIT/Hyde	Glide	Lead Finder	GOLD[a]	DOCK 6
Mean	0.72	0.72	0.74	0.74[b] (0.70)	0.70	0.60
SD	0.16	0.17	0,19	0.15	-	0.17
Median	0.69	0.73	0.76	0.76	0.69	0.56
Min	0.27	0.26	0.29	0.39	0.33	0.29
Max	0.96	0.95	0.99	0.96	0.95	0.96

[a] mit ChemPLP-Bewertung
[b] mit Filterprotokoll ähnlich zum Pharmakophoransatz

rekten Vergleich ordnet sich cRAISE demnach, ohne einen weiteren Eingriff durch den Anwender, grundlegend zwischen Gold bzw. LeadFinder und Dock ein. Jedoch ist auch beim Screening die adäquate Proteinpräparierung ein entscheidender Faktor für ein erfolgreiches Screening. Nach Schneider *et al.*[347] lassen sich die 40 Zielstrukturen nach Qualitätskriterien in vier Klassen einteilen: Q1-Strukturen besitzen keine oder nur geringe strukturelle Unstimmigkeiten, Q2-Strukturen weisen Unstimmigkeiten im gegebenen Referenzligand auf bzw. benötigen Wassermoleküle, um ein vernünftiges Resultat zu erzielen. Q3- und Q4-Strukturen weisen schlechte Elektronendichte oder Kristallpackungseffekte auf. Betrachtet man nach diesen Kriterien ausschließlich die 20 qualitativ besten Q1-Strukturen (vgl. Q1-AUC Tabelle 8.4), so kann die Anreicherung um 5% verbessert werden. Die individuellen Screening-Resultate aller Zielstrukturen sind im Anhang A.3 zu finden. Dort sind die frühen und globalen Anreicherungsleistungen detailliert aufgeführt und die Strukturen bezüglich ihrer Qualität klassifiziert.

8.2.2 Anreicherung unterschiedlicher Proteinfamilien

Betrachtet man die Anreicherungsleistung bzgl. unterschiedlicher Proteinfamilien, kann dies Aufschluss darüber geben, ob die Methode einzelne Familien besser als andere handhaben kann. Abbildung 8.11 zeigt die frühe und globale Anreicherung in Abhängigkeit zu den im DUD definierten Proteinfamilien der Metalloenzyme, der Kernrezeptoren, der Kinasen, der Serinproteasen, der Folatenzyme und zu der anderer Proteinstrukturen. Für die acht Kernrezeptor-Strukturen zeigt cRAISE mit einem durchschnittlichen AUC von 0,77 die beste Anreicherungsleistung. Im Vergleich zu den anderen Proteinfamilien besteht diese Klasse ausschließlich aus qualitativ guten Q1-Strukturen, was somit als Indikator für die gute Leistung gewertet werden kann. Die neun Kinasen im Datensatz stellen schwierige Zielproteine dar. Mit dieser Proteinklasse haben gängige strukturbasierte Screening-Methoden generell Probleme. Die Zielstrukturen zeichnen sich dadurch aus, dass die Ligandbindung durch eine entscheidende Transformation des Proteinrückgrats begleitet wird. Da cRAISE eine derartige Proteinflexibilität nicht modelliert, ist das Screening hier mit Schwierigkeiten bei der Platzierung und Bewertung der Liganden konfrontiert. Daher stellt die Proteinklasse mit einem durchschnittlichen AUC von 0,56 die Klasse mit der schlechtesten Anreicherungsleistung dar. Zudem haben lediglich zwei Strukturen Q1-Qualität. Auf den Metalloenzymen erreicht cRAISE eine durchschnittliche Anreicherung von 62%. Die Zielstrukturen sind ebenso wie die Serinproteasen (durchschnittlicher AUC von 0,61) und die Klasse der Folatenzyme (durchschnittlicher AUC von 0,65), die mit respektive nur vier, drei und zwei Strukturen im Datensatz vertreten sind, unterrepräsentiert. Auf dieser Basis lässt sich

ohne zu starke Schlüsse zu ziehen nicht sagen, ob cRAISE diese Klassen besser oder schlechter als andere handhaben kann. Mit den restlichen 14 Strukturen, die keiner Proteinfamilie zugeordnet sind, ist die Spanne möglicher Anreicherungsleistungen von der S-Adenosylhomocysteinhydrolase mit einem AUC von 0,95 (sahh, Q1) bis zur Acetylcholinesterase mit einem AUC von 0,47 (ache, Q2), in vollem Umfang ausgeschöpft. In dieser Klasse sind Strukturen jeglicher Güte vertreten. Mit einem durchschnittlichen AUC von 0,66 kann die Leistung auf dieser Klasse somit als die generell zu erwartende Anreicherungsleistung von cRAISE gewertet werden, wenn nicht weiter in den Screening-Prozess eingegriffen wird und keine weitere Proteinpräparierung durch den Anwender erfolgt.

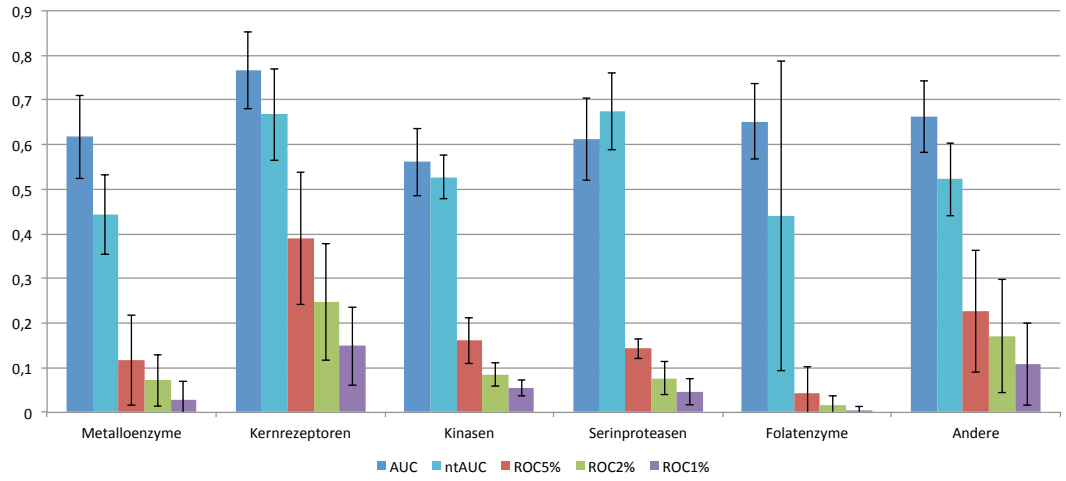


Abbildung 8.11: Anreicherungsleistung auf unterschiedlichen Proteinfamilien und unter Annahme der Nullhypothese. Fehlerbereiche entsprechen 95%-Konfidenzintervallen.

8.2.3 Zielstrukturspezifische Anreicherung

Abbildung 8.11 zeigt ebenfalls die durchschnittliche Anreicherungsleistung einzelner Proteinklassen bei Annahme der Nullhypothese. Unter der Voraussetzung, dass eine Zielstruktur zu einer perfekten Anreicherung mit einem tAUC von 1,0 und eine Anti-Zielstruktur zu einer zufälligen Auswahl von Aktiven, also einem ntAUC-Wert von 0,5 führt, ist der $\Delta\text{AUC} = \text{tAUC} - \text{ntAUC}$ nahe 0,5 und die Anreicherung somit höchst strukturspezifisch. ΔAUC-Werte nahe null werden dann erhalten, wenn Zielstruktur und Anti-Zielstruktur Aktive einer Bibliothek gleichermaßen anreichern. Dies kann als Indiz gewertet werden, dass die Screening-Methode die Strukturinformation missachtet und

Aktive alleinig aufgrund der Molekülinformation extrahiert. Mit den von den Symposiumorganisatoren gestellten Anti-Zielstrukturen ist die Anreicherung mit einem durchschnittlichen ntAUC-Wert von 0.55 wie erwartet nahe der einer zufälligen Auswahl. Für die Metalloproteine (durchschnittlicher ΔAUC 0,17), die Kernrezeptoren (durchschnittlicher ΔAUC 0,10), die Folatenzyme (durchschnittlicher ΔAUC 0,16) und die nicht weiter klassifizierten Proteine (durchschnittlicher ΔAUC 0,14) ist ein Austausch der Zielstruktur oft deutlich merkbar und führt zu einer schlechteren Anreicherung. Beispiele hierfür sind der Austausch von Phosphodiesterase V (pde5, tAUC 0,76) durch die Catechol-O-Methyltransferase (comt, ntAUC 0,51), der Austausch von RXR-alpha (rxr, tAUC 0,80) durch den Androgenrezeptor (ar, ntAUC 0,46) oder der Austausch der β-Glykogenphosphorylase (gpb, tAUC 0,85) durch die HIV reverse Transkriptase (hivrt, ntAUC 0,33). Gelegentlich führt ein Austausch der Zielstruktur allerdings dazu, dass die Anti-Zielstruktur die Aktiven der eigentlichen Zielstruktur ähnlich gut oder sogar besser anreichert. Das Screening des Estrogenrezeptoragonists (er_agonist) erreichte einen sehr guten AUC von 0,91. Der Austausch durch die Struktur des Mineralcorticoidrezeptors (mr) konnte die Aktiven des Estrogenrezeptoragonisten jedoch ebenfalls gut anreichern (ntAUC 0,86). Grund hierfür ist, dass beide Zielproteine Liganden mit einem sehr ähnlichen Steroidgrundgerüst binden. Tatsächlich ist der Schnitt beider aktiven Mengen auch nicht leer. Ein Ligand (ZINC03814383) ist in beiden aktiven Mengen vorhanden. Ein nichtleerer Schnitt der Ziel- und Anti-Ziel-Aktiven ist auch bei der Paarung von Glutocorticoidrezeptor (gr) und Progesteronrezpetor (pr) (6 gemeinsame Aktive) und der Tyrosinkinase SRC (src) mit der Fibrolastwachstumsfaktorrezeptorkinase (fgfr1) (84 gemeinsame Aktive) festzustellen. Der Datensatz ist somit zur Untersuchung einer zielstrukturspezifischen Anreicherung zum Teil suboptimal. Wird Thrombin dazu genutzt, Faktor-Xa-Aktive zu docken (ntAUC 0,70), kann sogar eine bessere Anreicherung als mit der eigentlichen Zielstruktur festgestellt werden (tAUC 0,64). Tatsächlich zeigen einige der Liganden bzgl. beider Zielstrukturen Kreuzreaktivität, sodass dieses Ergebnis durchaus plausibel ist.[362, 363] Zudem wurden, als für Faktor-Xa noch keine Kristallstrukturen vorhanden waren, Thrombinstrukturen ersatzweise für das strukturbasierte Design von Faktor-Xa-Inhibitoren genutzt.[364] Dies geschah ebenso für das Design von Kinaseinhibitoren, für die Zielstrukturen oft schwer zu kristallisieren sind.[365, 366] Mit diesem Hintergrundwissen ist es also nicht erstaunlich, dass Aktive manchmal auch mit einer Anti-Zielstruktur identifiziert werden können und für die Familie der Kinasen nur ein durchschnittlicher ΔAUC von 0,03 beobachtet werden kann.

8.3 Laufzeit und Selektivität

8.3.1 Generelles Laufzeitverhalten

Auf den einmalig präparierten $\text{ZINC}_{\text{CL1M}}$-, $\text{ZINC}_{\text{CL2M}}$- und $\text{ZINC}_{\text{CL3M}}$-Indizes wurde die Laufzeit gemäß Abschnitt 7.3 evaluiert. Für die leitstrukturähnlichen Bibliotheken wurden durchschnittlich 239 Konformationen pro Molekül und 108 Moleküldeskriptoren pro Konformer generiert und im Index hinterlegt. Zur Erstellung einer Indexpartition mit 2 500 Molekülen mussten dafür im Mittel einmalig 9,2 Stunden investiert werden. Die Konformergenerierung benötigte mit 49% den größten Anteil der aufgewandten Präparierungszeit, gefolgt von der Deskriptorberechnung mit 31% und der Deskriptorindexierung mit 20%. Für die Screening-Läufe wurden die S-Adenosylhomocysteinhydrolase (sahh), die β-Glykogenphosphorylase (gpb), das humane Hitzeschockprotein 90 (hsp90), der Estrogenrezeptoragonist (er_agonist), der Faktor Xa (fxa) und die Dehydrofolatreduktase (dhfr) aus dem DUD_{ACS} als Zielstrukturen gewählt. Sie repräsentieren untere und obere Schranken bzgl. der Anzahl der Anfragedeskriptoren (sahh: 10 677, gpb: 37 579, hsp90: 19 637, er_agonist 13 042, fxa: 31 510, dhfr: 36 943). Die Laufzeiten, die mit diesen Anfragen innerhalb der groß angelegten Screening-Experimente gemessen wurden, sind in Tabelle 8.6 zusammengefasst. Ist die maximale Rechenkapazität gewährleistet, so kann auf einem Rechencluster von 200 Kernen ein groß angelegtes Screening von mehreren Millionen Molekülen innerhalb weniger Stunden bis Tage realisiert werden. Für ein einzelnes, starres Docking sind durchschnittlich Zeiten im Millisekunden- und für ein flexibles Docking im Sekundenbereich aufzuwenden.

Tabelle 8.6: Zeitmessungen auf $\text{ZINC}_{\text{CL1M/2M/3M}}$ Bibliotheken. Durchschnittliche Laufzeit pro Konformation t_s, Molekül t_f, parallele Laufzeit t_p und Selektivität σ.

	t_s [s]	t_f [s]	t_p [h] (1M)	t_p [h] (2M)	t_p [h] (3M)	σ
sahh	0.01	2.75	3.65	7.13	10.40	0.13
gpb	0.04	9.73	12.58	26.22	38.78	0.87
hsp90	0.04	8.89	13.05	25.75	37.03	1.38
er_agonist	0.03	8.23	10.83	23.22	34.30	1.54
fxa	0.09	21.24	30.37	59.00	89.37	5.40
dhfr	0.10	24.81	34.12	63.32	103.37	6.10

Abbildung 8.12 stellt die Abhängigkeit der Laufzeit zum Umfang der gescreenten Molekülbibliothek gegenüber. Generell sind mit steigendem Bibliotheksumfang entsprechend erhöhte Laufzeiten zu erwarten. Allerdings können für unterschiedliche Zielproteine signifikante Laufzeitunterschiede auftreten.

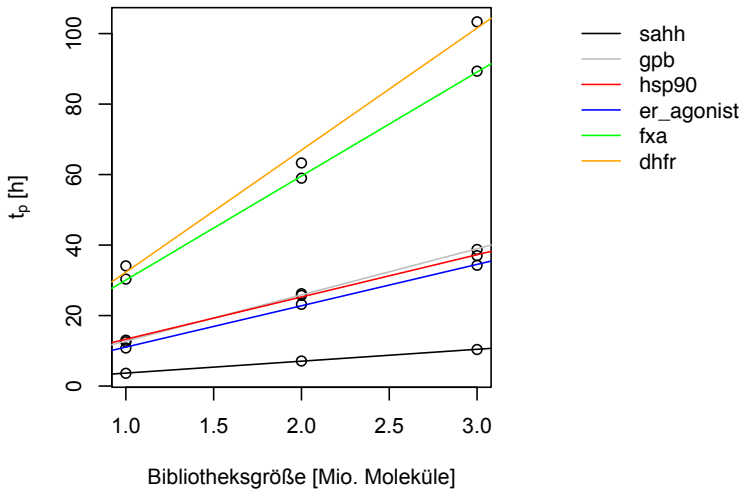


Abbildung 8.12: Parallele Laufzeit t_p gegen Größe der $\text{ZINC}_{\text{CL1M/2M/3M}}$ Bibliotheken.

8.3.2 Zusammenhang zur Selektivität einer Anfrage

Wie Abbildung 8.13 zeigt, werden im Mittel – auf Basis einer Zielstruktur – sehr ähnliche Laufzeiten für Indexpartitionen erhalten. Dies lässt sich dadurch erklären, dass die Moleküle umfangreicher Bibliotheken bzgl. ihrer Eigenschaften wie Größe, Flexibilität und Hydrophilität, normalverteilt sind. Unter dieser Voraussetzung ist die Anzahl extrahierter Molekülдeskriptoren und somit die Selektivität auf den Indexpartitionen annähernd gleich. Zwischen Zielproteinen kann die Selektivität der Anfragen jedoch variieren. Für die geringe Selektivität eines Zielproteins kann eine erhöhte Anzahl von Anfragedeskriptoren ursächlich sein. Dies ist ein notwendiger, aber nicht hinreichender Grund. Die Konstellation der Anfragedeskriptoren bestimmt viel mehr die Selektivität. Vergleicht man die Verteilung der Anfragedeskriptortypen (vgl. Abbildung 8.14), stellt man für Proteine mit hohem σ tendenziell eine erhöhte Anzahl der Typen 5 und 8 fest. Beide Deskriptortypen zeichnen sich dadurch aus, dass sie anhand von nur einer gerichteten Donor- bzw. Akzeptorinteraktionsstelle und zwei ungerichteten, hydrophoben Interaktionsstellen gebildet werden (vgl. Tabelle 6.1). Sie sind im Vergleich zu allen anderen Deskriptortypen weniger restriktiv, da weniger Freiheitsgrade, nämlich Interaktionsrichtungen, während des Abgleichs herangezogen werden. Die Eigenschaft viele hydrophobe Deskriptoren abzuleiten ist jedoch auch kein Alleinstellungsmerkmal wenig selektiver Zielproteine. Dies zeigt eine Gegenüberstellung von gpb und fxa. Beide Zielproteine besitzen vergleichbar viele hydrophobe Anfragedeskriptoren, unterscheiden sich aber mit einem σ von 0,87 und 5,40 deutlich bezüglich der Anfrageselektivität. Das selektivere Protein besitzt sogar rund 6 000 Deskriptoren mehr. Die aktiven Bindetaschen

von fxa und gpb unterscheiden sich aber in ihrer Form. fxa liegt offen zugänglich an der Oberfläche des Proteins, gpb vergraben im Proteininneren. Dadurch bleibt ein Großteil der Bulk-Strahlen der fxa-Anfragen unbeschränkt, sodass der Deskriptorabgleich eine geringere Restriktion ausübt. Dagegen beschränkt die gpb-Bindetasche die Strahlen maßgeblich, sodass der Abgleich weniger passende Indexdeskriptoren extrahiert.

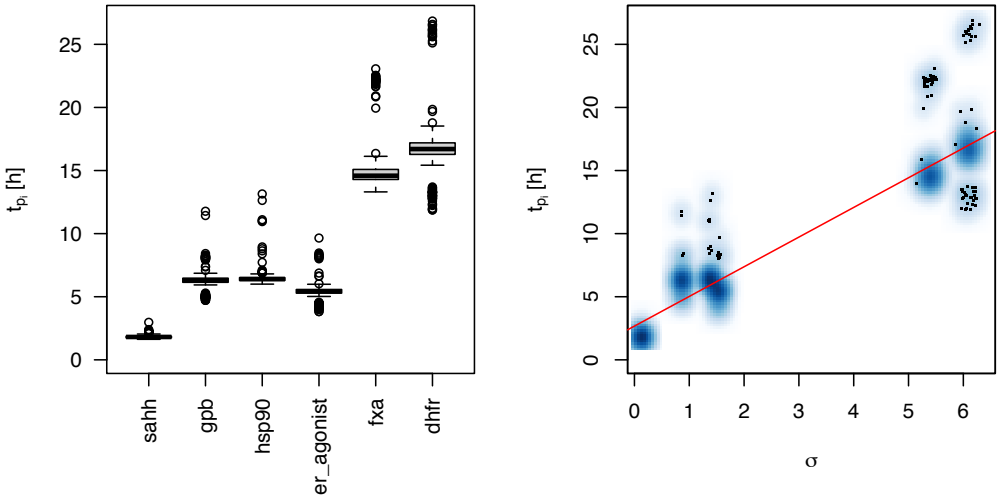


Abbildung 8.13: Laufzeiten t_{p_i} gescreenter Indexpartitionen (links). Boxen umschließen 50% der Datenpunkte um den Median, Whisker entsprechen dem 1,5-fachen des Interquartilabstands. Die Laufzeit ist primär von der Selektivität σ abhängig (rechts). Alle anderen im Diagramm sichtbaren Gruppierungen der Berechnungszeiten lassen sich eindeutig auf drei verschiedene Rechnerarchitekturen des HPC-Clusters zurückführen.

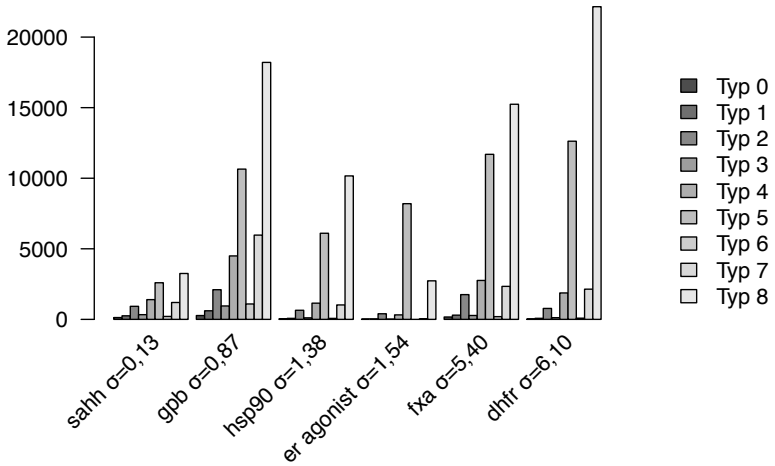


Abbildung 8.14: Verteilung der Anfragedeskriptortypen bei unterschiedlichem σ.

8.3.3 Laufzeitbestimmende Schritte

Der Screening-Prozess lässt sich im Wesentlichen in drei Phasen gliedern – die Rezeptorpräparierung, der Deskriptorabgleich und die Trefferprozessierung. Die Rezeptorpräparierung umfasst die Proteininitialisierung und -präparierung, die Formulierung einer aktiven Bindetasche, die Generierung von Proteininteraktionsstellen, die Anfragedeskriptorberechnung und die Vorberechnung der frühen Bewertungsinformation (vgl. Abschnitt 6.5). Im Vergleich zu den beiden anderen, nimmt die initiale Phase des Screening-Prozesses einen verschwindend geringen Anteil zur Gesamtlaufzeit ein. Für die betrachteten Proteine betrug sie lediglich 7 bis 14 Sekunden. Wie Abbildung 8.15 zeigt dominiert innerhalb eines Screening-Laufs der indexbasierte Deskriptorabgleich die Laufzeit mit 74% (dhfr) bis 99% (sahh). Die Trefferprozessierung nimmt dementsprechend die verbleibenden 26% bis 1% der Gesamtlaufzeit ein. Ihr Aufwand ist für unterschiedliche Zielstrukturen variabel und maßgeblich durch die Selektivität bestimmt. Sie entscheidet wie viele Treffer tatsächlich weiter prozessiert und durch die Bewertungshierarchie gereicht werden müssen. Da die Selektivität einer Anfrage vor einem Screening nicht

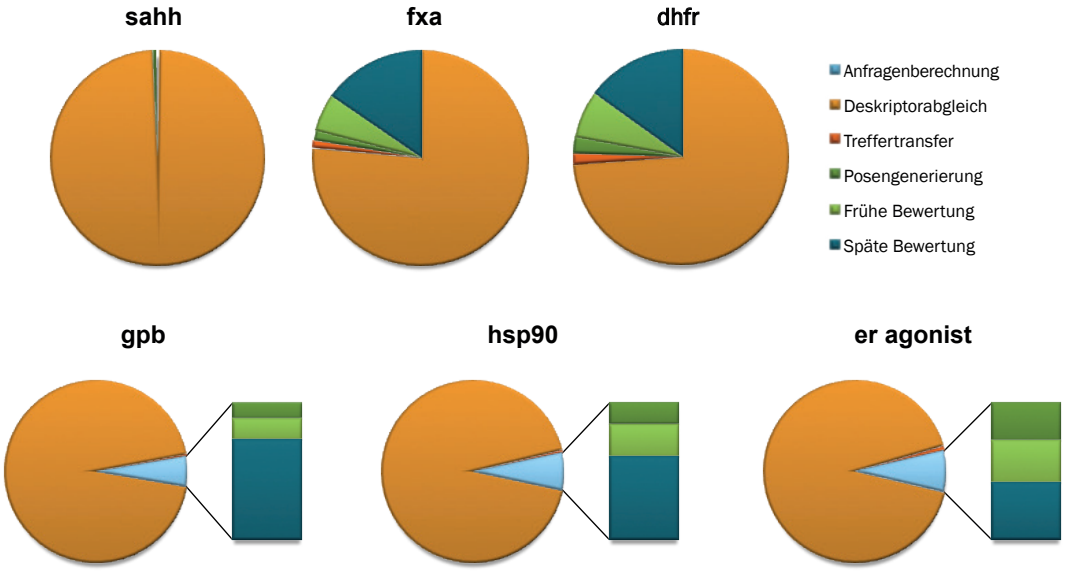


Abbildung 8.15: Anteile der Rezeptorpräparierung, des Deskriptorabgleichs (inkl. Treffertransfer) und der Trefferprozessierung zur Gesamtlaufzeit. Die Trefferprozessierung gliedert sich weiter in Posengenerierung, frühe und späte Bewertungsphase.

absehbar ist, wurde in cRAISE die Bewertungshierarchie gemäß Abschnitt 6.11 eingeführt. Um auch bei wenig restriktiven Proteinen zu gewährleisten, dass alle Treffer prozessiert werden können, forciert sie früh eine Posenreduktion. Da der Umfang der

Posen je nach Anwendungsfall mehr oder weniger reduziert wird, ist der Aufwand einzelner Komponenten der Bewertungshierarchie verschieden. Dies zeigt ein Vergleich der gpb, des hsp90 und des er_agonist, bei denen die Trefferprozessierung insgesamt einen ähnlichen Anteil zur Gesamtlaufzeit beansprucht, der Aufwand zur Posengenerierung, zur frühen und späten Bewertung jedoch variiert. Tabelle 8.7 listet detailliert die mittlere Laufzeit der ausgeführten Berechnungen auf. Je nach Anwendung unterscheidet sich die Häufigkeit ihrer Ausführung. Durch intelligentes Traversieren der Treffer reduziert cRAISE die Anzahl der recht teuren Molekül- und Konformerinitialisierungen auf ein Minimum. Im schlimmsten Fall wird die interne Repräsentation jedes Moleküls und jedes Konformers einmalig aufgebaut. Eine Transformation wird für jeden Deskriptortreffer berechnet. Inklusive der Reinitialisierung von Moleküldeskriptorkoordinaten ist der Aufwand hierfür zwei Größenordnungen geringer. Dies ist auch für den Überlapptest der Fall, der zuerst auf allen Posen ausgeführt wird. Der Test auf Enthaltensein im aktiven Volumen ist verhältnismäßig teuer wird aber bereits auf einer reduzierten Anzahl von Posen ausgeführt. Die aufwändige Wasserstoffbrückennetzwerkoptimierung und die atombasierte Bewertung werden auf einer eingeschränkten Posenmenge ausgeführt. Dies sind höchstens die tausend besten, früh bewerteten Posen pro Molekül.

Tabelle 8.7: Mittlere Laufzeit t einzelner Komponenten der Bewertungshierarchie.

Komponente	Berechnungsbasis	t [ms]
Posengenerierung		
Molekülinitialisierung	à Molekül	0.903
Konformerinitialisierung	à Konformer	0.044
Reinitialisierung der Dreieckskoordinaten	à Pose	0.003
Transformation	à Pose	0.008
Frühe Bewertung		
Gitterbasierter Überlapptest	à Pose	0.003
Test auf Enthaltensein in der konvexen Hülle	à Pose	0.155
Gitterbasierte Bewertung	à Pose	0.045
Späte Bewertung		
Clustering	à Pose	0.097
Wasserstoffbrückennetzwerkoptimierung	à Pose	0.844
Atombasierte Bewertung	à Pose	0.340

8.4 Geleitete Vorhersagen

8.4.1 Pharmakophorgeleitete Bindungmodusvorhersage

Mit dem $\text{Astex}_{\text{ACS}}$ und den gemäß Abschnitt 7.4.3 automatisch abgeleiteten Pharmakophormodellen wurden geleitete Bindungsmodusvorhersagen durchgeführt. Hierbei übte jeweils ein Pharmakophormodell implizit Druck auf die Posenbewertung aus, indem Posen, die den gegebenen Merkmalen widersprachen, durch das Modell verworfen wurden. Dadurch etablierten sich pharmakophorerfüllende Posen früher in der Bewertungsliste, falls zuvor bereits nicht die gewöhnliche Vorhersage eine pharmakophorerfüllende Pose auf dem ersten Rang präsentierte. Die Bewertung der Posen selbst blieb von der Pharmakophorinformation unberührt. Abbildung 8.16 stellt einige Modelle und die damit identifizierten Top-Posen dar. Jede der Posen wird auf einem niederen Rang identifiziert, wenn die Platzierung nicht durch die gezeigten Inklusionen geleitet wird. Die automatisch abgeleiteten Modelle können manchmal jedoch extrem gespannte Ligandkonformationen erfordern, um alle Merkmale optimal zu erfüllen. Dies ist auf die strukturbasierte Pharmakophormodellierung zurückzuführen, die Merkmale nicht auf Kristallligand- sondern relativ zu Proteinatomkoordinaten positioniert. Um diesem Problem zu begegnen, ermöglicht cRAISE die Verarbeitung relaxierter Modelle. Sie besitzen die Eigenschaft, dass die Anzahl zu erfüllender Merkmale N_e geringer als die Gesamtzahl aller definierten Merkmale ist. Ein reduziertes N_e schwächt den Druck während des Pharmakophorabgleichs ab, indem lediglich die Erfüllung einer N_e-dimensionierten Submenge von Merkmalen gefordert wird. So können Posen nahe des nativen Bindemodus erhalten werden, selbst wenn einige Merkmale geometrisch validen Ligandkonformationen widersprechen. Abbildung 8.17 zeigt den Effekt, der durch die Relaxierung erreicht wird und stellt die Docking-Erfolgsraten mit linear steigendem N_e dar. Generell leiten Modelle die Vorhersage verstärkt, wenn bis zu 80% der Inklusionen erfüllt werden müssen (blaue Balken). Gleichzeitig werden partiell gedockte Posen und Docking-Fehlschläge reduziert (grüne Balken). Allerdings sind die 80%-Modelle für einige Strukturen bereits zu strikt, da zugleich nicht alle Merkmale erfüllt werden können (rote Balken). Der Trend, mit steigendem N_e weiter die Vorhersagen zu leiten, kehrt sich deshalb um. Diese Fehlschläge können abgefangen werden, indem N_e strukturspezifisch reduziert wird. Dies ist als Teil der Pharmakophormodellierung zu betrachten, bevor ein geleitetes Docking durchgeführt wird. Um zu demonstrieren welches Resultat mit einem adäquat gewählten N_e erhalten werden kann, wurde für jede Struktur des $\text{Astex}_{\text{ACS}}$ ein gutes Model gewählt, d. h. ein RMSD_{Top}-minimierendes N_e^g (N_e^g-Balken). Typischerweise erfordert dies die Reduzierung des Werts um 5–25% vom Maximalwert. Dadurch werden Posen erhalten, die gespannte Modelle zumindest zu 75%–95% erfüllen.

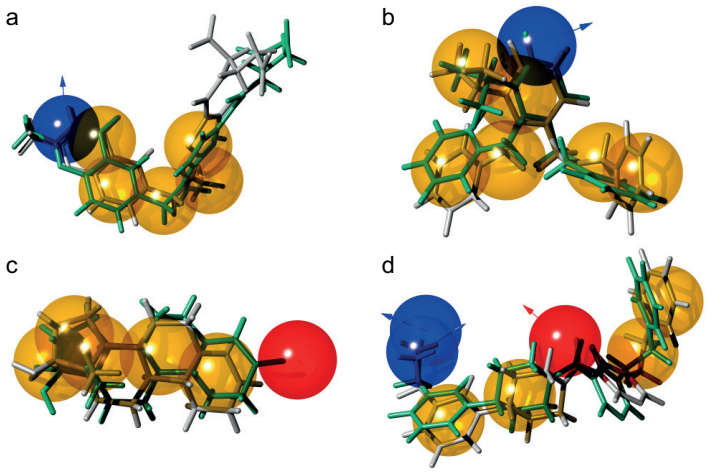


Abbildung 8.16: Pharmakophorgeleitete Vorhersagen (grün) werden nah zum nativen Bindemodus (grau) in den vorderen Rängen identifiziert. Donorinklusionen (blau), Akzeptorinklusionen (rot), hydrophobe Inklusionen (gold). **a**: 1hvy_3, **b**: 1jla_1, **c**: 1sqn_1 und **d**: 2bm2_2 werden mit den Modellen jeweils auf Rang 1 geführt (nicht geleitet: Rang 147, Rang 5, Rang 5 und Rang 143).

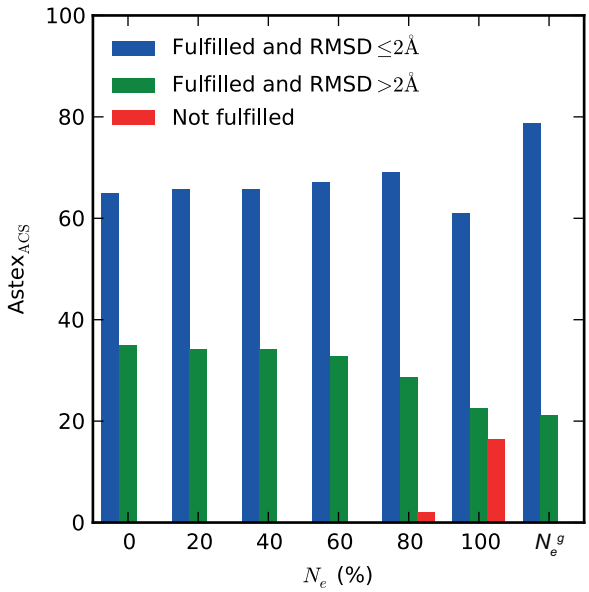


Abbildung 8.17: Erfolgsraten auf dem Astex$_{\text{ACS}}$ (n=146) bei steigender Anzahl geforderter Inklusionen N_e (blau). Top-Posen weichen weniger vom nativen Bindemodus ab je mehr Merkmale erfüllt werden (grün). Zu strikte Modelle führen zu Docking-Fehlschlägen (rot). Die adäquate Relaxierung von N_e kann dies verhindern (N_e^g).

Für die N_e^g-Modelle listet Tabelle 8.8 die Erfolgsraten geleiteter Vorhersagen detailliert auf. Im Vergleich zu den nicht geleiteten Vorhersagen (vgl. Tabelle 8.3) kann die Rate für den Docking-Erfolg insgesamt um 15% gesteigert werden, wenn zuvor für jede Bindetasche ein gutes Pharmakophormodell erstellt wird. Posen mit einem RMSD von mehr als 2 Å (grüne Balken) entsprechen vor allem partiell gedockten Posen mit einem RMSD von weniger als 3 Å. Sie sind das Resultat der relaxierten Pharmakophormodelle, die Posen nur partiell abdecken. Im Fall von `1hvy` (vgl. Abbildung 8.16) ermöglicht das relaxierte Modell beispielsweise einem flexiblen Teil des Liganden unbeschränkt Regionen der Bindetasche zu explorieren, während der Rest durch die Merkmale fest in der Bindetasche verankert wird. Posen, die völlig vom nativen Bindemodus abweichen, können mit den Modellen fast vollständig verhindert werden, sodass Docking-Fehlschläge auf ein Minimum reduziert werden. Zudem werden die Vorhersagen durch die Verwendung der Modelle nicht nur verbessert, sondern auch einige der Diskrepanzen ausgeräumt, die zuvor zwischen den Vorhersagen in unterschiedlichen Bindetaschen multimerer Proteine entstehen konnten. Um generell den Unterschied zwischen geleiteter und nicht geleiteter Vorhersage zu bestimmen, wurden RMSD-Differenzen zwischen den Docking-Resultaten ohne und mit N_e^g-Model, gebildet und Paardifferenztests durchgeführt. Tabelle 8.8 ist durch die Wahrscheinlichkeiten p für die Signifikanz der Vergleiche und durch die Stärke des beobachteten Effekts mittels Cohen's d ergänzt. Beide Werte sind für verschiedene Ränge dargestellt und zeigen, dass mit den Modellen ein signifikanter Unterschied und ein mittlerer Effekt auf den vorderen Rängen erreicht wird. Mit zunehmendem Rang werden die Unterschiede jedoch geringer. Dies ist nicht erstaunlich, da die Wahrscheinlichkeit dieselbe Pose ohne die Nutzung eines Pharmakophormodells auf einem unteren Rang zu finden ebenso steigt. Die Beobachtungen stützen die Behauptung, dass pharmakophorgeleitete Vorhersagen die Posengenerierung steuern und die Posenbewertung angemessen beeinflussen.

Tabelle 8.8: Erfolgsraten (%) beim mit N_e^g-Modellen geleiteten Redocking auf dem $\text{Astex}_{\text{ACS}}$ n=85 (n=146). Paardifferenzentest p und Cohen's d für n=146 Bindetaschen.

Rank	≤ 1.0 Å	≤ 2.0 Å	≤ 3.0 Å	p	d
1	35 (32)	85 (80)	97 (95)	<0.001	0.483
5	41 (40)	91 (87)	97 (95)	0.002	0.194
20	48 (45)	93 (91)	99 (99)	0.151	0.087
32	51 (47)	93 (93)	100 (99)	0.227	0.068
all	52 (49)	95 (95)	100 (100)	0.562	0.031

8.4.2 Pharmakophorgeleitete Anreicherung bioaktiver Moleküle

Für die pharmakophorgeleiteten Screening-Experimente wurden gemäß Abschnitt 7.4.3 automatisch Modelle von den gegebenen Protein-Ligand-Komplexen des DUD_{ACS} abgeleitet. Während des Screenings nutzte sie cRAISE, um zu verifizieren, ob die angereicherten Moleküle eine Mindestzahl gemeinsamer Merkmale besitzen. Ein striktes Model erwartete, dass nahezu alle Inklusionen erfüllt werden, um Aktive mit dem dadurch reflektierten, spezifischen Bindungsmuster zu identifizieren. Ein relaxiertes Modell erlaubte es, Toleranz beim Pharmakophorabgleich einzuführen und verschiedene Merkmalskombinationen des Modells auszuwerten. Der Grad der Relaxierung beeinflusste hierbei die Anreicherung der Aktiven. Abbildung 8.18 demonstriert dies am Beispiel der Neuraminidase (na), dem humanen Hitzeschockprotein 90 (hsp90), dem Estrogenrezeptorantagonisten (er_antagonist), der Hydroxymethylglutaryl-CoA-Reduktase (hmga), der Dihydrofolatreduktase (dhfr) und der Adenosindeaminase (ada). Je mehr Inklusionen erfüllt werden mussten, desto stärker wurde die globale Anreicherung in Form des AUCs (grüne Kurven). Im Gegensatz dazu reicherten strikte Modelle ausschließlich Submengen von Aktiven an, diese jedoch oft früher (blaue Kurven). Generell veranschaulichen die Experimente, dass mittels der Verwendung von Pharmakophormodellen der Screening-Prozess extern kontrolliert werden kann. In dieser Arbeit war es jedoch nicht das Ziel, ein Werkzeug zur Pharmakophormodellierung zu entwickeln, sondern die Möglichkeit zu etablieren, ein Screening mit bereits adäquat präparierten Modellen zu steuern. Um zu zeigen was cRAISE in diesem Fall leisten kann, wurde für jede Zielstruktur des DUD_{ACS} ein gutes Modell gewählt, d. h. es wurde ein N_e^g bestimmt, das die globale Anreicherung optimierte (rote Kurven in Abbildung 8.18). Innerhalb der durchgeführten Experimente rangierte dieser Wert zwischen 10% und 95% der im Modell insgesamt definierten Merkmale. Abbildung 8.19 zeigt alle Anreicherungswerte der Screening-Läufe auf dem Datensatz, die mit den N_e^g-Modellen erhalten wurden (blaue Balken). Im Vergleich zu den nicht geleiteten Vorhersagen (graue Balken) konnten die geleiteten Vorhersagen zumeist die frühe und die globale Anreicherung signifikant verbessern. Im Fall von hmga bewirkte die Nutzung des Pharmakophormodells sogar eine Verbesserung des AUCs um 36%, sodass mithilfe des Pharmakophormodells aus dem Screening ein stark gerichteter Prozess wurde, während das nicht geleitete Screening die Aktiven nur zufällig anreichern konnte. Wenige Modelle führten zu verschlechterten AUC-Werten. Hier war zu beobachten, dass die automatische Modellgenerierung mit problematischen Komplexen konfrontiert war, die zu zweifelhaften Merkmalsdefinitionen führte. Die N_e^g-Modelle lassen sich weiter verbessern indem man die N_e^g-Kombination bestimmt, die in der Lage ist, alle Aktiven zu bemerken. Die Ergebnisse spiegeln daher

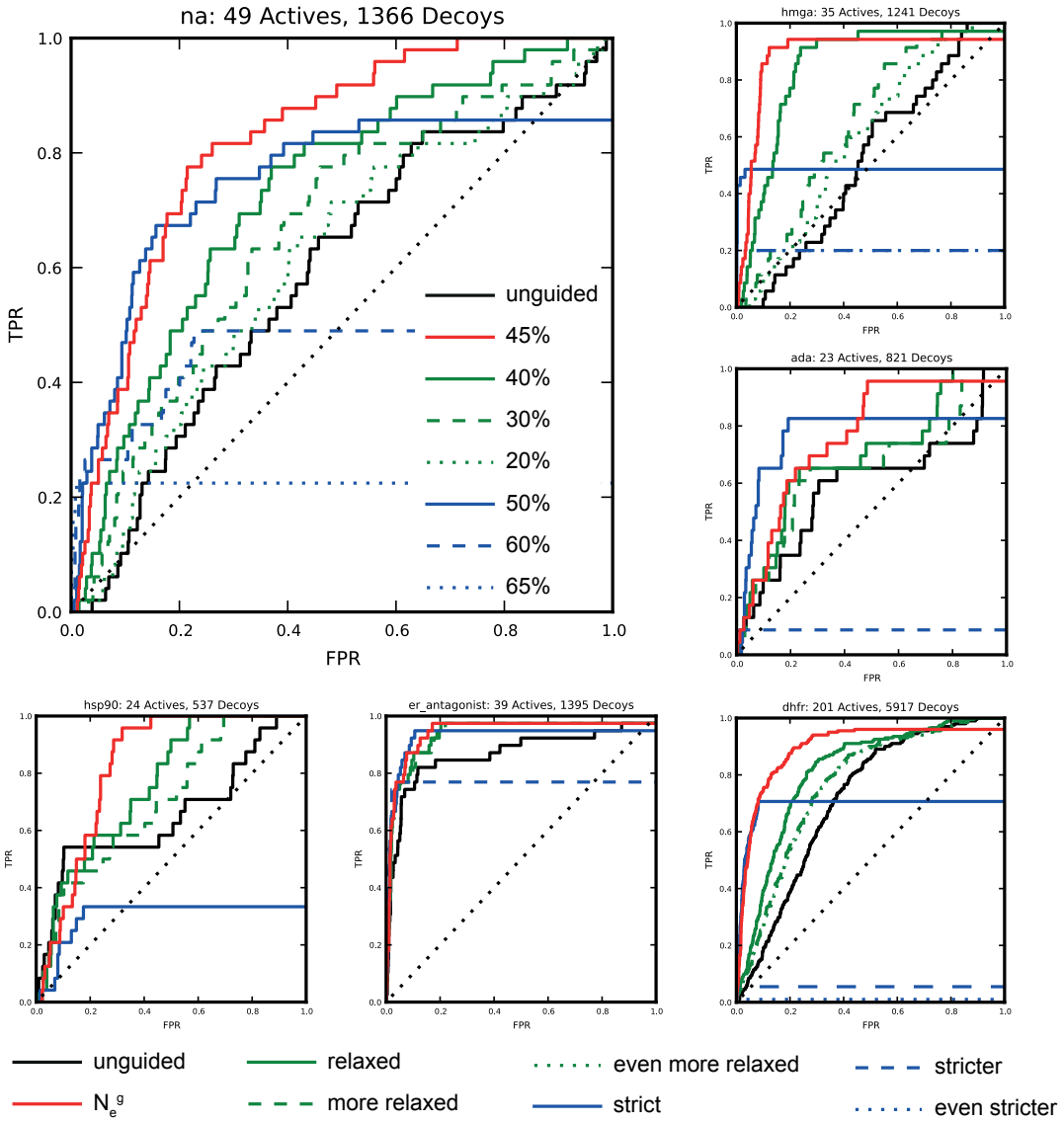


Abbildung 8.18: Einfluss strikter und relaxierter Modelle auf die Anreicherung. Mit steigendem N_e wird im Vergleich zu gewöhnlichen Vorhersagen (schwarze Kurven) die globale Anreicherung gesteigert (grüne Kurven). Strikte Modelle reichern Submengen der Aktiven früh an (blaue Kurven). Rote Kurven wurden mit guten N_e^g-Modellen erhalten.

realistische Kennzahlen wider, welches Resultat zumindest von einem pharmakophorgeleiteten Screening erwartet werden kann. Tabelle 8.9 fasst die geleitete Anreicherungsleistung auf dem gesamten Datensatz zusammen und zeigt die statistische Information bzgl. der Anreicherungsmetriken.

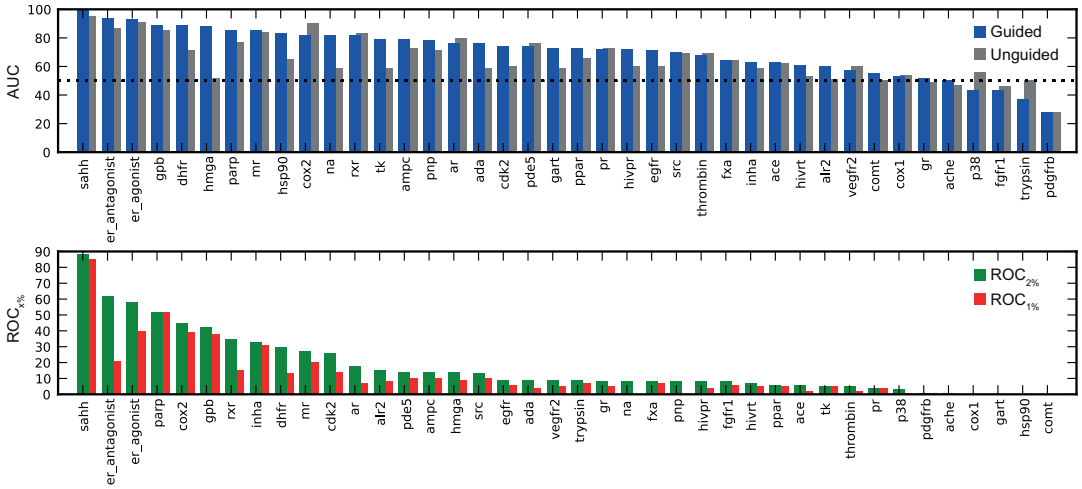


Abbildung 8.19: $\mathrm{ROC}_{1\,\%}$, $\mathrm{ROC}_{2\,\%}$ und AUC mit guten Modellen auf dem $\mathrm{DUD}_{\mathrm{ACS}}$.

Tabelle 8.9: Anreicherungsleistung auf dem $\mathrm{DUD}_{\mathrm{ACS}}$ Datensatz unter Verwendung guter Pharmakophormodelle. Fehlerbereiche entsprechen 95%-Konfidenzintervallen.

	$\mathrm{ROC}_{1\%}$	$\mathrm{ROC}_{2\%}$	AUC
Mean	0.123 ($\pm$0.055)	0.177 ($\pm$0.064)	0.704 ($\pm$0.052)
SD	0.173	0.201	0.164
Median	0.060	0.090	0.730
Min	0.000	0.000	0.280
Max	0.850	0.880	0.990

8.4.3 Laufzeit und Selektivität

Unter Nutzung der $\mathrm{ZINC}_{\mathrm{CL1M}}$-, $\mathrm{ZINC}_{\mathrm{CL2M}}$- und $\mathrm{ZINC}_{\mathrm{CL3M}}$-Bibliotheken wurde in Analogie zu den in Abschnitt 8.3 durchgeführten Experimenten der Einfluss von Pharmakophormodellen auf die Laufzeit von cRAISE evaluiert. Für die dort aufgeführten Zielstrukturen wurden N_e^g-Modelle gewählt, mit denen bereits gute Anreicherungswerte auf dem $\mathrm{DUD}_{\mathrm{ACS}}$ erzielt werden konnten. Tabelle 8.10 umfasst die Zielstrukturdaten, die für die Experimente genutzt wurden und vergleicht die Anzahl der Anfragedeskriptoren beim Screening mit und ohne Pharmakophormodell. Die Zahlen verdeutlichen, dass hydrophile Inklusionsmerkmale vom Typ Donor, Akzeptor oder Hydrophil deutlich die Anzahl der Anfragedeskriptoren der anvisierten Bindungstasche und somit entscheidend den Suchraum reduzieren. Die Modelle beinhalteten durchschnittlich vier dieser Merkmale, die den Umfang der Anfragen (durchschnittlich 30 000) auf ein Viertel reduzierten. Die Laufzeiten, die innerhalb der groß angelegten, pharmakophorgeleiteten Screening-

Tabelle 8.10: Anzahl der Anfragedeskriptoren mit und ohne Pharmakophor.

	mit Pharmakophor	ohne Pharmakophor
sahh	5678	10677
gpb	15433	37579
hsp90	2934	19637
fxa	3996	31510
er_agonist	7756	13042
dhfr	7491	36943

Experimente beobachtet wurden, sind in Tabelle 8.11 zusammengefasst. Zudem ist die Selektivität σ für die pharmakophorbeschränkten Anfragen gelistet. Im Vergleich zu den nicht geleiteten Vorhersagen (vgl. Tabelle 8.6, Abschnitt 8.3) wurde die Selektivität der Anfragen durch die Pharmakophormodelle deutlich verbessert. Mit den verwendeten Modellen konnten deshalb bis zu siebenmal schnellere Screening-Läufe durchgeführt werden. Da σ – wie bereits beschrieben – mit der Laufzeit korreliert, erklärt dies die wesentliche Beschleunigung des geleiteten Screening-Prozesses. Darüber hinaus beeinflusst die Wahl eines adäquaten N_e die Anzahl der Anfragen nicht, sodass striktere Modelle grundsätzlich nicht zu einer weiteren Reduzierung des Suchraums führen. Um die Anzahl eingehaltener Inklusionen zu verifizieren, müssen Moleküle tatsächlich in der Bindetasche platziert werden. Da dies in der Bewertungshierarchie aber vor den frühen und späten Bewertungsphasen geschieht, kann die Erhöhung von N_e die Anzahl tatsächlich bewerteter Posen reduzieren. Striktere Modelle führen also dazu, dass teure Bewertungsberechnungen entfallen.

Tabelle 8.11: Zeitmessungen und Selektivität auf den $\text{ZINC}_{\text{CL1M/2M/3M}}$ Bibliotheken unter Verwendung guter N_e^g-Pharmakophormodelle. Durchschnittliche Laufzeiten pro Konformation t_s, pro Molekül t_f, parallele Laufzeit t_p und Anfragenselektivität σ.

	t_s [s]	t_f [s]	t_p [h] (1M)	t_p [h] (2M)	t_p [h] (3M)	σ
sahh	0.01	1.20	1.66	3.32	5.00	0.07
gpb	0.01	2.80	3.73	7.73	11.70	0.35
hsp90	0.01	1.23	1.66	3.43	5.15	0.21
fxa	0.01	3.32	4.38	8.93	13.87	0.75
er_agonist	0.02	5.68	7.62	15.10	24.05	1.28
dhfr	0.02	4.49	5.98	12.15	18.70	1.06

8.4.4 Molekülprofilgeleitetes Screening

Die von cRAISE propagierte Strategie, eine Molekülbibliothek einmalig zu präparieren und die abgeleitete Molekülinformation in Form des Deskriptorindex statisch für wiederholte Anfragen vorzuhalten, scheint für gewisse Anwendungen nachteilig zu sein. Nachteile entstehen insbesondere dann wenn anwendungsbezogen Molekülmengen vom Screening ausgeschlossen werden sollen, da sie im Kontext des betrachteten Zielproteins keine potentiellen Wirkstoffkandidaten darstellen (vgl. Abschnitt 2.5.3 und Abschnitt 2.5.4). Die Nutzung eines Molekülprofils zum geleiteten Screening begegnet diesem Problem. Es verwirft während des Screenings Moleküle, die den dadurch gegebenen Moleküleigenschaften widersprechen. So wird die Neuberechnung des Deskriptorindex auf einer zuvor eingeschränkten Molekülmenge überflüssig gemacht. Der geleitete Ansatz ist jedoch nur dann methodisch vertretbar, wenn eine Molekülfilterung während des Screenings, im Vergleich zu einer vor- oder nachbereitenden Filterung, keinen Laufzeitnachteil mit sich bringt. Um zu demonstrieren, dass dies tatsächlich der Fall ist, wurde exemplarisch das in Abschnitt 7.4.3 beschriebene Molekülprofil definiert. Es reduziert die $\text{ZINC}_{\text{CL3M}}$-Bibliothek um rund 75% auf 707 770 Moleküle mit eingeschränkten Leitstruktureigenschaften. Wurde das Profil zur simultanen Filterung der Bibliothek während des Screenings des Estrogenrezeptoragonisten (er_agonist) eingesetzt, so reduzierte sich die Anzahl angereicherter Moleküle um denselben Faktor. Die hierbei gemessenen Zeiten bestätigten das erwartete Laufzeitverhalten eines molekülprofilgeleiteten Screenings. Im Vergleich zum nicht geleiteten Screening (vgl. Tabelle 8.6) reduzierte sich die durchschnittliche Zeit für ein flexibles Docking t_f von 8,23 auf nun 2,06 Sekunden und die parallele Laufzeit t_p von zuvor 34,30 auf 9,25 Stunden. Die Laufzeit wurde also in gleichem Maß reduziert wie die Bibliothek eingeschränkt wurde. Das Experiment verdeutlicht, dass es generell möglich ist, Molekülprofile zur zeitgleichen Filterung im Screening zu nutzen, ohne dass dafür ein gesteigerter Zeitaufwand notwendig wird oder die Notwendigkeit entsteht, einen neuen Index auf einer eingeschränkten Bibliothek zu etablieren. Der einmalig präparierte Index kann so tatsächlich statisch vorgehalten werden und die Investitionen, die zu seiner Erstellung aufgebracht werden müssen, lohnen sich durch seine breite Anwendbarkeit in den unterschiedlichsten Screening-Szenarien.

8.5 Über die Relevanz der Bulk-Beschreibung

Abweichend von dem in Kapitel 7 skizzierten Ablauf der Evaluierung sei an dieser Stelle eine kurze Studie über die Relevanz der Bulk-Beschreibung des cRAISE-Deskriptors

beschrieben. Während der Finalisierung dieses Kapitels wurde ein kleiner aber entscheidender Fehler in der Platzierungsroutine von cRAISE festgestellt. Bei der Ausrichtung der Bulk-Beschreibung am Basisdreieck des Deskriptors (vgl. Abschnitt 6.6) drehte die zweite Transformation den Ikosaederstrahl nicht, wie eigentlich angedacht, in die Dreiecksebene hinein, sondern fälschlicherweise aus ihr heraus. Dieser Fehler führte dazu, dass die Bulk-Beschreibung nicht tranformationsinvariant und der Deskriptorabgleich somit abhängig von der ursprünglichen Orientierung eines Moleküls war. Die Invarianz der Deskriptoren bezüglich der Transformation bildet jedoch die Grundlage für den validen Deskriptorabgleich im Docking mit cRAISE. Mit dem Hintergrundwissen über die bereits geleisteten Erfolgsraten der Redocking- und Anreicherungsexperimente bot der Fehler deshalb Anlass dazu, die Relevanz der Bulk-Beschreibung zu hinterfragen. Um deren Einfluss bei der Platzierung zu untersuchen, wurde der Fehler behoben und die in Abschnitt 7.1 beschriebenen Redocking-Experimente und die in Abschnitt 7.2 beschriebenen Anreicherungs-Experimente wiederholt. Im Vergleich zu den bis dato durchgeführten Docking-Läufen bestand der wesentliche Unterschied mit den nun tatsächlich transformationsinvarianten Deskriptoren darin, dass maßgeblich mehr Deskriptortreffer erhalten wurden. Für viele Vorhersagen waren sie um ein Zwei- bis Dreifaches erhöht. Die nativen Moleküle des $\text{Astex}_{\text{ACS}}$ konnten also wesentlich einfacher in die Bindetasche eingepasst werden, da die Strahlen nun besser die Form der Moleküle reflektierten. Der durchschnittliche $\text{RMSD}_{\text{min Pose}}$ wurde dadurch leicht auf 0,89 Å verbessert. Da die massiv erhöhten Trefferzahlen für den effizienten Einsatz im Screening so allerdings nicht tragbar waren, mussten die Abgleichsparameter angezogen werden. Um den Durchsatz der Methode zu gewährleisten, wurden sie derart eingestellt, dass die Trefferzahlen durchschnittlich den Trefferzahlen der ursprünglichen Vorhersagen entsprachen. Die finalen Parameterwerte sind in Abschnitt 6.9 aufgeführt. Mit ihnen verbesserten sich die Erfolgsraten beim Redocking in den vorderen Rängen um 2–3% (vgl. Tabelle 8.3). Die durchschnittlichen Anreicherungswerte konnten in den Screening-Experimenten um 1% gesteigert werden (vgl. Tabelle 8.4). Generell erschwerten die nicht transformationsinvarianten Deskriptoren zwar die Platzierung von Molekülen, die Güte der Posen konnte zuvor aber mittels weicherer Abgleichsparameter gewährleistet werden. Die Resultate weichen nur geringfügig von den bisher vorgestellten Ergebnissen ab und ändern nicht die bereits getroffenen qualitativen Aussagen. Bei den nun folgenden Experimenten wurde der Fehler dennoch ausgeräumt.

8.6 Berücksichtigung von Zustandsänderungen

8.6.1 Abhängigkeit vom Eingabezustand

Zur Evaluierung des cRAISE-Multizustandsansatzes wurden die Experimente gemäß Abschnitt 7.5 durchgeführt. Zunächst wurde die Abhängigkeit der statischen Platzierung und Bewertung bei unterschiedlichen Eingabezuständen quantifiziert. Für jede Bindetasche der relevanten $\text{Astex}_{\text{ACS}}$-Strukturen wurden realistische Histidin- und Ligandzustände enumeriert. Bei Metalloproteinen, bei denen Histidine häufig Metallionen koordinieren, wurden Imidazolringe nahe des Ions im optimalen, koordinierenden Zustand gehalten. Unter Erhaltung des Eingabezustands wurde dann jeder Ligand- gegen jeden Rezeptorzustand statisch gedockt. Die Box-Whisker-Plots in Abbildung 8.20 fassen die Resultate dieser Kreuz-Docking-Studie zusammen und stellen die Unterschiede dar, die variierende Eingabezustände beim statischen Docking verursachten. Tabelle 8.12

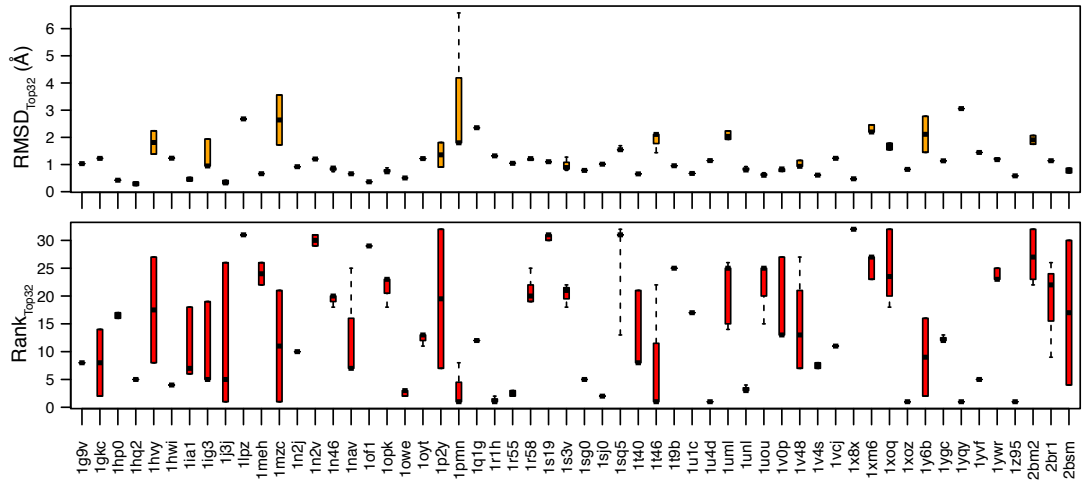


Abbildung 8.20: Einfluss von Zustandsvariationen auf die Posengenerierung und -bewertung beim statischen Redocking relevanter $\text{Astex}_{\text{ACS}}$-Komplexe. $\text{RMSD}_{\leq \text{Top32}}$ (oben) und Ränge der RMSD-besten Pose (unten) für jeden Kreuz-Docking-Lauf. Whisker repräsentieren die beste bzw. schlechteste Zustandskombination.

listet die relevanten $\text{Astex}_{\text{ACS}}$-Strukturen auf und gibt die Anzahl der Rezeptor- und Ligandzustände sowie der Zustandskombinationen an, die innerhalb der Kreuz-Docking-Studie berücksichtigt wurden. Für elf Komplexe beeinflussten die Eingabezustände die Platzierung der Liganden (gelbe Boxen). In seltenen Fällen ist der RMSD-Unterschied $\Delta\text{RMSD}_{\leq \text{Top32}}$ zwischen der besten und ungünstigsten Kombination größer 2 Å, d. h. die ungünstigste Kombination führte zu einem Docking-Fehlschlag, während die beste Kombination erfolgreich den Bindungsmodus reproduzieren konnte. Die Auswirkungen auf

Tabelle 8.12: Anzahl enumerierter Eingabezustände relevanter $\text{Astex}_{\text{ACS}}$-Komplexe, die innerhalb der Kreuz-Docking-Studie berücksichtigt wurden.

ID	Rezeptor	Ligand	Kombinationen	ID	Rezeptor	Ligand	Kombinationen
1g9v	3	1	3	1sg0	54	1	54
1gkc	3	2	6	1sj0	3	1	3
1hp0	1	2	2	1sq5	9	1	9
1hq2	9	2	18	1t40	27	1	27
1hvy	27	2	54	1t46	3	4	12
1hwi	9	1	9	1t9b	27	1	27
1ia1	3	3	9	1u1c	3	1	3
1ig3	3	3	9	1u4d	1	2	2
1j3j	3	3	9	1uml	9	2	18
1lpz	9	1	9	1unl	1	3	3
1meh	2	1	2	1uou	3	1	3
1mzc	27	2	54	1v0p	3	3	9
1n2j	9	1	9	1v48	3	2	6
1n2v	1	2	2	1v4s	3	2	6
1n46	3	1	3	1vcj	3	1	3
1nav	3	1	3	1x8x	3	1	3
1of1	9	1	9	1xm6	9	1	9
1opk	1	3	3	1xoq	27	2	54
1owe	9	1	9	1xoz	3	1	3
1oyt	3	1	3	1y6b	1	2	2
1p2y	1	2	2	1ygc	3	1	3
1pmn	1	4	4	1yqy	6	2	6
1q1g	9	2	18	1yvf	3	1	3
1r1h	3	1	3	1ywr	3	3	9
1r55	3	2	6	1z95	3	1	3
1r58	27	1	27	2bm2	3	2	6
1s19	9	1	9	2br1	1	3	3
1s3v	1	3	3	2bsm	3	2	6

die Platzierung ist der Tatsache geschuldet, dass der statische Docking-Ansatz einen Ligand auf Basis von Interaktionsstellen platziert, deren Typen vom gegebenen Eingabezustand abhängen. Sind diese Interaktionsstellen essenziell zur Verankerung des Liganden und ihre Interaktionstypen nicht komplementär, dann wird das Docking fehlschlagen. Im Gegensatz zu dieser extremen, eher seltenen Situation, ist der RMSD-Unterschied $\Delta\text{RMSD}_{\leq\text{Top32}}$ zwischen den meisten Kombinationen unter 0,5 Å oder sogar nicht erkennbar. In diesen Situationen ist die Platzierung erfolgreich, da der Ligand

über statische Interaktionsstellen positioniert werden kann. Allerdings wurden bei diesen Komplexen Auswirkungen auf die Posenbewertung beobachtet (vgl. Abbildung 8.20, rote Boxen). Werden ungünstige Zustandskombinationen genutzt, dann ist die Posenbewertung mit ungünstigen Protein-Ligand-Kontakten konfrontiert, was dazu führt, dass eigentlich gute Posen schlechter bewertet und auf niederen Rängen landen. Solche Rangvariationen können mit einer integrierten Optimierung wie Protoss, das den optimalen Zustand für jede Pose wählt bevor sie anschließend bewertet wird, abgefangen werden. Allerdings kann die Methode allein nicht die Diskrepanzen während der Platzierung lösen. Das Experiment demonstriert, dass die Zielsetzung des Ansatzes, konsistente Ergebnisse für jeden gegebenen Eingabezustand zu liefern, berechtigt ist. Wurde für die Kreuz-Docking-Experimente der dynamische Multizustandsansatz verwendet, konnten weder RMSD noch Rangabweichungen in den Resultaten beobachtet werden.

8.6.2 Ensemble-Docking vs. cRAISE-Multizustandsansatz

Um variierende Zustände mit Hilfe eines klassischen Docking-Ansatzes zu lösen, muss jeder Molekülzustand gegen jeden Rezeptorzustand gedockt, die Ausgabe vereinigt und die erhaltenen Posen gemäß der Bewertungsfunktion erneut sortiert werden (vgl. Abschnitt 3.8). Abbildung 8.21 vergleicht das Ergebnis dieser naiven Ensemble-Docking-Strategie, die auf den zuvor erhaltenen Kreuz-Docking-Resultaten angewendet wurde, mit dem Resultat des cRAISE-Multizustandsansatzes auf den 56 relevanten $\text{Astex}_{\text{ACS}}$-Komplexen. Prinzipiell realisiert die cRAISE-Methode dieselbe Aufgabe wie die eines

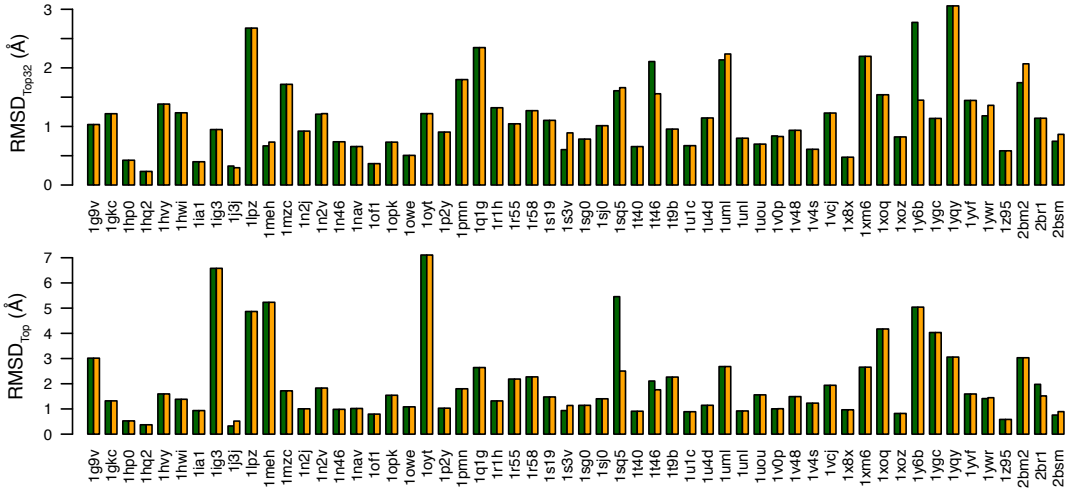


Abbildung 8.21: Vergleich des Ensemble-Dockings (orange) und des Multizustandsansatzes (grün). $\text{RMSD}_{\leq \text{Top32}}$ (oben), RMSD_{Top} (unten).

Ensemble-Ansatzes mit einigen Adaptierungen, die zu leicht veränderten Vorhersagen führen: Da die Verankerung von Posen über statische Interaktionsstellen für unterschiedliche Eingabezustände identische Posen produziert, häuft die simple Vereinigung der Resultate ein Vielfaches identischer Posen an. Im Komplex entgegengesetzte Multizustandsatome induzieren zudem Posenduplikate. Da beide Atome als Donor und Akzeptor fungieren können, werden Posen erzeugt, die sich nur darin unterscheiden, dass jeweils ein Zustand durch sein entsprechendes Gegenstück komplementiert wird. Der cRAISE-Ansatz vermeidet Duplikate beider Arten. Ohne ein weiteres, aufwändigeres Protokoll zur Nachbearbeitung der vereinigten Resultate kann der Ensembleansatz diese redundanten Posen nicht vermeiden. Darüber hinaus richtet die Postoptimierung der statischen Docking-Methode, die für das Kreuz-Docking genutzt wurde, lediglich frei rotierbare Wasserstoffe aus, um das Wasserstoffbrückennetzwerk zwischen Rezeptor und Pose zu optimieren. Dagegen darf der Multizustandsansatz in der Postoptimierungsphase auch Zustände adaptieren. Dies hat zu Folge, dass selbst bei vergleichbaren Posenkoordinaten die Netzwerke, das Optimierungsresultat und somit auch die Bewertung der Posen verschieden sein können. Bis auf diese Unterschiede, die in den dargestellten Ergebnissen offensichtlich werden, bestätigt das Experiment dennoch die qualitative Vergleichbarkeit beider Ansätze.

8.6.3 Laufzeitvergleich

Grundsätzlich erfordert die erhöhte Anzahl an Freiheitsgraden, um zusätzlich Rezeptor- und Ligandzustände zu prozessieren, einen erhöhten Rechenaufwand. Um den Aufwand, den cRAISE hierfür investiert, generell einordnen zu können, wurden die Laufzeiten unterschiedlicher Screening-Ansätze verglichen:

1. das statische Screening, das lediglich den Grundzustand von Rezeptor und Molekülen annimmt und während des Prozesses vollständig erhält,
2. die naive Ensemblestrategie, die initial Rezeptor- und Molekülzustände enumeriert und statische Docking-Läufe jeder möglichen Zustandskombination durchführt
3. und der cRAISE-Multizustandsansatz.

Die dafür notwendigen cRAISE Einzel- und Multizustandsindizes der Bibliotheken wurden einmalig präpariert und mit den in Abschnitt 7.5 beschriebenen, mutmaßlich rechenintensiven DUD-E-Zielstrukturen angefragt. Tabelle 8.13 listet die Anzahl der Rezeptor- und Molekülzustände, die innerhalb der Screening-Läufe von der Ensemble- und Multizustandsstrategie berücksichtigt wurden. Tabelle 8.14 listet die Laufzeiten, die in den Experimenten gemessen wurden.

Tabelle 8.13: Anzahl betrachteter Zustände rechenintensiver DUD-E-Mengen.

Zielstruktur	Rezeptorzustände[a]	Molekülzustände[b]	Kombinationen[c]
aces	3 (3)	1.34	4.02 (4.02)
ada	27 (3)	1.62	43.74 (4.86)
cxcr4	27 (9)	1.35	36.45 (12.15)
def	3 (3)	1.72	5.16 (5.16)
hdac8	27 (9)	1.72	46.44 (15.48)
hmdh	9 (9)	1.35	12.15 (12.15)
mapk2	3 (3)	1.58	4.74 (4.74)
met	27 (27)	1.34	36.18 (36.18)
mmp13	3 (3)	1.48	4.44 (4.44)
pde5a	27 (9)	1.33	35.91 (11.97)
pnph	9 (9)	1.75	15.75 (15.75)
pur2	27 (9)	1.64	44.28 (14.76)
sahh	81 (9)	1.82	147.42 (16.38)
vgfr2	3 (3)	1.40	4.20 (4.20)

[a] Zustände aktiver Residuen (Für Wasser zugängliche Rezeptorzustände)
[b] Durchschnittliche Anzahl der Zustände pro Molekül in der Bibliothek
[c] Durchschnittliche Anzahl von Zustandskombinationen pro Docking-Lauf

Tabelle 8.14: Zeitmessungen auf rechenintensiven DUD-E-Mengen.

	Grundzustand		Ensemble[a]		Multizustand	
Zielstruktur	t_s [s]	t_f [s]	t_s [s][a]	t_f [s][a]	t_s [s]	t_f [s]
aces	.10	23.35	.40 (.40)	93.86 (93.86)	.10	24.64
ada	.11	26.33	4.81 (.53)	1151.67 (127.96)	.15	37.87
cxcr4	.42	100.27	15.30 (5.10)	3654.84 (1218.28)	.49	116.87
def	.13	31.36	.67 (.67)	161.81 (161.81)	.23	53.89
hdac8	.10	23.39	4.64 (1.54)	1086.23 (362.07)	.14	33.99
hmdh	.21	44.32	2.55 (2.55)	538.48 (538.48)	.27	56.78
mapk2	.19	38.95	.90 (.90)	184.62 (184.62)	.22	45.36
met	.06	14.29	2.17 (2.17)	517.01 (517.01)	.07	16.33
mmp13	.13	29.26	.57 (.57)	129.91 (129.91)	.17	38.84
pde5a	.13	29.70	4.66 (1.55)	1066.52 (355.50)	.16	38.70
pnph	.24	52.87	3.78 (3.78)	832.70 (832.70)	.37	80.78
pur2	.19	46.31	8.41 (2.80)	2050.60 (683.53)	.25	59.39
sahh	.04	9.11	5.89 (.65)	1342.99 (149.22)	.06	12.78
vgfr2	.06	15.57	.25 (.25)	65.39 (65.39)	.07	18.15

[a] durch Einzel-Docking im Grundzustand × Anzahl Kombinationen abgeschätzt

Im Vergleich zum statischen Docking eines Grundzustands, erhöht sich die Laufzeit des Multizustandsansatzes bemerkenswerterweise kaum. Ein erhöhter Aufwand um einen Faktor von bis zu 1,5 ist allerdings notwendig, um die erhöhte Anzahl an Anfragedeskriptoren für Zielstrukturen mit vielen Rezeptorzuständen zu verarbeiten. Trotzdem benötigt ein starres Multizustands-Docking lediglich den Bruchteil einer Sekunde und ein flexibles Multizustands-Docking (mit durchschnittlich 240 Konformationen pro Molekül) kann innerhalb von mehreren Sekunden bis zu etwa einer Minute realisiert werden. Im Gegensatz dazu werden um mehrere Größenordnungen erhöhte Laufzeiten erwartet, wenn dieselbe Aufgabe mit dem naiven Ensembleansatz bewerkstelligt wird. Dies ist nicht überraschend, da speziell für histidinreiche Bindetaschen oder Bindetaschen, die Kofaktoren mit multiplen Zuständen enthalten, der naive Ansatz die Wiederholung einer beträchtlichen Anzahl individueller, statischer Docking-Läufe erfordert. Im schlimmsten Fall, der einer Brute-Force-Enumeration aller möglichen Zustände aktiver Residuen entspricht, kann ein Ensemble-Docking-Lauf eines einzelnen Moleküls bis zu mehrere Stunden dauern. Diese Zeit macht eine Anwendung des Ensembleansatzes für ein umfangreiches Screening praktisch unmöglich. Dagegen erlaubt der cRAISE-Ansatz die Prozessierung großer Bibliotheken. Selbst wenn der naive Ansatz nur die Zustände der aktiven Residuen aufzählt, für die ein Multizustandsatom tatsächlich auf der wasserzugänglichen Oberfläche vorzufinden ist, werden für flexible Docking-Läufe noch immer bis zu 20 Minuten benötigt. Die naive Ensemblestrategie ist bei simultaner Berücksichtigung von Rezeptorzuständen somit nicht in der Lage bezüglich der Laufzeit und so der praktischen Anwendbarkeit mit dem cRAISE-Ansatz mitzuhalten.

8.6.4 Ricin A in Komplex mit Neopterin

Der optimale Zustand von Rezeptor und Ligand offenbart sich häufig erst, wenn beide Komponenten im Komplex betrachtet werden (Holozustand). In einem Screening-Szenario ist der gebundene Zustand beider Komponenten allerdings *a priori* unbekannt. Objektive Zielstruktur- und Bibliothekspräparierungen sollten daher zumindest einen unabhängigen Grundzustand (Apozustand) des betrachteten Proteins und der Moleküle annehmen, um die Vorhersagen nicht bereits vorab in eine bestimmte Richtung zu lenken. Allerdings können die ungebundenen Apozustände vom eigentlich im Komplex gebundenen Holozustand abweichen. Beispiele für solche Situationen wurden bereits in der Literatur diskutiert.[239, 289] Abbildung 8.22 illustriert solch einen Fall am Beispiel von Ricin A im Komplex mit Neopterin (PDB: `1br5`[356]). Die Pterin-Gruppe nimmt in Wasser gelöst das wahrscheinlichere 3H-Tautomer an, das auch dem durch cRAISE angenommenen Grundzustand des Liganden entspricht. Es wurde behauptet, dass der Ligand

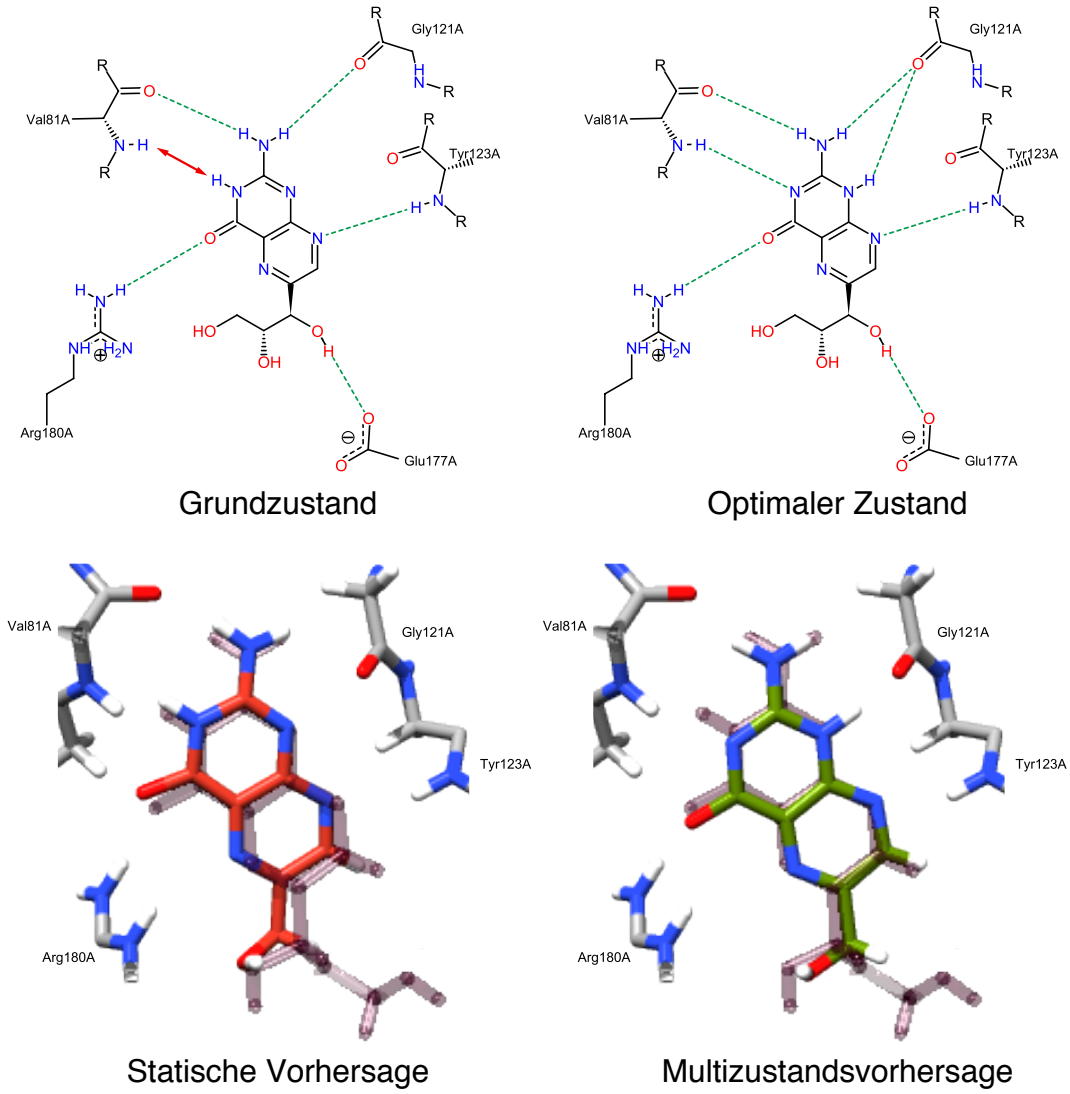


Abbildung 8.22: Vorhergesagter Zustand von Neopterin beim Multizustands-Docking in Ricin A. Oben: Die Pterin-Gruppe von Neopterin nimmt in Wasser das wahrscheinlichere 3H-Tautomer an (links). Der Zustand entspricht dem von cRAISE angenommenen Grundzustand, führt mit Ricin A jedoch zum unvorteilhaften Kontakt (roter Pfeil). Das unwahrscheinlichere 1H-Tautomer bindet an Ricin A (rechts). Unten: Das 1H-Tautomer wird auch mit dem Multizustandsansatz vorhergesagt (grün). Ein statisches Docking des Liganden im Grundzustand platziert das Molekül ähnlich über statische Interaktionsstellen (rot), wird jedoch schlechter bewertet. Das schattierte 3H-Tautomer des Kristallliganden ist zur Referenz dargestellt.

im Komplex allerdings das weniger wahrscheinlichere 1H-Tautomer annimmt.[239, 289]

Die Verwendung eines einzelnen Grundzustands kann somit nicht das optimale Docking-Resultat garantieren. Der von cRAISE propagierte Multizustandsansatz geht dieses Problem an. Ist es notwendig, so wird der Grundzustand von Rezeptor und/oder Ligand geändert. Abbildung 8.22 vergleicht die Top-Posen, die mit dem konventionellen, statischen und dem Multizustandsansatz unter Verwendung der 3H-Tautomer-Eingabe erhalten wurden. Die Posen unterscheiden sich nicht merklich bezüglich der resultierenden Atomkoordinaten, da auch im statischen Docking der Grundzustand über statische Interaktionsstellen verankert werden konnte. Dies ist in diesem Fall ausreichend, um eine nah native Pose zu erhalten. Allerdings kann der Multizustandsansatz den Zustand des Liganden zum eigentlich weniger wahrscheinlicheren 1H-Tautomer ändern. Dadurch können zwei zusätzliche Wasserstoffbrücken identifiziert werden, die die Bewertung des Liganden verbessern. Das diskutierte Beispiel verdeutlicht, dass es nicht ausreicht den wahrscheinlichsten Grundzustand von Liganden anzunehmen, sondern dass es notwendig ist, während des Dockings unterschiedliche Ligandzustände zu berücksichtigen. Dieser Umstand ist im Allgemeinen akzeptiert und wird von gängigen Methoden im Screening durch die Anreicherung von Molekülbibliotheken durch zuvor enumerierte Molekülzustände gelöst.

8.6.5 ALDR mit IDD594- und Fidarestat-ähnlichen Inhibitoren

Rezeptorseitig ist die Notwendigkeit verschiedene Zustände während des Dockings zu betrachten allerdings wenig akzeptiert. Gewöhnlicherweise wird vor einem Screening einmalig, anhand eines bereits bekannten Komplexes ein Rezeptorzustand eingestellt, der zu diesem individuellen Liganden passt. Die ähnliche Stabilität der wahrscheinlichsten Zustände des Imidazols legt jedoch nahe, dass verschiedene Liganden bei Interaktion mit Histidin unterschiedliche Zustände auf Rezeptorseite induzieren können. Die Abwesenheit von Wasserstoffkoordinaten in gewöhnlichen Kristallstrukturen erschwert es allerdings, solche Situationen zu beobachten. Wie in Abschnitt 7.5.3 beschrieben konnte anhand von äußerst hoch aufgelösten Strukturen dennoch das System der Aldosereduktase (ALDR) identifiziert werden, bei dem vermutlich dieser Fall eintritt. Abbildung 8.23 veranschaulicht, dass es hier notwendig ist, den Histidinzustand zu ändern, damit für unterschiedliche Liganden das typische Wasserstoffbrückenmuster mit den drei Schlüsselresiduen Tyr48, His110 und Trp111 in der aktiven Bindetasche etabliert werden kann. Zudem zeigt die Abbildung die von cRAISE vorhergesagten Rezeptorzustände, für die Top-Posen beim Redocking von IDD594 in `1us0` und von Fidarestat in `1pwm`. Für die Bindung mit dem IDD594-Inhibitor schlug cRAISE das $N^{\epsilon 2}$-H-Tautomer für His110 vor,

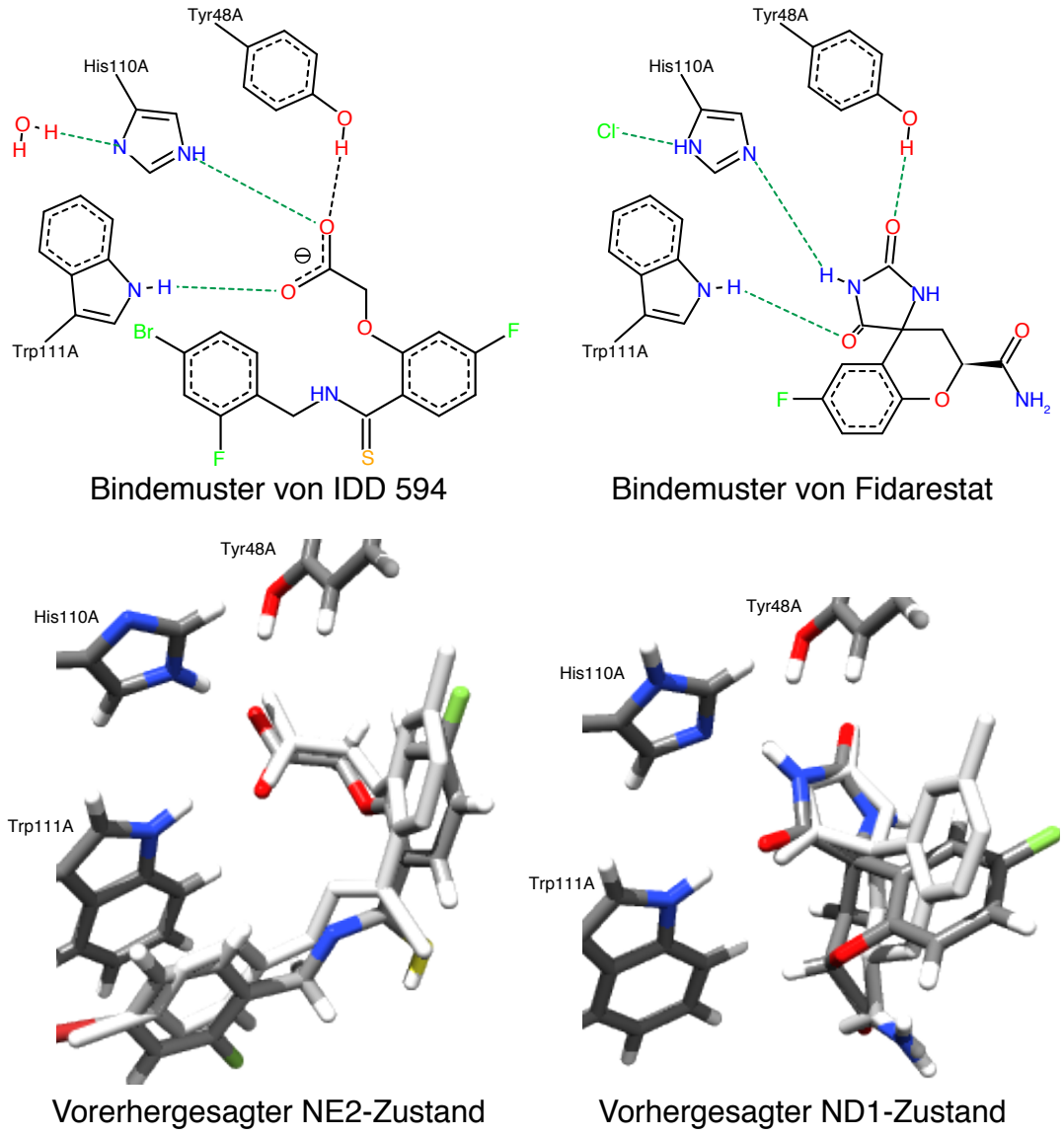


Abbildung 8.23: Vorhergesagte Rezeptorzustände beim Multizustands-Docking in ALDR: Top-Posen und Zustandsvorhersage beim Redocking von IDD594 (links, PDB: `1us0`) und Fidarestat (rechts, PDB: `1pwm`). Unterschiedliche His110-Zustände werden vorhergesagt. Kristallliganden (grau) zur Referenz.

wohingegen für Fidarestat das $N^{\delta 1}$-H-Tautomer vorhergesagt wurde. Durch die individuell vorhergesagten Rezeptorzustände kann das Histidin eine Wasserstoffbrücke zum Carboxylat von IDD594 bzw. zum 1'-positionierten Stickstoffatom der Spirohydantoin-Gruppe von Fidarestat etablieren. Da in beiden Fällen die direkte Umgebung des His110

ein essentieller Bestandteil des proteininternen Wasserstoffbrückennetzwerks darstellt, musste die aktive Bindungstasche mit einem Radius von 7,5 Å definiert werden. Bei geringerem Radius wurde im Fall von IDD594 ein positives Histidin angenommen, da die PROTOSS-Optimierung geladene Imidazole in der Nähe von negativ geladenen Carboxylaten forciert. Das Wassermolekül in der direkten Umgebung verhindert dagegen die Ladung. Für die Bindung von Fidarestat wurde in der Literatur ein wesentlich komplexerer Bindungsvorgang beschrieben.[360] Demnach induziert das Chloridion ein positiv geladenes Histidin, das das Wasserstoff an der 1'-Postion des Liganden akzeptiert und so wiederum einen negativ geladenen Liganden induziert. Tatsächlich forciert PROTOSS nicht die Ladung von Histidinen aufgrund von nah gelegenen Chloridionen. Zudem wird der als äußerst instabil betrachtete, negativ geladene Ligandzustand weder von cRAISE noch von PROTOSS berücksichtigt. Dennoch scheint der vorhergesagte Holozustand von Protein und Ligand durchaus plausibel, da er zumindest einen Übergangszustand im komplexen Bindungsmechanismus darstellt.

8.6.6 Bewertung von Zuständen

Der Gewinn einer adäquaten Zustandsselektion im virtuellen Screening hängt entscheidend von der genutzten Bewertungsfunktion ab, die letztendlich diesen Vorgang leitet. Bewertungsfunktionen, die durch Beiträge dominiert sind, die hauptsächlich die Komplementarität der Protein- und Ligandform bewerten, werden allerdings kaum einen Effekt von Zustandsvariationen innerhalb des Screenings bemerken. Tatsächlich gehört die cRAISE-Funktion (vgl. Abschnitt 6.10) zu dieser Klasse der Bewertungsfunktionen. In Screening-Experimenten wurde daher nur ein geringfügiger Nutzen für die Anreicherung festgestellt. Nichtsdestotrotz wird die Notwendigkeit, Rezeptor- und Ligandzustände innerhalb des Screening-Prozesses zu bestimmen durch die Beobachtungen gestützt, die innerhalb der in Abschnitt 7.5.4 beschriebenen Experimenten gemacht wurden. Hierbei wurde das HID- und das HIE-Ensemble der ALDR (vgl. Abschnitt 7.5.3) mit HYDE[347, 361] in vier unterschiedlichen Szenarien bewertet. In jedem der Experimente sollte für jeden bereits bekannten IDD594- und Fidarestat-ähnlichen Inhibitor die HYDE-Bewertungsfunktion den passenden der beiden Rezeptorzustände wählen. Welcher der Zustände von HYDE im betrachteten Kontext präferiert wurde, wurde durch $\Delta\Delta G_{\mathrm{Hyde}} = \Delta G_{\mathrm{Hyde}}(\mathrm{HIE110}) - \Delta G_{\mathrm{Hyde}}(\mathrm{HID110})$ angezeigt. Abbildung 8.24 stellt solch eine HYDE-Bewertung und die daraus gefolgerte Wahl des passenden Rezeptorzustands für die Kristallkomplexe der ALDR gebunden mit IDD594 (PDB: `1us0`) bzw. Fidarestat (PDB: `1pwm`) dar. Von Experiment zu Experiment stieg die betrachtete Problemkomplexität und somit auch die Anforderung an die Bewertungsfunktion.

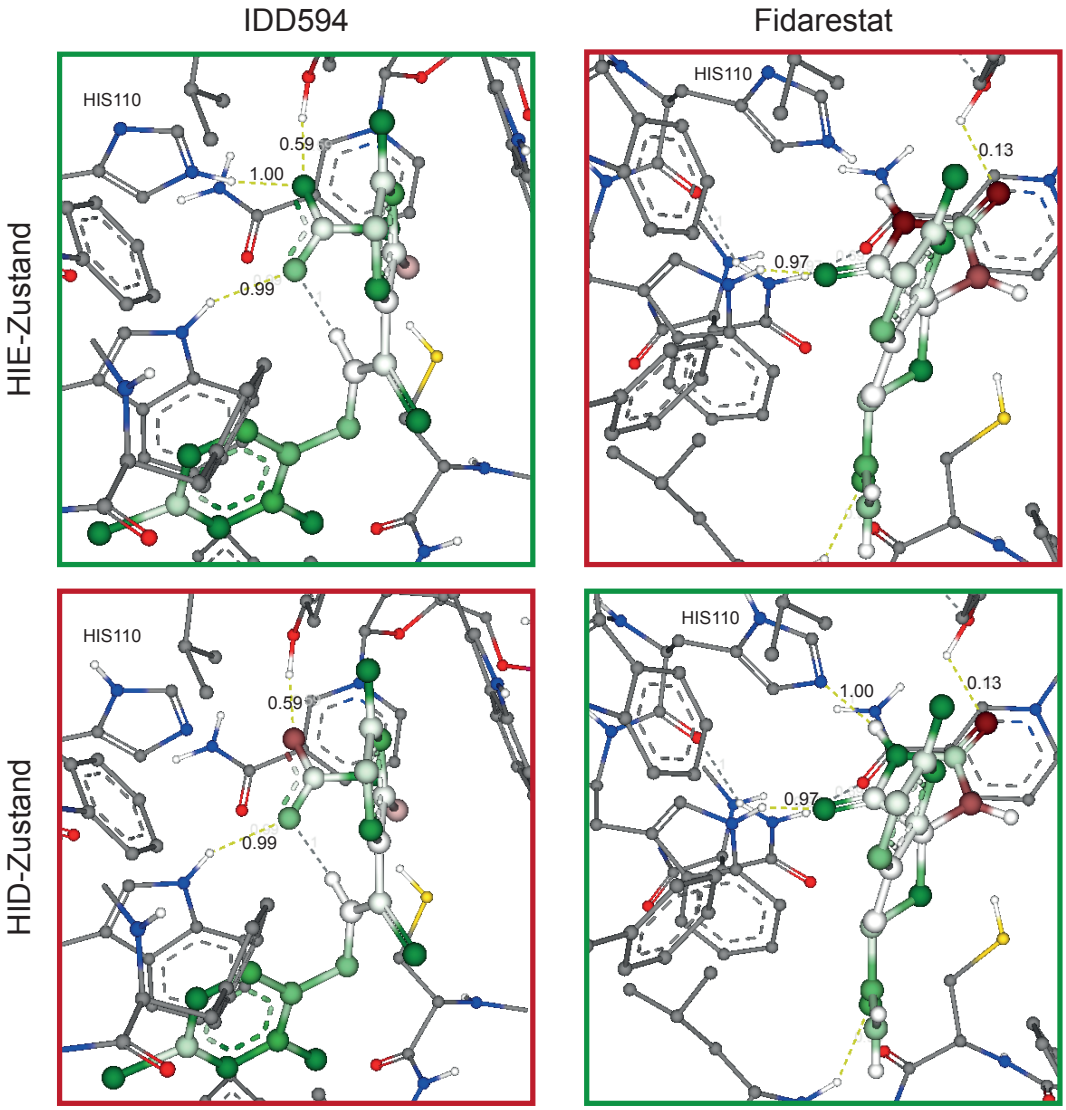


Abbildung 8.24: Hyde-Bewertung von IDD594 (`1us0`) und Fidarestat (`1pwm`) im HID- und HIE-Zustand der ALDR. Grün gefärbte Atome liefern einen günstigen, rote einen ungünstigen Beitrag zur Bewertung. Die dargestellten Werte geben den relativen Beitrag der Wasserstoffbrücken an. IDD594 wird im HIE-Zustand mit $-56,539$, im HID-Zustand mit $-45,381$ bewertet. Mit einem $\Delta\Delta G_{\text{Hyde}}$ von $-11,158$ wird der HIE-Zustand bevorzugt. Fidarestat wird im HIE-Zustand mit -5.768 und im HID-Zustand mit $-16,962$ bewertet. Mit einem $\Delta\Delta G_{\text{Hyde}}$ von 11,194 wird der HID-Zustand präferiert.

Abbildung 8.25 stellt das Resultat dieser Experimente in Form der Verteilungen der $\Delta\Delta G_{\text{Hyde}}$-Werte dar. Abbildung 8.25 a) zeigt das einfachste Szenario, bei dem jeder

Kristallligand mit seiner assoziierten Proteinkristallstruktur bewertet wurde. In Abbildung 8.25 b) wurde ein eher Screening-ähnlicheres Szenario konstruiert. Jeder Kristallligand der überlagerten Ensemblestrukturen wurde in einer Proteinstruktur bewertet (PDB: `1us0`). Abbildung 8.25 c) visualisiert $\Delta\Delta G_{\text{Hyde}}$-Werte, die nach erneuter Bewertung der Top-Posen aus statischen Redocking-Läufen der individuellen Komplexe erhalten wurden. Zuletzt sind in Abbildung 8.25 d) die $\Delta\Delta G_{\text{Hyde}}$-Werte der Top-Posen gezeigt, wenn jeder der ALDR-Aktiven des DUD-E gegen den `1us0`-Rezeptor statisch gedockt wurde. Gemäß der Erkenntnisse über das Bindungsmuster der ALDR mit ihren Inhibitoren sollte HYDE bei IDD594-ähnlichen Inhibitoren stets den HIE110-Zustand des Rezeptors präferieren. Dies ist bei einem negativen $\Delta\Delta G_{\text{Hyde}}$-Wert angezeigt. Für Fidarestat-ähnliche Inhibitoren sollte HYDE dagegen den HIE110-Zustand wählen, was bei einen positiven $\Delta\Delta G_{\text{Hyde}}$-Wert der Fall ist. Bei der Bewertung der Kristallstrukturen konnten genau diese Präferenzen beider Inhibitorklassen für den entsprechenden Rezeptorzustand festgestellt werden. Mit Erhöhung der Problemkomplexität, die zu immer ungenaueren Ligandkoordinaten führte, nahm die Tendenz für einen präferierten Rezeptorzustand jedoch stetig ab. Da die HYDE-Bewertungsfunktion keine Terme zur Abschätzung der sterischen Anordnung von Ligand und Rezeptor enthält und auch keine Torsionsspannungen in der Ligandkonformationen erkennt, wurden bei den Experimenten b)–d) vor der eigentlichen Bewertung zunächst mit einem numerischen Algorithmus die Koordinaten der überlagerten Liganden bzw. der Posen innerhalb der aktiven Bindetasche optimiert. Die Optimierung war allerdings nicht immer in der Lage, die Koordinatenabweichungen abzufangen, was notwendig gewesen wäre, um angemessen den Rezeptorzustandswechsel durch HYDE zu bemerken. Zudem beeinflussten die Rezeptorzustände den Optimierungsprozess selbst, da dieser bei ungünstiger Rezeptor-Ligand-Zustandskombination die ursprünglich gut ausgerichteten Liganden in andere Minima drückte. Dadurch konnten die unterschiedlichen Rezeptorzustände durch HYDE nicht immer objektiv miteinander verglichen werden. Ein anderes Problem ist in der innewohnenden Proteinflexibilität der Aldosereduktase zu finden. Während IDD594-ähnliche Inhibitoren in einer eher offenen Rezeptorkonformation binden, binden Fidarestat-ähnliche in einer eher geschlossenen. Weder cRAISE noch die numerische Optimierungsmethode berücksichtigen momentan Proteinflexibilität. Für die Experimente b) und c) wurde mit `1us0` zwar eine offene Rezeptorkonformation gewählt, sodass alle Liganden in der Tasche Platz finden. Für das tendenziell kleinere Fidarestat werden dadurch jedoch mehr unabgesättigte Atome erhalten, was HYDE rigoros bestraft.

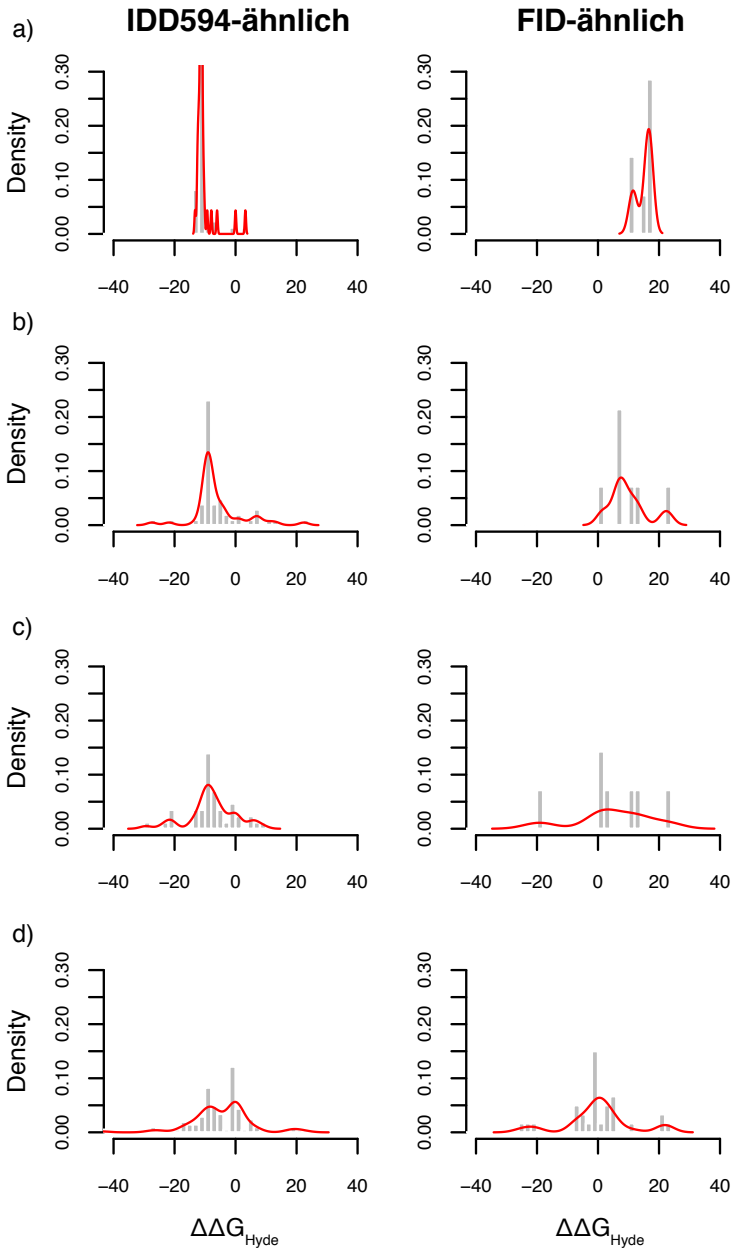


Abbildung 8.25: Relative Häufigkeiten der $\Delta\Delta G_{\text{Hyde}}$ zwischen HIE110- und HIE110-Rezeptorzustand bei Bewertung der ALDR mit bekannten Aktiven. Bewertung von a) Kristallkomplexen, b) `1us0` mit optimierten Ligandüberlagerungen, c) optimierten Top-Posen aus statischem Redocking und d) mit optimierten Top-Posen statisch gedockter DUD-E ALDR-Aktiven in `1us0`. Die Experimente a)-c) umfassen 43 IDD594- und 7 Fidarestat-ähnliche Inhibitoren, d) 129 IDD594-, 30 Fidarestat-ähnliche Inhibitoren.

9 Fazit und Ausblick

9.1 Zusammenfassung

cRAISE ist ein Werkzeug zum strukturbasierten virtuellen Screening, das auf den Konzepten von TrixX aufbaut. In dieser Arbeit wurden die Modelle von TrixX überarbeitet und erneut, auf Basis der Naomi-Bibliothek, implementiert. Die Neuentwicklung legte das Fundament zur Realisierung einer pharmakophor- und molekülprofilgeleiteten Suche, die zugleich die Proteinstruktur nutzt, um die Vorhersagen in eine gewünschte Richtung zu lenken. Zudem konnten weitere Freiheitsgrade in den Docking-Prozess integriert werden. Die indexbasierte Auswertung dieser Freiheitsgrade ermöglicht es, Protomere von Protein und Ligand simultan zu berücksichtigen und dennoch die Anwendbarkeit zum Screening umfangreicher Molekülbibliotheken zu gewährleisten.

9.1.1 Indexbasiertes virtuelles Screening

Die Naomi-basierte Implementierung ermöglicht eine einheitliche Handhabung von Molekül- und Proteininformation während der Präparierungs- und Screening-Phase. Dies ist insbesondere in einem zweigeteilten Screening-Prozess wichtig, bei dem ein wiederholter aber widersprüchlicher Aufbau der internen Molekülrepräsentation Folgeberechnungen maßgeblich stören kann. Die Zweiteilung erlaubt es, einen indexbasierten Ansatz zu verfolgen, der den Einsatz der Methode im Screening umfangreicher Molekülbibliotheken möglich macht. Mit Millionen von Molekülen konfrontiert garantiert der einmalig präparierte cRAISE-Deskriptorindex einen raschen Erhalt von Molekülinformation in den darauffolgenden Screening-Läufen. Durch den Deskriptorabgleich können plausible Bindungsmodusvorhersagen getätigt und eine Vorauswahl von potentiellen Molekülen aus Bibliotheken getroffen werden. Um von der indexbasierten Screening-Technologie zu profitieren, muss die vorberechnete Information permanent gespeichert und während ihrer kompletten Lebenszeit unverändert bleiben. Damit sich die Investitionen

zum Aufbau der Indexstruktur lohnen, muss der Index wiederholt, in verschiedenen Screening-Projekten genutzt werden können. Dies ist nur dann möglich, wenn die indexbasierte Suche eine breite Anwendbarkeit gewährleistet und diversen Anforderungen in den unterschiedlichsten Screening-Szenarien genügt. Die Methoden, die in dieser Arbeit vorgestellt wurden, bieten eine vielseitige Schnittstelle, um flexible Anfragen auf der statischen Bibliothek für verschiedene Screening-Projekte zu unterstützen.

9.1.2 Pharmakophor- und molekülprofilgeleitete Vorhersagen

Die pharmakophorgeleitete Posengenerierung und das molekülprofilgeleitete Screening sind insbesondere dann nützlich, wenn Hypothesen über entscheidende Schlüsselinteraktionen und/oder physikochemische Eigenschaften von potentiellen oder ungeeigneten Wirkstoffkandidaten bei Planung eines Screening-Projektes vorliegen. In solchen Situationen bietet cRAISE die Möglichkeit, sich auf besonders interessante Vorhersagen zu konzentrieren. Eine gegebene Pharmakophordefinition lenkt den Prozess derart, dass nur Molekülinformation extrahiert wird, die pharmakophorerfüllende Posen produzieren kann. Der pharmakophorgeleitete Ansatz führt dabei zu einer effektiven Suchraumreduktion und minimiert den Aufwand, der für einzelne Docking-Berechnungen geleistet werden muss. Posen, die den gegebenen Merkmalsdefinitionen widersprechen, werden entweder gar nicht erzeugt oder verworfen bevor sie aufwändig bewertet werden. Die Prozedur lässt dadurch pharmakophorerfüllende Posen hervortreten, ohne dass die zugrundeliegende Bewertungsfunktion angepasst werden muss. Die Einschränkung des Suchraums führt aber nicht zu einem Qualitätsverlust. Ganz im Gegenteil. Sind die Modelle adäquat präpariert, bieten sie die Chance, Bindungsmodusvorhersagen entscheidend zu verbessern. Die präsentierten Anreicherungsstudien zeigen, dass die Anreicherung sowohl früh als auch global verstärkt werden kann. Die globale Verstärkung ist auf die vorgestellten relaxierten Modelle zurückzuführen, die verschiedene Merkmalskombinationen evaluieren. Strikte Pharmakophormodelle können aber dennoch Moleküle extrahieren, die lediglich ein spezifisches Bindungsmuster etablieren. Insgesamt demonstriert die Methode wie extern Kontrolle über das strukturbasierte virtuelle Screening ausgeübt und zugleich eine drastische Beschleunigung des Prozesses erreicht werden kann. Durch die Effizienz des Ansatzes ist so eine Plattform geschaffen, um bereits aufgestellte Hypothesen in breitem Umfang zu testen. Durch repetitive Anwendung von Pharmakophormodellierung und pharmakophorgeleitetem Screening können beispielsweise initial vage Hypothesen verworfen, bestätigt oder verfeinert werden. Mit der Integration einer geleiteten Suche auf Basis von Molekülinformation sind zudem die Konzepte zweier Welten vereint. Molekülprofile erlauben es, konstitutionelle oder topologische

Bedingungen an Liganden zu konstatieren. Die zusätzlichen Auflagen beschränken die indexbasierte Suche weiter und filtern Moleküle während eines Screening-Laufes, ohne dass hierbei irgendein Effizienzverlust merkbar wird. cRAISE ist somit ein erster Schritt hin zu einer synergetischen virtuellen Screening-Plattform, die als Hybrid struktur- und ligandbasierte Techniken vereint und den Einsatz in einer Vielzahl unterschiedlichster Screening-Kampagnen erlaubt.

9.1.3 Vorhersage molekularer und makromolekularer Zustände

Gängige Molekül- bzw. Proteindateiformate können nur einen von mehreren möglichen Zuständen repräsentieren. Auch wenn sie sich nur geringfügig in der Position weniger Wasserstoffkoordinaten und einzelner Bindungstypen unterscheiden, beeinflussen die so dargestellten Tautomere und Protonierungszustände die physikochemischen Eigenschaften von Proteinen und Molekülen. Wenn strukturbasierte Methoden anhand der Eigenschaften eine Molekülplatzierung oder -bewertung vornehmen, können unterschiedliche Eingabezustände deshalb zu unterschiedlichen Vorhersagen führen. Da unterschiedliche Zustände aber zuletzt ein und dasselbe Molekül bzw. Protein repräsentieren, sollten die leicht unterschiedlichen Eingaben auch stets zum selben Ergebnis führen. Eine Normalisierung der Eingaben, indem nur ein wahrscheinlicher Grundzustand angenommen wird, ist nicht ausreichend, da er vom tatsächlich gebundenen Zustand im Protein-Ligand-Komplex abweichen kann.[239, 289] Eine protein- und ligandseitige Integration von Protomerfreiheitsgraden in den Docking-Prozess ist somit unumgänglich. Sie kann effizient Zustandsabhängigkeiten auflösen und unabhängig vom Eingabezustand, konsistente Vorhersagen machen. cRAISE bietet die Option, während geleiteter und nicht geleiteter Suchen, diese Freiheitsgrade zu berücksichtigen. Der Multizustands-Docking-Ansatz betrachtet sinnvolle Zustände von Ligand und Rezeptor. Ausgehend von einem wohldefinierten Grundzustand ändert er implizit den Zustand eines oder beider Komponenten im generierten Komplex, wenn dies die Etablierung von hydrophilen Interaktionen innerhalb des Wasserstoffbrückennetzwerks favorisiert. Atom-, Bindungstypen und Wasserstoffkoordinaten werden explizit angepasst, nachdem ein Molekül in die anvisierte Proteinbindetasche eingepasst wurde. Neben den Schweratomkoordinaten des Bindungsmodus sagt der Ansatz somit den angenommen Zustand im Protein-Ligand-Komplex voraus. cRAISE wurde dafür entwickelt, die Aufgabe eines Ensembleansatzes mit zuvor enumerierten Protein- und Ligandzuständen zu realisieren, ohne dabei dessen Nachteile erkennen zu lassen. Tatsächlich muss bei der Anwendung des Multizustandsansatzes weder Protein, noch Molekülbibliothek diesbezüglich präpariert werden. Zudem erzielt die Methode vergleichbare Vorhersagen, jedoch weitaus effizienter. Die indexbasierte

Strategie realisiert eine implizite Zustandsselektion auf Deskriptorebene, extrahiert lediglich passende Molekülzustände und vermeidet so die Evaluierung jeder möglichen Rezeptor-Ligand-Zustandskombination. Besonders wenn histidinreiche Rezeptoren oder aktive Bindetaschen betrachtet werden, die Kofaktoren mit innewohnenden Zuständen besitzen, ist das Screening mit mehreren Rezeptormodellen im Ensembleansatz normalerweise nicht praktikabel. In solchen Fällen ermöglicht cRAISE Screening-Projekte zu initiieren, die bislang nicht angegangen werden konnten.

9.2 Limitierungen

Obwohl mit der Entwicklung von cRAISE entscheidende Probleme des strukturbasierten virtuellen Screenings umfangreicher Molekülbibliotheken angegangen werden konnten, stoßen die hier entwickelten Methoden in bestimmten Bereichen an ihre Grenzen.

9.2.1 Proteinflexibilität

Neben der expliziten Modellierung rotierbarer, terminaler Wasserstoffbrückendonoren, -akzeptoren und Protomeren, modelliert cRAISE bis zu einem gewissen Grad Proteinflexibilität, indem es eine weiche Rezeptorstruktur abbildet. Beim Abgleich der Bulk-Strahlen und bei der Detektion von Überlappungen werden leichte Abweichungen toleriert. Die Bewertungsfunktion gewährt, durch weiche Repulsionsterme, ebenso leichte Überlappungen von Protein und Ligand. Eine derartige, indirekte Weitung der aktiven Bindetasche ist in einigen Fällen jedoch nicht ausreichend, wie es das Screening gegen die ALDR zeigt (vgl. Abschnitt 8.6.6). Ein Vergleich der ALDR-Strukturen offenbart eine flexible Schleife des Proteinrückgrats, die bei größeren Liganden eine geöffnete und bei kleineren Liganden eine geschlossene Konformation annimmt. Da cRAISE von der gegebenen Proteinkonformation abhängig ist, muss, um erfolgreich Platzierungen für beide Klassen von Aktiven zu erhalten, eine geöffnete Proteinbindetasche gewählt werden. Die geöffnete Proteinkonformation führt aber vor allem bei der Bewertung kleinerer Liganden dazu, dass Atome nicht abgesättigt werden. Das Beispiel verdeutlicht die Notwendigkeit, Proteinfreiheitsgrade im Docking-Prozess explizit zu integrieren. Dieses Thema ist seit Jahren Gegenstand von Forschungsprojekten.[94, 95] Um explizit Proteinflexibilität zu modellieren, muss ein Ensemble- oder Induced-Fit-Docking-Ansatz verfolgt werden, der zu Beginn eines Screenings die Flexibilität der Rezeptorstruktur analysiert, die weiteste Konformation der Bindetasche zur Platzierung der Liganden wählt und dann, vor der Bewertung, die Rezeptorstruktur individuell an die Form der Posen anpasst. Seitenkettenflexibilität und die Flexibilität des Proteinrückgrats werden

mittlerweile von einigen Methoden berücksichtigt. Die massiv erhöhte Anzahl der Freiheitsgrade führt jedoch zu einem erhöhten Rechenaufwand, sodass sie im Allgemeinen nicht zum Screening umfangreicher Molekülbibliotheken eingesetzt werden.

9.2.2 Bewertungsfunktion

Die cRAISE-Bewertungsfunktion wurde dazu entwickelt, die Passung von Posen zu evaluieren, um frühzeitig Überlappungen in Docking-Berechnungen zu registrieren und unpassende Posen auszusortieren. Im Screening eingesetzt hat sie jedoch Schwierigkeiten, unterschiedliche Moleküle zu vergleichen. Dies liegt vor allem daran, dass einige Beiträge nur sehr grob eingeschätzt werden oder völlig unberücksichtigt bleiben. Ionische (bis auf Metallinteraktionen) und aromatische Beiträge werden zu den Wasserstoffbrücken bzw. zu den hydrophoben Interaktionen gezählt und nicht weiter unterschieden. Ebenso wird auch nicht zwischen starken und schwachen Wasserstoffbrücken unterschieden. Letztere werden von der Bewertungsfunktion gar nicht detektiert. Zumindest auf Bewertungsebene wäre eine weitere Klassifizierung, mit einer entsprechenden Parametrisierung, vermutlich angebracht. Die cRAISE-Bewertungsfunktion berücksichtigt zudem keine entropischen Beiträge und andere, ungünstig wirkende Effekte. Beispielsweise wird die Spannung von Ligandtorsionen nur durch einen empirischen Parameter mit der Anzahl rotierbarer Bindungen verrechnet und nicht geometrisch, durch die Abweichung von optimalen Torsionen, eingeschätzt. Die Änderung eines Zustands von Protein oder Ligand fließt ebenso nicht in die Bewertung ein. Dies kann dazu führen, dass energetisch ungünstigere Zustände überschätzt werden. Vor allem beim Zustandswechsel können kooperative Effekte wirken, sodass Einzelbeiträge verstärkt oder abgeschwächt werden. Die Summation konstanter Einzelbeiträge ignoriert diese Effekte. (De-)solvatisierungen bildet cRAISE nicht ab. Diese Beiträge können aber dabei helfen, unterschiedliche Liganden besser einzuschätzen. Allerdings sollte dabei das Protein die Möglichkeit besitzen, sich an den Liganden anzupassen, sodass unabgesättigte Atome abgesättigt werden können und die Beiträge nicht zu einer Fehleinschätzung des Liganden führen.

9.2.3 Ligandoptimierung

In cRAISE ist eine numerische Postoptimierungsroutine integriert, die, aufgrund der erhöhten Laufzeit, allerdings nicht standardmäßig ausgeführt wird. Sie steht aber optional zur Verfügung. Häufig räumt eine Optimierung Abweichungen aus dem Weg, die durch die diskrete Platzierungsmethode verursacht werden und schafft dadurch eine bessere Grundlage zur Bewertung von Interaktionsgeometrien. In manchen Fällen driftet eine Pose während ihrer Optimierung jedoch derart ab, dass das Resultat nicht

mehr der eigentlich zu bewerteten Pose entspricht. Es stellt sich also die Frage, ob der Aufwand dieser Option den Nutzen rechtfertigt. Ligandoptimierungen wären aber zusammen mit der pharmakophorgeleiteten Screening-Methode interessant. Die Sphären der Pharmakophormerkmale stellen bei ihrer Auswertung sehr harte Kriterien dar und bereits geringfügige Platzierungsabweichungen können darüber entscheiden, ob die Hypothese erfüllt wird oder nicht. Eine Routine, die Ligandkoordinaten optimiert, mit dem Ziel möglichst viele Merkmale zu erfüllen, könnte die Nachteile aus dem Weg räumen. Sie könnte dafür sorgen, dass der Ligand nicht in ein anderes Minimum abdriftet und den Suchraum der Optimierungsroutine derart beschränkt wird, dass sie effizient auch zum Einsatz im Screening umfangreicher Molekülbibliotheken eingesetzt werden kann.

9.2.4 Abhängigkeit molekularer Eigenschaften vom Zustand

Die Molekülregistrierung des cRAISE-Multizustandsansatzes normalisiert die Eingaben, indem ein wahrscheinlicher Grundzustand angenommen wird. Moleküleigenschaften werden anhand der normalisierten Eingabe bestimmt. Einige der registrierten Moleküleigenschaften hängen allerdings von diesem Zustand ab. Dies betrifft insbesondere die Gesamtladung, die Anzahl potentieller Wasserstoffbrückendonoren und -akzeptoren, aber auch den berechneten Oktanol/Wasserverteilungskoeffizienten cLogP[367] und die topologische polare Oberfläche TPSA[368]. Ursache für die Diskrepanz ist, dass die Berechnung auf Valenzzustandsinformation zurückgreift. Es wurde bereits gezeigt, dass die Berechnung von Moleküleigenschaften für Moleküle, die tautomere Formen besitzen, nicht eindeutig ist.[239] Es stellt sich die Frage, ob die berechneten Eigenschaften als Entität des Zustands und nicht als Entität des Moleküls betrachtet werden sollten. In diesem Falle muss die Berechnung der Eigenschaften angepasst werden, da unterschiedliche Zustände von cRAISE nicht in der Datenbank registriert werden. Eine Zustandsenumeration kann aber auch während der Berechnung erfolgen und Variationen z. B. als Wertebereiche in der Datenbank registriert werden.

9.2.5 Modellierung weiterer Zustände

cRAISE baut auf dem VSC-Modell von Naomi auf, das unterschiedliche Zustände durch unterschiedliche Valenzzustandsfolgen repräsentiert, die auf Atomtypebene jedoch identische Atomtypfolgen besitzen. Dies impliziert, dass ein Zustandswechsel nie die Geometrie eines Atoms ändert. Für die Anwendung im virtuellen Screening ist diese Annahme in der Regel ausreichend, da dadurch Zustandsvariationen beschrieben werden, die keine große Überwindung einer Energiebarriere erfordern und so relativ häufig vorzufinden sind.[289] Tatsächlich können Moleküle aber auch Zustände einnehmen, die

eine Änderung der Atomgeometrie zur Folge haben. Ein klassisches Beispiel hierfür sind Keto-Enol-Tautomerien, die zentrale Atome von einem sp3- in einen sp2-hybridisierten Zustand überführen und so die Gestalt eines Moleküls ändern. Derartige Zustandsänderungen werden von cRAISE nicht berücksichtigt.

9.3 Weitere Anwendungen der RAISE-Technologie

Während der Entwicklung von cRAISE entstanden in Nebenprojekten andere Screening-Methoden, die einzelne Komponenten in einem anderen Kontext verwenden. Sie demonstrieren die generelle Einsetzbarkeit der in dieser Arbeit entwickelten Techniken.

9.3.1 Inverses strukturbasiertes Screening

Ein inverses strukturbasiertes virtuelles Screening soll für ein kleines Molekül potentielle Zielstrukturen identifizieren und dadurch ein möglichst vollständiges Zielstrukturprofil etablieren. Voraussetzung hierfür ist eine Bibliothek dreidimensionaler Proteinstrukturen. Ein Zielstrukturprofil kann dazu beitragen, die Proteinselektivität eines Moleküls vorherzusagen, um frühzeitig Seiteneffekte potentieller Wirkstoffkandidaten auszuschließen oder um Strukturen zu identifizieren, die bislang nicht als Teil des Wirkmechanismus erkannt wurden. Ebenso können für einen etablierten Wirkstoff neue Anwendungsgebiete erschlossen oder die Entwicklung von Wirkstoffen unterstützt werden, die mehrere Zielstrukturen angehen. Betrachtet man nicht nur humane Proteine, können Zielstrukturvorhersagen auch in der Biotechnologie eingesetzt werden, um den Einfluss neu entwickelter Stoffe auf andere Organismen einzuschätzen. Wird eine Docking-basierte Methode zum inversen Screening genutzt, kann neben der potentiellen Zielstruktur, gleichzeitig der Bindungsmodus des Moleküls aufgedeckt werden. Allerdings stellt die inverse Verwendung etwas andere Anforderungen. Es müssen Proteine präpariert und in der Bibliothek verwaltet werden. Zumindest Zehntausende von Proteinstrukturen sollten gehandhabt werden können, weshalb die Proteinpräparierung vollständig automatisiert verlaufen muss. Zudem müssen Proteinstrukturen bewertet und als Resultat des Screenings in der Hitliste präsentiert werden. Auf Basis der RAISE-Technologie wurde hierfür iRAISE[369] entwickelt. Es nutzt einzelne Komponenten und Konzepte von cRAISE und weiterentwickelte Methoden, um den Anforderungen bei inverser Anwendung gerecht zu werden. Wie cRAISE ist iRAISE eine zweigeteilte Prozedur, die eine Präparierungs- und Screening-Phase realisiert. Die Präparierungsphase muss ebenso nur einmalig durchgeführt und die abgeleitete Information kann wiederholt für verschiedene Screening-Läufe verwendet werden. Im Gegensatz zu cRAISE präpariert

iRAISE jedoch Proteininformation. Ausgehend von einer Bibliothek dreidimensionaler Proteinstrukturen werden Bindetaschen von Liganden im Komplex bestimmt. RAISE-Deskriptoren werden dann für alle Bindetaschen berechnet und im Index hinterlegt. Die Proteinstrukturen werden in einer relationalen Datenbank gespeichert, die die Information der Komplexe und der zuvor bestimmten Bindetaschen verwaltet. Für ein Anfragemolekül generiert die Screening-Phase Konformere, für die RAISE-Anfragedeskriptoren berechnet werden. Die Anfragedeskriptoren werden mit den Indexdeskriptoren abgeglichen und jede resultierende Pose wird innerhalb einer Scoring-Kaskade bewertet. Das Resultat ist eine Liste bewerteter Zielstrukturen mit vorhergesagten Bindungsmodi des Anfragemoleküls in den getroffenen Zielstrukturen, die entsprechend der iRAISE-Bewertung sortiert ist.

9.3.2 Bindetaschenvergleich

Ähnlichkeiten zwischen Proteinen können für Anwendungen im Kontext des Wirkstoffentwurfs hilfreich sein. Ist ein Protein mit bekannter Funktion gegeben, so kann dieses Wissen auf ähnliche Proteine mit unbekannter Funktion transferiert werden und Hinweise auf neue Zielstrukturen, Kreuzreaktivitäten oder Seiteneffekte liefern. In biotechnologischen Anwendungen kann die Ähnlichkeit Aufschlüsse über die Substratspezifität von Enzymen oder über potentielle Mutationsstellen zur Enzymoptimierung geben. Unter der Voraussetzung, dass ähnliche Liganden in ähnliche Bindetaschen binden, modellieren Methoden zum Bindetaschenvergleich die Ähnlichkeit indirekt über die Ligandbindung. Die drei wesentlichen Bestandteile dieser Methoden sind die Kodierung der für die Ligandbindung verantwortlichen Bindetaschenmerkmale, eine Ähnlichkeitssuche und eine Bewertungsphase. Initial wird die Komplexität des Vergleichs reduziert, indem die Bindetasche vereinfacht durch entscheidende Merkmale repräsentiert wird. Anschließend werden ähnliche Bindetaschenrepräsentationen identifiziert. Zuletzt quantifiziert eine Bewertungsfunktion deren Ähnlichkeit zur Referenz. TrixP[370] ist eine Methode zum indexbasierten Bindetaschenvergleich, die die RAISE-Technologie und Funktionalitäten von cRAISE verwendet. TrixP nutzt wie cRAISE zur Anfrage die Bindetasche eines Proteins, gleicht jedoch die abgeleiteten Anfragedeskriptoren mit Deskriptoren ab, die wie bei iRAISE zuvor von Proteinbindetaschen abgeleitet wurden. Diese können zuvor automatisch mit DoGSite[97] für die Bibliotheksproteine vorhergesagt werden. Ein wesentlicher Unterschied zu cRAISE und iRAISE besteht darin, dass während des Deskriptorabgleichs nicht die Komplementarität der Deskriptoren bewertet wird, sondern die Deskriptoren von Anfrage und Index tatsächlich abgeglichen werden. Zudem berücksichtigt TrixP indirekt Proteinflexibilität, indem nur ein Teil der Bulk-Strahlen zum

Abgleich herangezogen wird. Deskriptortreffer identifizieren Bindetaschen der Bibliothek, die auf die Anfrage superpositioniert werden. Die dafür notwendige Transformation wird anhand von Clustern von Deskriptortreffern bestimmt. Um die Ähnlichkeit der getroffenen Bindetaschen bezüglich der Anfrage zu quantifizieren, wird die Ausrichtung der Interaktionsstellen in den überlagerten Bindetaschen bewertet.

9.4 Perspektiven

Grundlegende Komponenten von cRAISE wurden in andere Methoden überführt. Der geleitete Screening- und der Multizustandsansatz wurden bislang allerdings nicht adaptiert. Sie bieten noch Potential für innovative Neuentwicklungen.

9.4.1 Anwendungsmöglichkeiten des geleiteten Screening-Ansatzes

Ein statischer Proteindeskriptorindex, wie er von iRAISE und TrixP propagiert wird, hat den Vorteil, rasch Proteininformation während der Screening-Phase zu extrahieren. Die Investitionen zum Aufbau der Indexstruktur lohnen sich auch in diesen Anwendungsbereichen nur, wenn der Index wiederholt, in unterschiedlichen Screening-Projekten genutzt werden kann. Die Konzepte zur geleiteten Suche könnten daher auch auf iRAISE und TrixP übertragen werden, um vielfältige Arten von Anfragen umzusetzen. Hierfür müssten während der Präparierungsphase Protein- und Bindetascheneigenschaften berechnet und in der Datenbank registriert werden. Zudem muss die Möglichkeit geschaffen werden, anhand der Eigenschaften ein Proteinprofil zu definieren und dieses zur Formulierung von Bereichs- oder Existenzanfragen auf den registrierten Eigenschaften zu nutzen. So können Teilmengen der Proteinbibliothek während des Screenings ausgeschlossen und eine Neuetablierung des Proteinindex vermieden werden. Bereits die einfache Kennzeichnung der Proteine anhand der bekannten/unbekannten Funktion, des Organismus oder der subzellulären Lage würde genügen, um vielfältige Screening-Projekte und Vergleiche zu unterstützen. Einfache Komplexeigenschaften, wie das Vorkommen von Metallionen, spezifische Kofaktoren oder charakteristische Aminosäureanordnunungen wie katalytische Diaden und Triaden definieren bereits Klassen von Bindetaschen und können als Ausschlusskriterium im Screening dienen. Ergänzt mit Kennzahlen, die Volumen, zugängliche Oberfläche, Hydrophilität oder Hydrophobizität einer Bindetasche beschreiben, ist die Voraussetzung geschaffen, um anwendungsbezogen diverse Fragestellungen zu beantworten.

Die Anwendung eines Pharmakophormodells ist nicht auf das strukturbasierte VS beschränkt.[371] Ein Pharmakophormodell kann auch zum inversen Screening oder Bin-

detaschenvergleich eingesetzt werden.[372, 373] Bewertungsfunktionen, die kleine Moleküle im strukturbasierten VS bewerten, haben im inversen Ansatz Schwierigkeiten, Proteine zu bewerten. Eine geometrische Einschätzung der Ähnlichkeit von Bindetaschen wird häufig durch die Proteinflexibilität erschwert. Pharmakophorgeleitete Vorhersagen bieten die Chance, diese Hindernisse zu überwinden. Dafür kann der pharmakophorgeleitete Ansatz von cRAISE mit geringeren Anpassungen auf iRAISE bzw. TrixP übertragen werden. Das Modell muss zunächst invertiert werden, damit die Inklusion von Proteinatomen überprüft werden kann. In beiden Fällen kann das Modell die Anzahl der Anfragedeskriptoren deutlich reduzieren, sodass eine entscheidende Beschleunigung der Prozesse zu erwarten ist. Die globale Überprüfung der Hypothese erfolgt für jeden Deskriptortreffer. Beim inversen Ansatz müssen die Pharmakophormerkmale jedoch inklusive des Moleküls in die getroffene Bindetasche transformiert werden.

9.4.2 Anwendungsmöglichkeiten des Multizustandsansatzes

Da Pharmakophormodelle nur wenige Merkmale definieren, um essentielle Interaktionen zu repräsentieren, kann die Multizustandsmethode besonders beim pharmakophorgeleiteten VS von großem Wert sein. Hier kann eine Zustandsänderung dazu führen, dass ein Molekül die Hypothese fälschlicherweise nicht erfüllt. cRAISE berücksichtigt bereits Rezeptor- und Ligandzustände bei der Anwendung mit Pharmakophormodellen. Dabei wählt die proteinseitige Enumeration der Multizustandsdeskriptoren nur solche, die zumindest lokal das Modell erfüllen. Die globale Testung überprüft bei Donor- und Akzeptormerkmalen nicht nur den komplementären Atomtyp, sondern auch ob ein in der Sphäre enthaltenes Posenatom ein Multizustandsatom ist. Inwiefern das pharmakophorgeleitete SBVS davon profitiert, bleibt noch zu zeigen.

Das Konzept der indexbasierten Zustandsselektion auf Deskriptorebene, wie es in dieser Arbeit eingeführt wurde, kann als allgemeine Vorlage dienen, um Tautomere und Protonierungszustände in andere, merkmalsbasierte Screening-Methoden zu integrieren. Während die Methode ohne weitere Adaptierungen auf das inverse VS übertragbar ist, kann es auch für merkmalsbasierte Ähnlichkeitssuchen im ligandbasierten VS, beim Bindetaschenvergleich oder der Pharmakophormodellierung nützlich sein. Ähnlichkeitssuchen erfordern jedoch eine andere Realisierung der expliziten Zustandsselektion. Prinzipiell könnte der Referenzzustand übertragen werden. Die Frage ist, ob dies überhaupt notwendig ist. Zumindest die Bewertungsfunktion sollte aber alternative Zustände registrieren und sie dann identisch bewerten. Grundsätzlich zeigt der cRAISE-Ansatz wie mehr Freiheitsgrade effizient in den Screening-Prozess integriert werden können.

Literaturverzeichnis

[1] Schellhammer, I. and Rarey, M. Trixx: structure-based molecule indexing for large-scale virtual screening in sublinear time. *J Comput Aid Mol Des* **21**(5), 223–238, May (2007). 3, 245

[2] Schlosser, J. and Rarey, M. Beyond the virtual screening paradigm: structure-based searching for new lead compounds. *J Chem Inf Model* **49**(4), 800–809, Apr (2009). 3, 37, 63, 85, 245

[3] Urbaczek, S., Kolodzik, A., Fischer, J. R., Lippert, T., Heuser, S., Groth, I., Schulz-Gasch, T., and Rarey, M. Naomi: on the almost trivial task of reading molecules from different file formats. *J Chem Inf Model* **51**(12), 3199–3207, Dec (2011). 4, 63, 65, 245

[4] Ehrlich, P. Ueber den zusammenhang von chemischer constitution und wirkung. *Munchen Med Wochen* **-**, 1654–1655 (1898). 7

[5] Ehrlich, P. Die grundlagen der experimentellen chemotherapie. *Angew Chem* **23**(1), 2–8 (1910). 7

[6] Hüntelmann, A. C. 1910. transformationen eines arzneistoffes—vom 606 zum salvarsan. In Arzneimittel des 20. Jahrhunderts. 13 historische Skizzen von Lebertran bis Contergan., Eschenbruch, N., Balz, V., Klöppel, U., and Hulverscheidt, M., editors, 17–52. transcript Verlag, Bielefeld (2009). 7

[7] Alanine, A., Nettekoven, M., Roberts, E., and Thomas, A. W. Lead generation–enhancing the success of drug discovery by investing in the hit to lead process. *Comb Chem High Throughput Screen* **6**(1), 51–66, Feb (2003). 8

[8] Steinmeyer, A. The hit-to-lead process at schering ag: strategic aspects. *Chemmedchem* **1**(1), 31–36, Jan (2006). 8

[9] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. Principles of early drug discovery. *Brit J Pharmacol* **162**(6), 1239–1249, Mar (2011). 8

[10] U. S. Food and Drug Administration. Drug approval process. `http://www.fda.gov/downloads/Drugs/ResourcesForYou/Consumers/UCM284393.pdf`. accessed Nov 11, 2013. 9

[11] Voet, D. and Voet, J. G. *Biochemistry*. Wiley, 4th edition, (2010). 10

[12] Skipper, L. Proteins | overview. In Encyclopedia of Analytical Science, Worsfold, P., Townshend, A., and Poole, C., editors, 344–352. Elsevier, Oxford2nd edition (2005). 10

[13] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Res* **28**(1), 235–242, Jan (2000). 10

[14] Goldenberg, D. Protein folding and assembly. In Encyclopedia of Biological Chemistry, Lennarz, W. J. and Lane, M. D., editors, 625–631. Academic Press, Waltham (2013). 10

[15] Charton, M. and Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology* **99**(4), 629–644 (1982). 10, 13

[16] Singh, J., Petter, R. C., Baillie, T. A., and Whitty, A. The resurgence of covalent drugs. *Nat Rev Drug Discov* **10**(4), 307–317, Apr (2011). 11

[17] Klebe, G. and Böhm, H. J. Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J Recept Signal Transduct Res* **17**(1-3), 459–473 (1997). 13

[18] Bissantz, C., Kuhn, B., and Stahl, M. A medicinal chemist's guide to molecular interactions. *J Med Chem* **53**(14), 5061–5084, Jul (2010). 13, 14, 15, 16

[19] Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber Dtsch Chem Ges* **27**(3), 2985–2993 (1894). 14

[20] McNaught, A. D. and Wilkinson, A., editors. *IUPAC. Compendium of Chemical Terminology, (the "Gold Book").* Blackwell Scientific Publications, Oxford, Oxford, 2nd edition, (1997). XML online corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. 14

[21] Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D. C., Crabtree, R. H., Dannenberg, J. J., Hobza, P., et al. Definition of the hydrogen bond (iupac recommendations 2011). *Pure Appl Chem* **83**(8), 1637–1641 (2011). 14

[22] Panigrahi, S. K. and Desiraju, G. R. Strong and weak hydrogen bonds in the protein-ligand interface. *Proteins* **67**(1), 128–141, Apr (2007). 15

[23] Kumar, S. and Nussinov, R. Close-range electrostatic interactions in proteins. *Chembiochem* **3**(7), 604–617, Jul (2002). 15

[24] Archibald, S. and Smith, R. 3.22 – protein-binding metal complexes: Noncovalent and coordinative interactions. In Comprehensive Inorganic Chemistry II, Reedijk, J. and Poeppelmeier, K., editors, 661–682. Elsevier Amsterdam2nd edition (2013). 15

[25] Dokmanić, I., Sikić, M., and Tomić, S. Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. *Acta Crystallogr D Biol Crystallogr* **64**(Pt 3), 257–263, Mar (2008). 15

[26] Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M., and Tanabe, K. Origin of attraction and directionality of the pi/pi interaction: model chemistry calculations of benzene dimer interaction. *J Am Chem Soc* **124**(1), 104–112, Jan (2002). 15

[27] Salonen, L. M., Ellermann, M., and Diederich, F. Aromatic rings in chemical and biological recognition: energetics and structures. *Angew Chem Int Ed Engl* **50**(21), 4808–4842, May (2011). 15, 16

[28] Ma, J. C. and Dougherty, D. A. The cation–π interaction. *Chem Rev* **97**(5), 1303–1324, Aug (1997). 16

[29] Sirimulla, S., Bailey, J. B., Vegesna, R., and Narayan, M. Halogen interactions in protein-ligand complexes: Implications of halogen bonding for rational drug design. *Journal of Chemical Information and Modeling* **0**(0), null (2013). 16

[30] Politzer, P., Murray, J. S., and Clark, T. Halogen bonding and other σ-hole interactions: a perspective. *Phys Chem Chem Phys* **15**(27), 11178–11189, Jul (2013). 16

[31] Pal, D. and Chakrabarti, P. Non-hydrogen bond interactions involving the methionine sulfur atom. *J Biomol Struct Dyn* **19**(1), 115–128, Aug (2001). 16

[32] Iwaoka, M. and Isozumi, N. Hypervalent nonbonded interactions of a divalent sulfur atom. implications in protein architecture and the functions. *Molecules* **17**(6), 7266–7283 (2012). 16

[33] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**(1-3), 3–26, Mar (2001). 16, 17

[34] Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of alogp and clogp methods. *The Journal of Physical Chemistry A* **102**(21), 3762–3772, May (1998). 17

[35] Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *J Comb Chem* **1**(1), 55–68, Jan (1999). 17

[36] Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **45**(12), 2615–2623, Jun (2002). 17

[37] Oprea, T. I., Davis, A. M., Teague, S. J., and Leeson, P. D. Is there a difference between leads and drugs? a historical perspective. *J Chem Inf Comput Sci* **41**(5), 1308–1315 (2001). 17, 18

[38] Teague, Davis, Leeson, and Oprea. The design of leadlike combinatorial libraries. *Angew Chem Int Ed Engl* **38**(24), 3743–3748, Dec (1999). 18

[39] Hann, M. M. and Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **8**(3), 255–263, Jun (2004). 18

[40] Congreve, M., Carr, R., Murray, C., and Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* **8**(19), 876–877, Oct (2003). 18

[41] Yusof, I. and Segall, M. D. Considering the impact drug-like properties have on the chance of success. *Drug Discovery Today* **18**(13-14), 659–666, Jul (2013). 18

[42] Cummings, M. D., Arnoult, E., Buyck, C., Tresadern, G., Vos, A. M., and Wegner, J. K. *Preparing and Filtering Compound Databases for Virtual and Experimental Screening*, 35–59. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 18, 19

[43] Pajouhesh, H. and Lenz, G. R. Medicinal chemical properties of successful central nervous system drugs. *NeuroRx* **2**(4), 541–553, Oct (2005). 18

[44] von Nussbaum, F., Brands, M., Hinzen, B., Weigand, S., and Häbich, D. Antibacterial natural products in medicinal chemistry – exodus or revival? *Angew Chem Int Ed Engl* **45**(31), 5072–5129, Aug (2006). 18

[45] O'Shea, R. and Moser, H. E. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J Med Chem* **51**(10), 2871–2878, May (2008). 18

[46] Bolognesi, M. L. Polypharmacology in a single drug: Multitarget drugs. *Current Medicinal Chemistry* **20**(13), 1639–1645, Mar (2013). 19

[47] Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. V. Data completeness–the achilles heel of drug-target networks. *Nat Biotechnol* **26**(9), 983–984, Sep (2008). 19

[48] Langer, T. and Hoffmann, R. D., editors. *Pharmacophores and Pharmacophore Searches*. Methods and Principles in Medicinal Chemistry. Wiley Blackwell (John Wiley & Sons), , Jul (2006). 20, 44

[49] Kier, L. B. Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. *Mol Pharmacol* **3**(5), 487–494, Sep (1967). 20

[50] Kier, L. B. *Molecular Orbital Theory in Drug Research (Medicinal chemistry)*. Academic Press Inc, (1971). 20

[51] Van Drie, J. H. Monty kier and the origin of the pharmacophore concept. *Internet Electron J Mol Des* **6**, 271–279 (2007). 20

[52] Wermuth, C. G., Ganellin, C. R., Lindberg, P., and Mitscher, L. A. Glossary of terms used in medicinal chemistry (iupac recommendations 1998). *Pure Appl Chem* **70**(5), 1129–1143 (1998). 20

[53] Petukh, M., Stefl, S., and Alexov, E. The role of protonation states in ligand-receptor recognition and binding. *Curr Pharm Des* **19**(23), 4182–4190 (2013). 22, 54

[54] Sayle, R. So you think you understand tautomerism? *Journal of Computer-Aided Molecular Design* **24**(6-7), 485–496 (2010). 23, 55

[55] Koshland Jr, D. Application of a theory of enzyme specificity to protein synthesis. *P Natl Acad Sci Usa* **44**(2), 98 (1958). 24

[56] Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., and Greyson, D. The cost of drug development: a systematic review. *Health Policy* **100**(1), 4–17 (2011). 25

[57] Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**(3), 203–214, Mar (2010). 25

[58] Bleicher, K. H., Böhm, H.-J., Müller, K., and Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* **2**(5), 369–378, May (2003). 25

[59] Talele, T. T., Khedkar, S. A., and Rigby, A. C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr Top Med Chem* **10**(1), 127–141 (2010). 25

[60] Buckle, D. R., Erhardt, P. W., Ganellin, C. R., Kobayashi, T., Perun, T. J., Proudfoot, J., and Senn-Bilfinger, J. Glossary of terms used in medicinal chemistry. part ii (iupac recommendations 2013). *Pure and Applied Chemistry* **85**(8), 1725–1758, Jul (2013). 25

[61] Walters, W., Stahl, M. T., and Murcko, M. A. Virtual screening: an overview. *Drug Discovery Today* **3**(4), 160–178, Apr (1998). 26

[62] Boehm, M. *Virtual Screening of Chemical Space: From Generic Compound Collections to Tailored Screening Libraries*, 1–33. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 26

[63] Sotriffer, C., editor. *Virtual Screening: Principles, Challenges, and Practical Guidelines*. Methods and Principles in Medicinal Chemistry. Wiley Blackwell (John Wiley & Sons), , Jan (2011). 26

[64] Kirchmair, J., Distinto, S., Schuster, D., Spitzer, G., Langer, T., and Wolber, G. Enhancing drug discovery through in silico screening: strategies to increase true positives retrieval rates. *Curr Med Chem* **15**(20), 2040–2053 (2008). 26, 30

[65] Koeppen, H., Kriegl, J., Lessel, U., Tautermann, C. S., and Wellenzohn, B. *Ligand-Based Virtual Screening*, 61–85. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 27, 33, 57

[66] Eckert, H. and Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* **12**(5-6), 225–233, Mar (2007). 27

[67] Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* **11**(13-14), 580–594, Jul (2006). 27, 58

[68] Bron, C. and Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* **16**(9), 575–577, Sep (1973). 28

[69] Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **32**(5), 922–923, Sep (1976). 28

[70] Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **34**(5), 827–828, Sep (1978). 28

[71] Brint, A. T. and Willett, P. Algorithms for the identification of three-dimensional maximal common substructures. *Journal of Chemical Information and Modeling* **27**(4), 152–158, Nov (1987). 28

[72] Lemmen, C. and Lengauer, T. Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des* **14**(3), 215–232, Mar (2000). 28

[73] Bajorath, J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **1**(11), 882–894, Nov (2002). 30

[74] Barnard, J. and Downs, G. Chemical fragment generation and clustering software. *Journal of Chemical Information and Modeling* **37**(1), 141–142, Jan (1997). 30, 32

[75] Daylight-fingerprint. `http://www.daylight.com/dayhtml/doc/theory/theory.finger.html`. Daylight Chemical Information Systems, Inc., Laguna Niguel, CA 92677. 30, 32

[76] McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.-F., and Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* **47**(4), 1504–1519 (2007). 30

[77] Krüger, D. M. and Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* **5**(1), 148–158, Jan (2010). 30

[78] Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50**(5), 742–754, May (2010). 30

[79] Schneider, Neidhart, Giller, and Schmid. "scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew Chem Int Ed Engl* **38**(19), 2894–2896, Oct (1999). 30

[80] Renner, S., Noeske, T., Parsons, C. G., Schneider, P., Weil, T., and Schneider, G. New allosteric modulators of metabotropic glutamate receptor 5 (mglur5) found by ligand-based virtual screening. *Chembiochem* **6**(4), 620–625, Apr (2005). 31

[81] Renner, S. and Schneider, G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **1**(2), 181–185, Feb (2006). 31

[82] Grant, J. A., Gallardo, M. A., and Pickup, B. T. A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *Journal of Computational Chemistry* **17**(14), 1653–1666, Nov (1996). 31

[83] Rush, 3rd, T. S., Grant, J. A., Mosyak, L., and Nicholls, A. A shape-based 3-d scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* **48**(5), 1489–1495, Mar (2005). 31

[84] Nicholls, A. and Grant, J. A. Molecular shape and electrostatics in the encoding of relevant chemical information. *J Comput Aided Mol Des* **19**(9-10), 661–686 (2005). 31

[85] Nicholls, A., McGaughey, G. B., Sheridan, R. P., Good, A. C., Warren, G., Mathieu, M., Muchmore, S. W., Brown, S. P., Grant, J. A., Haigh, J. A., Nevins, N., Jain, A. N., and Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J Med Chem* **53**(10), 3862–3886, May (2010). 31

[86] Willett, P., Barnard, J., and Downs, G. Chemical similarity searching. *Journal of Chemical Information and Modeling* **38**(6), 983–996, Nov (1998). 32

[87] Yuriev, E. and Ramsland, P. A. Latest developments in molecular docking: 2010-2011 in review. *Journal of Molecular Recognition* **26**(5), 215–239, May (2013). 33

[88] Yuriev, E., Agostino, M., and Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J Mol Recognit* **24**(2), 149–164 (2011). 33

[89] Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J* **14**(1), 133–141, Mar (2012). 33

[90] Rognan, D. *Docking Methods for Virtual Screening: Principles and Recent Advances*, 153–176. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 33

[91] Tuccinardi, T. Docking-based virtual screening: recent developments. *Comb Chem High Throughput Screen* **12**(3), 303–314, Mar (2009). 33

[92] Kroemer, R. T. Structure-based drug design: docking and scoring. *Curr Protein Pept Sci* **8**(4), 312–328, Aug (2007). 33

[93] Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**(2), 269–288, Oct (1982). 34

[94] Henzler, A. M. and Rarey, M. In pursuit of fully flexible protein-ligand docking: Modeling the bilateral mechanism of binding. *Molecular Informatics* **29**(3), 164–173 (2010). 35, 60, 204

[95] Henzler, A. M. and Rarey, M. *Protein Flexibility in Structure-Based Virtual Screening: From Models to Algorithms*, 223–244. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 35, 60, 204

[96] Cavasotto, C. N. *Handling Protein Flexibility in Docking and High-Throughput Docking: From Algorithms to Applications*, 245–262. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 35

[97] Volkamer, A., Griewel, A., Grombacher, T., and Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J Chem Inf Model* **50**(11), 2041–2052, Nov (2010). 35, 208

[98] Laurie, A. T. R. and Jackson, R. M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr Protein Pept Sci* **7**(5), 395–406, Oct (2006). 35

[99] Fradera, X. and Mestres, J. Guided docking approaches to structure-based design and screening. *Curr Top Med Chem* **4**(7), 687–700 (2004). 36

[100] Klebe, G. and Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J Comput Aided Mol Des* **8**(5), 583–606, Oct (1994). 36, 54, 55

[101] Sadowski, J. and Boström, J. Mimumba revisited: torsion angle rules for conformer generation derived from x-ray structures. *J Chem Inf Model* **46**(6), 2305–2309 (2006). 36, 55

[102] Dunbrack, Jr, R. and Karplus, M. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol* **230**(2), 543–574, Mar (1993). 36

[103] Lovell, S. C., Word, J. M., Richardson, J. S., and Richardson, D. C. The penultimate rotamer library. *Proteins* **40**(3), 389–408, Aug (2000). 36

[104] Leach, A. R. and Lemon, A. P. Exploring the conformational space of protein side chains using dead-end elimination and the a* algorithm. *Proteins: Structure, Function, and Bioinformatics* **33**(2), 227–239 (1998). 36

[105] Lin, J.-H., Perryman, A. L., Schames, J. R., and McCammon, J. A. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers* **68**(1), 47–62, Jan (2003). 36

[106] Amaro, R. E., Baron, R., and McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des* **22**(9), 693–705, Sep (2008). 36

[107] Knegtel, R. M., Kuntz, I. D., and Oshiro, C. M. Molecular docking to ensembles of protein structures. *J Mol Biol* **266**(2), 424–440, Feb (1997). 36

[108] Claussen, H., Buning, C., Rarey, M., and Lengauer, T. Flexe: efficient molecular docking considering protein structure variations. *J Mol Biol* **308**(2), 377–395, Apr (2001). 36

[109] Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J., and Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins* **46**(1), 34–40, Jan (2002). 36

[110] Schapira, M., Abagyan, R., and Totrov, M. Nuclear hormone receptor targeted virtual screening. *J Med Chem* **46**(14), 3045–3059, Jul (2003). 36

[111] Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W., and Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J Mol Biol* **337**(5), 1161–1182, Apr (2004). 36

[112] Sotriffer, C. A. and Dramburg, I. ïn situ cross-docking"to simultaneously address multiple targets. *J Med Chem* **48**(9), 3122–3125, May (2005). 36

[113] Zavodszky, M. I. and Kuhn, L. A. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci* **14**(4), 1104–1114, Apr (2005). 37

[114] Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **49**(2), 534–553, Jan (2006). 37

[115] Mizutani, M. Y., Takamatsu, Y., Ichinose, T., Nakamura, K., and Itai, A. Effective handling of induced-fit motion in flexible docking. *Proteins* **63**(4), 878–891, Jun (2006). 37

[116] Cavasotto, C. N., Kovacs, J. A., and Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc* **127**(26), 9632–9640, Jul (2005). 37

[117] Rueda, M., Bottegoni, G., and Abagyan, R. Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J Chem Inf Model* **49**(3), 716–725, Mar (2009). 37

[118] DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J Med Chem* **31**(4), 722–729, Apr (1988). 37

[119] Kearsley, S., Underwood, D., Sheridan, R., and Miller, M. Flexibases: A way to enhance the use of molecular docking methods. *Journal of Computer-Aided Molecular Design* **8**(5), 565–582 (1994). 37

[120] McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., and Brown, F. K. Gaussian docking functions. *Biopolymers* **68**(1), 76–90, Jan (2003). 37

[121] Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**(3), 470–489, Aug (1996). 38, 40, 87

[122] Welch, W., Ruppert, J., and Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol* **3**(6), 449–462, Jun (1996). 38

[123] Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15**(5), 411–428, May (2001). 38

[124] Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **46**(4), 499–511, Feb (2003). 38

[125] Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., and Johnson, A. P. ehits: a new fast, exhaustive flexible ligand docking system. *J Mol Graph Model* **26**(1), 198–212, Jul (2007). 38, 60

[126] Abagyan, R., Totrov, M., and Kuznetsov, D. Icm — a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry* **15**(5), 488–506, May (1994). 39

[127] McMartin, C. and Bohacek, R. S. Qxp: powerful, rapid computer algorithms for structure-based drug design. *J Comput Aided Mol Des* **11**(4), 333–344, Jul (1997). 39

[128] Morris, G. M., Goodsell, D. S., Huey, R., and Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of autodock 2.4. *J Comput Aided Mol Des* **10**(4), 293–304, Aug (1996). 39, 56

[129] Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19**(14), 1639–1662 (1998). 39, 56

[130] Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**(3), 727–748, Apr (1997). 39, 42, 56

[131] Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., and Eldridge, M. D. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins* **33**(3), 367–382, Nov (1998). 39

[132] Korb, O., Stützle, T., and Exner, T. E. Plants: Application of ant colony optimization to structure-based drug design. In Ant Colony Optimization and Swarm Intelligence, Dorigo, M., Gambardella, L., Birattari, M., Martinoli, A., Poli, R., and Stützle, T., editors, volume 4150 of *Lecture Notes in Computer Science*, 247–258. Springer Berlin Heidelberg (2006). 39, 56

[133] Korb, O., Stützle, T., and Exner, T. E. An ant colony optimization approach to flexible protein-ligand docking. *Swarm Intelligence* **1**(2), 115–134, Nov (2007). 39, 56

[134] Korb, O., Stützle, T., and Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with plants. *J Chem Inf Model* **49**(1), 84–96, Jan (2009). 39, 41, 56

[135] Chen, H.-M., Liu, B.-F., Huang, H.-L., Hwang, S.-F., and Ho, S.-Y. Sodock: swarm optimization for highly flexible protein-ligand docking. *J Comput Chem* **28**(2), 612–623, Jan (2007). 39

[136] Meier, R., Pippel, M., Brandt, F., Sippl, W., and Baldauf, C. Paradocks: a framework for molecular docking with population-based metaheuristics. *J Chem Inf Model* **50**(5), 879–889, May (2010). 39

[137] Liu, Y., Zhao, L., Li, W., Zhao, D., Song, M., and Yang, Y. Fipsdock: a new molecular docking technique driven by fully informed swarm optimization algorithm. *J Comput Chem* **34**(1), 67–75, Jan (2013). 39

[138] Alonso, H., Bliznyuk, A. A., and Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med Res Rev* **26**(5), 531–568, Sep (2006). 39

[139] Floriano, W. B., Vaidehi, N., Zamanakos, G., and Goddard, 3rd, W. A. Hiervls hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J Med Chem* **47**(1), 56–71, Jan (2004). 40, 42

[140] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J Med Chem* **47**(7), 1739–1749, Mar (2004). 40, 42, 56

[141] Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J Med Chem* **47**(7), 1750–1759, Mar (2004). 40, 42, 56

[142] Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **8**(3), 243–256, Jun (1994). 40

[143] Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., and Freer, S. T. Molecular recognition of the inhibitor ag-1343 by hiv-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biol* **2**(5), 317–324, May (1995). 41

[144] Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Freer, S. T., and Rose, P. W. Monte carlo simulations of the peptide recognition at the consensus binding site of the constant fragment of human immunoglobulin g: the energy landscape analysis of a hot spot at the intermolecular interface. *Proteins* **48**(3), 539–557, Aug (2002). 41

[145] Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Freer, S. T., and Rose, P. W. Computational detection of the binding-site hot spot at the remodeled human growth hormone-receptor interface. *Proteins* **53**(2), 201–219, Nov (2003). 41

[146] Verkhivker, G. M. Computational analysis of ligand binding dynamics at the intermolecular hot spots with the aid of simulated tempering and binding free energy calculations. *J Mol Graph Model* **22**(5), 335–348, May (2004). 41

[147] Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. 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**(5), 425–445, Sep (1997). 41

[148] Murray, C. W., Auton, T. R., and Eldridge, M. D. Empirical scoring functions. ii. the testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of bayesian regression to improve the quality of the model. *J Comput Aided Mol Des* **12**(5), 503–519, Sep (1998). 41

[149] Gohlke, H., Hendlich, M., and Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **295**(2), 337–356, Jan (2000). 41

[150] Gohlke, H., Hendlich, M., and Klebe, G. Predicting binding modes, binding affinities and hot spots for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design* **20**(1), 115–144 (2000). 41

[151] Muegge, I. and Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **42**(5), 791–804, Mar (1999). 41

[152] Muegge, I., Martin, Y. C., Hajduk, P. J., and Fesik, S. W. Evaluation of pmf scoring in docking weak ligands to the fk506 binding protein. *J Med Chem* **42**(14), 2498–2503, Jul (1999). 41

[153] Muegge, I. Pmf scoring revisited. *J Med Chem* **49**(20), 5895–5902, Oct (2006). 41

[154] Weiner, P. K. and Kollman, P. A. Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **2**(3), 287–303 (1981). 42

[155] MacKerell, A. D., Bashford, D., Bellott, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The Journal of Physical Chemistry B* **102**(18), 3586–3616 (1998). 42

[156] Jones, G., Willett, P., and Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol* **245**(1), 43–53, Jan (1995). 42, 56

[157] Huey, R., Morris, G. M., Olson, A. J., and Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**(6), 1145–1152, Apr (2007). 42

[158] Skone, G., Voiculescu, I., and Cameron, S. Knowing when to give up: early-rejection stratagems in ligand docking. *J Comput Aided Mol Des* **23**(10), 715–724, Oct (2009). 43

[159] Charifson, P. S., Corkery, J. J., Murcko, M. A., and Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**(25), 5100–5109, Dec (1999). 43

[160] Raymond, J. W., Jalaie, M., and Bradley, M. P. Conditional probability: a new fusion method for merging disparate virtual screening results. *J Chem Inf Comput Sci* **44**(2), 601–609 (2004). 43

[161] Baber, J. C., Shirley, W. A., Gao, Y., and Feher, M. The use of consensus scoring in ligand-based virtual screening. *J Chem Inf Model* **46**(1), 277–288 (2006). 43

[162] Whittle, M., Gillet, V. J., Willett, P., and Loesel, J. Analysis of data fusion methods in virtual screening: similarity and group fusion. *J Chem Inf Model* **46**(6), 2206–2219 (2006). 43

[163] Whittle, M., Gillet, V. J., Willett, P., and Loesel, J. Analysis of data fusion methods in virtual screening: theoretical model. *J Chem Inf Model* **46**(6), 2193–2205 (2006). 43

[164] Tan, L., Geppert, H., Sisay, M. T., Gütschow, M., and Bajorath, J. Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **3**(10), 1566–1571, Oct (2008). 43

[165] Svensson, F., Karlén, A., and Sköld, C. Virtual screening data fusion using both structure- and ligand-based methods. *J Chem Inf Model* **52**(1), 225–232, Jan (2012). 43

[166] Sastry, G. M., Inakollu, V. S. S., and Sherman, W. Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimensional fingerprints, shape, and docking. *J Chem Inf Model* **53**(7), 1531–1542, Jul (2013). 43

[167] Discovery studio. `http://accelrys.com`. Accelrys, Inc., San Diego, USA. 44

[168] Small-molecule drug discovery suite. `http://www.schrodinger.com/smdd`. Schrödinger LLC, New York, USA. 44

[169] Moe. http://www.chemcomp.com. Chemical Computing Group, Montreal, Canada. 44, 53, 54, 55

[170] Ligandscout. `http://www.inteligand.com`. Inte:Ligand Software-Entwicklungs und Consulting GmbH, Wien, Österreich. 44

[171] Wallach, I. Pharmacophore inference and its application to computational drug discovery. *Drug Development Research* **72**(1), 17–25 (2011). 44

[172] Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* **15**(11), 444–450 (2010). 44, 52

[173] Langer, T. Pharmacophores in drug research. *Molecular Informatics* **29**(6-7), 470–475 (2010). 44

[174] Leach, A. R., Gillet, V. J., Lewis, R. A., and Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J Med Chem* **53**(2), 539–558, Jan (2010). 44, 46, 47, 48, 58

[175] Dror, O., Shulman-Peleg, A., Nussinov, R., and Wolfson, H. Predicting molecular interactions in silico: I. a guide to pharmacophore identification and its applications to drug design. *Current Medicinal Chemistry* **11**(1), 71–90, Jan (2004). 44

[176] Markt, P., Schuster, D., and Langer, T. *Pharmacophore Models for Virtual Screening*, 115–152. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 44, 53

[177] Laggner, C., Wolber, G., Kirchmair, J., Schuster, D., and Langer, T. *Chemoinformatics Approaches to Virtual Screening*, chapter Pharmacophore-based Virtual Screening in Drug Discovery, 76–119. Royal Society of Chemistry (2008). 44, 53

[178] Güner, O. F., editor. *Pharmacophore perception, development, and use in drug design*. International University Line, La Jolla, CA, (2000). 44

[179] Wolber, G. and Langer, T. Ligandscout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model* **45**(1), 160–169 (2005). 45, 51

[180] Wolber, G., Seidel, T., Bendix, F., and Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today* **13**(1-2), 23–29, Jan (2008). 45, 46, 47

[181] Spitzer, G. M., Heiss, M., Mangold, M., Markt, P., Kirchmair, J., Wolber, G., and Liedl, K. R. One concept, three implementations of 3d pharmacophore-based virtual screening: distinct coverage of chemical search space. *J Chem Inf Model* **50**(7), 1241–1247, Jul (2010). 45, 46

[182] Sanders, M. P., Barbosa, A. J., Zarzycka, B., Nicolaes, G. A., Klomp, J. P., de Vlieg, J., and Del Rio, A. Comparative analysis of pharmacophore screening tools. *Journal of chemical information and modeling* **52**(6), 1607–1620 (2012). 45

[183] Dixon, S. L., Smondyrev, A. M., Knoll, E. H., Rao, S. N., Shaw, D. E., and Friesner, R. A. Phase: a new engine for pharmacophore perception, 3d qsar model development, and 3d database screening: 1. methodology and preliminary results. *J Comput Aided Mol Des* **20**(10-11), 647–671 (2006). 45, 49, 51, 53

[184] Greene, J., Kahn, S., Savoj, H., Sprague, P., and Teig, S. Chemical function queries for 3d database search. *Journal of Chemical Information and Computer Sciences* **34**(6), 1297–1308 (1994). 47

[185] Todorov, N. P., Alberts, I. L., de Esch, I. J. P., and Dean, P. M. Quasi: a novel method for simultaneous superposition of multiple flexible ligands and virtual screening using partial similarity. *J Chem Inf Model* **47**(3), 1007–1020 (2007). 47

[186] Cheeseright, T. J., Mackey, M. D., Melville, J. L., and Vinter, J. G. Fieldscreen: virtual screening using molecular fields. application to the dud data set. *J Chem Inf Model* **48**(11), 2108–2117, Nov (2008). 47

[187] Cross, S. and Cruciani, G. Molecular fields in drug discovery: getting old or reaching maturity? *Drug Discov Today* **15**(1-2), 23–32, Jan (2010). 47

[188] Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **28**(7), 849–857, Jul (1985). 47, 51

[189] Vinter, J. G. Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. *J Comput Aided Mol Des* **8**(6), 653–668, Dec (1994). 47

[190] Pastor, M., Cruciani, G., McLay, I., Pickett, S., and Clementi, S. Grid-independent descriptors (grind): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* **43**(17), 3233–3243, Aug (2000). 47

[191] Good, A. C., Hodgkin, E. E., and Richards, W. G. Utilization of gaussian functions for the rapid evaluation of molecular similarity. *Journal of chemical information and computer sciences* **32**(3), 188–191 (1992). 48

[192] Taminau, J., Thijs, G., and De Winter, H. Pharao: pharmacophore alignment and optimization. *J Mol Graph Model* **27**(2), 161–169, Sep (2008). 48, 53

[193] Barnum, D., Greene, J., Smellie, A., and Sprague, P. Identification of common functional configurations among molecules. *J Chem Inf Comput Sci* **36**(3), 563–571 (1996). 48, 49

[194] Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., and Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. fingerprints for ligands and proteins (flap): theory and application. *J Chem Inf Model* **47**(2), 279–294 (2007). 48, 52, 56

[195] Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., and Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J Med Chem* **42**(17), 3251–3264, Aug (1999). 48

[196] Hurst, T. Flexible 3d searching: The directed tweak technique. *Journal of Chemical Information and Computer Sciences* **34**(1), 190–196 (1994). 49

[197] Martin, Y. C., Bures, M. G., Danaher, E. A., DeLazzer, J., Lico, I., and Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J Comput Aided Mol Des* **7**(1), 83–102, Feb (1993). 49

[198] Martin, Y. C. *Pharmacophore Perception, Development, and Use in Drug Design.*, chapter DISCO: what we did right and what we missed., 51–66. International University Line: La Jolla, CA (2000). 49

[199] Wolber, G., Dornhofer, A. A., and Langer, T. Efficient overlay of small organic molecules using 3d pharmacophores. *J Comput Aided Mol Des* **20**(12), 773–788, Dec (2006). 49, 53

[200] Smellie, A., Teig, S. L., and Towbin, P. Poling: promoting conformational variation. *Journal of Computational Chemistry* **16**(2), 171–187 (1995). 49, 55

[201] Jones, G., Willett, P., and Glen, R. C. *Pharmacophore Perception, Development, and Use in Drug Design*, chapter GASP: genetic algorithm superimposition program, 85–106. International University Line: La Jolla, CA (2000). 50

[202] Richmond, N. J., Abrams, C. A., Wolohan, P. R. N., Abrahamian, E., Willett, P., and Clark, R. D. Galahad: 1. pharmacophore identification by hypermolecular alignment of ligands in 3d. *J Comput Aided Mol Des* **20**(9), 567–587, Sep (2006). 50

[203] Sanders, M. P., McGuire, R., Roumen, L., de Esch, I. J., de Vlieg, J., Klomp, J. P., and de Graaf, C. From the protein's perspective: the benefits and challenges of protein structure-based pharmacophore modeling. *MedChemComm* **3**(1), 28–38 (2012). 50, 52

[204] Löwer, M. and Proschak, E. Structure-based pharmacophores for virtual screening. *Molecular Informatics* **30**(5), 398–404 (2011). 50

[205] Wang, R., Gao, Y., and Lai, L. Ligbuilder: a multi-purpose program for structure-based drug design. *Molecular modeling annual* **6**(7-8), 498–516 (2000). 50

[206] Böhm, H. J. Ludi: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aid Mol Des* **6**(6), 593–606, Dec (1992). 51, 87

[207] Böhm, H. J. The computer program ludi: a new method for the de novo design of enzyme inhibitors. *J Comput Aid Mol Des* **6**(1), 61–78, Feb (1992). 51, 87

[208] Kirchhoff, P. D., Brown, R., Kahn, S., Waldman, M., and Venkatachalam, C. Application of structure-based focusing to the estrogen receptor. *Journal of Computational Chemistry* **22**(10), 993–1003 (2001). 51

[209] Ortuso, F., Langer, T., and Alcaro, S. Gbpm: Grid-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics* **22**(12), 1449–1455, Jun (2006). 51

[210] Tintori, C., Corradi, V., Magnani, M., Manetti, F., and Botta, M. Targets looking for drugs: a multistep computational protocol for the development of structure-based pharmacophores and their applications for hit discovery. *J Chem Inf Model* **48**(11), 2166–2179, Nov (2008). 51, 58

[211] Cross, S., Baroni, M., Goracci, L., and Cruciani, G. Grid-based three-dimensional pharmacophores i: Flappharm, a novel approach for pharmacophore elucidation. *J Chem Inf Model* **52**(10), 2587–2598, Oct (2012). 51, 52

[212] Carlson, H. A., Masukawa, K. M., Rubins, K., Bushman, F. D., Jorgensen, W. L., Lins, R. D., Briggs, J. M., and McCammon, J. A. Developing a dynamic pharmacophore model for hiv-1 integrase. *J Med Chem* **43**(11), 2100–2114, Jun (2000). 51

[213] Miranker, A. and Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **11**(1), 29–34 (1991). 51

[214] Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., and Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**(21), 6177–6196, Oct (2006). 51

[215] Loving, K., Salam, N. K., and Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J Comput Aided Mol Des* **23**(8), 541–554, Aug (2009). 51

[216] Salam, N. K., Nuti, R., and Sherman, W. Novel method for generating structure-based pharmacophores using energetic analysis. *J Chem Inf Model* **49**(10), 2356–2368, Oct (2009). 51

[217] Barillari, C., Marcou, G., and Rognan, D. Hotspots-guided receptor-based pharmacophores (hspharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *J Chem Inf Model* **48**(7), 1396–1410, Jul (2008). 51

[218] Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., and Rognan, D. sc-pdb: an annotated database of druggable binding sites from the protein data bank. *J Chem Inf Model* **46**(2), 717–727 (2006). 51

[219] Meslamani, J., Rognan, D., and Kellenberger, E. sc-pdb: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **27**(9), 1324–1326, May (2011). 51

[220] Chen, J. and Lai, L. Pocket v.2: further developments on receptor-based pharmacophore modeling. *J Chem Inf Model* **46**(6), 2684–2691 (2006). 51

[221] McGregor, M. J. A pharmacophore map of small molecule protein kinase inhibitors. *J Chem Inf Model* **47**(6), 2374–2382 (2007). 52

[222] Deng, Z., Chuaqui, C., and Singh, J. Structural interaction fingerprint (sift): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem* **47**(2), 337–344, Jan (2004). 52

[223] Chuaqui, C., Deng, Z., and Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J Med Chem* **48**(1), 121–133, Jan (2005). 52

[224] Mordalski, S., Kosciolek, T., Kristiansen, K., Sylte, I., and Bojarski, A. J. Protein binding site analysis by means of structural interaction fingerprint patterns. *Bioorg Med Chem Lett* **21**(22), 6816–6819, Nov (2011). 52

[225] Kurczab, R. and Bojarski, A. J. New strategy for receptor-based pharmacophore query construction: A case study for 5-ht7 receptor ligands. *J Chem Inf Model* **0**, 0–0, Nov (2013). 52

[226] Cross, S., Baroni, M., Carosati, E., Benedetti, P., and Clementi, S. Flap: Grid molecular interaction fields in virtual screening. validation using the dud data set. *J Chem Inf Model* **50**(8), 1442–1450, Aug (2010). 52

[227] Oellien, F., Cramer, J., Beyer, C., Ihlenfeldt, W.-D., and Selzer, P. M. The impact of tautomer forms on pharmacophore-based virtual screening. *J Chem Inf Model* **46**(6), 2342–2354 (2006). 52, 55

[228] Hu, B. and Lill, M. A. Exploring the potential of protein-based pharmacophore models in ligand pose prediction and ranking. *J Chem Inf Model* **53**(5), 1179–1190, May (2013). 52

[229] Perola, E. and Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* **47**(10), 2499–2510, May (2004). 52, 112

[230] Takagi, T., Amano, M., and Tomimoto, M. Novel method for the evaluation of 3d conformation generators. *J Chem Inf Model* **49**(6), 1377–1388, Jun (2009). 52

[231] Braga, R. C. and Andrade, C. H. Assessing the performance of 3d pharmacophore models in virtual screening: how good are they? *Curr Top Med Chem* **13**(9), 1127–1138 (2013). 53

[232] Triballeau, N., Bertrand, H.-O., and Acher, F. *Pharmacophores and Pharmacophore Searches*, chapter Are You Sure You Have a Good Model?, 325–364. Methods and Principles in Medicinal Chemistry. Wiley Blackwell (John Wiley & Sons) (2006). 53

[233] Hecker, E. A., Duraiswami, C., Andrea, T. A., and Diller, D. J. Use of catalyst pharmacophore models for screening of large combinatorial libraries. *J Chem Inf Comput Sci* **42**(5), 1204–1211 (2002). 53

[234] Kurogi, Y. and Güner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr Med Chem* **8**(9), 1035–1055, Jul (2001). 53

[235] Güner, O., Clement, O., and Kurogi, Y. Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. *Curr Med Chem* **11**(22), 2991–3005, Nov (2004). 53

[236] Unity. `http://www.tripos.com`. Certara USA, Inc., St. Lois, MO 63101 USA. 53

[237] Koes, D. R. and Camacho, C. J. Pharmer: efficient and exact pharmacophore search. *J Chem Inf Model* **51**(6), 1307–1314, Jun (2011). 53

[238] Oprea, T. I. and Matter, H. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* **8**(4), 349–358, Aug (2004). 54

[239] Martin, Y. C. Let's not forget tautomers. *J Comput Aided Mol Des* **23**(10), 693–704, Oct (2009). 54, 56, 156, 193, 194, 203, 206

[240] Leach, A. *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall, (2001). 54

[241] Schwab, C. H. *Conformational Analysis and Searching*, 262–301. Wiley-VCH Verlag GmbH (2008). 54

[242] Smellie, A., Stanton, R., Henne, R., and Teig, S. Conformational analysis by intersection: Conan. *J Comput Chem* **24**(1), 10–20, Jan (2003). 54

[243] Bruccoleri, R. E. and Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26**(1), 137–168, Jan (1987). 54

[244] Gippert, G. P., Wright, P. E., and Case, D. A. Distributed torsion angle grid search in high dimensions: a systematic approach to nmr structure determination. *J Biomol NMR* **11**(3), 241–263, Apr (1998). 54

[245] Li, J., Ehlers, T., Sutter, J., Varma-O'brien, S., and Kirchmair, J. Caesar: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J Chem Inf Model* **47**(5), 1923–1932 (2007). 54

[246] O'Boyle, N. M., Vandermeersch, T., Flynn, C. J., Maguire, A. R., and Hutchison, G. R. Confab - systematic generation of diverse low-energy conformers. *J Cheminform* **3**, 8 (2011). 54

[247] Mohamadi, F., Richards, N. G. J., Guida, W. C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T., and Still, W. C. Macromodel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *Journal of Computational Chemistry* **11**(4), 440–467 (1990). 54, 55

[248] Rusinko, A., Sheridan, R. P., Nilakantan, R., Haraki, K. S., Bauman, N., and Venkataraghavan, R. Using concord to construct a large database of three-dimensional coordinates from connection tables. *Journal of Chemical Information and Computer Sciences* **29**(4), 251–255 (1989). 54, 55

[249] Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. Conformer generation with omega: algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J Chem Inf Model* **50**(4), 572–584, Apr (2010). 54, 55

[250] Watts, K. S., Dalal, P., Murphy, R. B., Sherman, W., Friesner, R. A., and Shelley, J. C. Confgen: a conformational search method for efficient generation of bioactive conformers. *J Chem Inf Model* **50**(4), 534–546, Apr (2010). 55

[251] Miteva, M. A., Guyon, F., and Tufféry, P. Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic Acids Res* **38**(Web Server issue), W622–W627, Jul (2010). 55

[252] Renner, S., Schwab, C. H., Gasteiger, J., and Schneider, G. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J Chem Inf Model* **46**(6), 2324–2332 (2006). 55

[253] Griewel, A., Kayser, O., Schlosser, J., and Rarey, M. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. *J Chem Inf Model* **49**(10), 2303–2311, Oct (2009). 55, 77, 92

[254] Schärfer, C., Schulz-Gasch, T., Hert, J., Heinzerling, L., Schulz, B., Inhester, T., Stahl, M., and Rarey, M. Confect: Conformations from an expert collection of torsion patterns. *Chemmedchem* **8**(10), 1690–1700, Aug (2013). 55, 77, 81, 245, 252

[255] Schärfer, C., Schulz-Gasch, T., Ehrlich, H.-C., Guba, W., Rarey, M., and Stahl, M. Torsion angle preferences in druglike chemical space: a comprehensive guide. *J Med Chem* **56**(5), 2016–2028, Mar (2013). 55, 63, 77, 78

[256] Chandrasekhar, J., Saunders, M., and Jorgensen, W. L. Efficient exploration of conformational space using the stochastic search method: application to β-peptide oligomers. *Journal of Computational Chemistry* **22**(14), 1646–1654 (2001). 55

[257] Saunders, M. Stochastic search for the conformations of bicyclic hydrocarbons. *Journal of Computational Chemistry* **10**(2), 203–208 (1989). 55

[258] McGarrah, D. and Judson, R. Analysis of the genetic algorithm method of molecular conformation determination. *Journal of Computational Chemistry* **14**(11), 1385–1395 (1993). 55

[259] Judson, R., Jaeger, E., Treasurywala, A., and Peterson, M. Conformational searching methods for small molecules. ii. genetic algorithm approach. *Journal of Computational Chemistry* **14**(11), 1407–1414 (1993). 55

[260] Glen, R. C. and Payne, A. W. A genetic algorithm for the automated generation of molecules within constraints. *J Comput Aided Mol Des* **9**(2), 181–202, Apr (1995). 55

[261] Vainio, M. J. and Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model* **47**(6), 2462–2474 (2007). 55

[262] Liu, X., Bai, F., Ouyang, S., Wang, X., Li, H., and Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics* **10**, 101 (2009). 55

[263] Leach, A. R. and Smellie, A. S. A combined model-building and distance-geometry approach to automated conformational analysis and search. *Journal of Chemical Information and Computer Sciences* **32**(4), 379–385 (1992). 55

[264] Crippen, G. M. Exploring the conformation space of cycloalkanes by linearized embedding. *Journal of Computational Chemistry* **13**(3), 351–361 (1992). 55

[265] Peishoff, C. E. and Dixon, J. S. Improvements to the distance geometry algorithm for conformational sampling of cyclic structures. *Journal of Computational Chemistry* **13**(5), 565–569 (1992). 55

[266] Havel, T. F. Distance geometry: Theory, algorithms, and chemical applications. *Encyclopedia of Computational Chemistry* **120**, 1–19 (1998). 55

[267] Chen, J., Im, W., and Brooks, 3rd, C. L. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J Comput Chem* **26**(15), 1565–1578, Nov (2005). 55

[268] Sun, Y. and Kollman, P. A. Conformational sampling and ensemble generation by molecular dynamics simulations: 18-crown-6 as a test case. *Journal of Computational Chemistry* **13**(1), 33–40 (1992). 55

[269] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science* **220**(4598), 671–680 (1983). 55

[270] Wilson, S. R. and Cui, W. Applications of simulated annealing to peptides. *Biopolymers* **29**(1), 225–235 (1990). 55

[271] ten Brink, T. and Exner, T. E. pk(a) based protonation states and microspecies for protein-ligand docking. *J Comput Aided Mol Des* **24**(11), 935–942, Nov (2010). 55

[272] Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. Epik: a software program for pk(a) prediction and protonation state generation for drug-like molecules. *J Comput Aided Mol Des* **21**(12), 681–691, Dec (2007). 55

[273] Milletti, F., Storchi, L., Sforna, G., Cross, S., and Cruciani, G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J Chem Inf Model* **49**(1), 68–75, Jan (2009). 55

[274] ChemAxon Kft., Budapest, Hungary. *Marvin User's Guide, Version 5.11.4*, (2012). 55

[275] OpenEye Scientific Software, Santa Fe, NM. *QUACPAC, v1.5.0*, (2012). 55

[276] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. Zinc: A free tool to discover chemistry for biology. *J Chem Inf Model* **52**(7), 1757–1768, Jun (2012). 55, 151, 152

[277] Irwin, J. J. and Shoichet, B. K. Zinc–a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**(1), 177–182 (2005). 55, 151

[278] Trepalin, S. V., Skorenko, A. V., Balakin, K. V., Nasonov, A. F., Lang, S. A., Ivashchenko, A. A., and Savchuk, N. P. Advanced exact structure searching in large databases of chemical compounds. *J Chem Inf Comput Sci* **43**(3), 852–860 (2003). 55

[279] Kenny, P. W. and Sadowski, J. *Structure Modification in Chemical Databases*, 271–285. Wiley-VCH Verlag GmbH & Co. KGaA (2005). 55

[280] Todorov, N. P., Monthoux, P. H., and Alberts, I. L. The influence of variations of ligand protonation and tautomerism on protein-ligand recognition and binding energy landscape. *J Chem Inf Model* **46**(3), 1134–1142 (2006). 55, 56

[281] Deng, W. and Verlinde, C. L. M. J. Evaluation of different virtual screening programs for docking in a charged binding pocket. *J Chem Inf Model* **48**(10), 2010–2020, Oct (2008). 55

[282] ten Brink, T. and Exner, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J Chem Inf Model* **49**(6), 1535–1546, Jun (2009). 55, 56

[283] Kalliokoski, T., Salo, H. S., Lahtela-Kakkonen, M., and Poso, A. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *J Chem Inf Model* **49**(12), 2742–2748, Dec (2009). 55, 56

[284] Milletti, F. and Vulpetti, A. Tautomer preference in pdb complexes and its impact on structure-based drug discovery. *J Chem Inf Model* **50**(6), 1062–1074, Jun (2010). 55, 56

[285] Park, M.-S., Gao, C., and Stern, H. A. Estimating binding affinities by docking/scoring methods using variable protonation states. *Proteins: Structure, Function, and Bioinformatics* **79**(1), 304–314, Jan (2011). 55

[286] Huang, N., Shoichet, B. K., and Irwin, J. J. Benchmarking sets for molecular docking. *J Med Chem* **49**(23), 6789–6801, Nov (2006). 56, 150

[287] Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., Felts, A. K., Halgren, T. A., Mainz, D. T., Maple, J. R., Murphy, R., Philipp, D. M., Repasky, M. P., Zhang, L. Y., Berne, B. J., Friesner, R. A., Gallicchio, E., and Levy, R. M. Integrated modeling program, applied chemical theory (impact). *J Comput Chem* **26**(16), 1752–1780, Dec (2005). 56

[288] Pospisil, P., Ballmer, P., Scapozza, L., and Folkers, G. Tautomerism in computer-aided drug design. *J Recept Signal Transduct Res* **23**(4), 361–371 (2003). 56

[289] Greenwood, J., Calkins, D., Sullivan, A., and Shelley, J. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design* **24**(6-7), 591–604 (2010). 56, 156, 193, 194, 203, 206

[290] Cole, J. C., Korb, O., Olsson, T. S. G., and Liebeschuetz, J. *The Basis for Target-Based Virtual Screening: Protein Structures*, 87–114. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 56

[291] Brünger, A. T. and Karplus, M. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins: Structure, Function, and Bioinformatics* **4**(2), 148–156 (1988). 57

[292] Bass, M. B., Hopkins, D. F., Jaquysh, W. A. N., and Ornstein, R. L. A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins: Structure, Function, and Bioinformatics* **12**(3), 266–277 (1992). 57

[293] Hooft, R. W., Sander, C., and Vriend, G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Structure, Function, and Bioinformatics* **26**(4), 363–376 (1996). 57

[294] Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology* **285**(4), 1735–1747 (1999). 57

[295] Li, X., Jacobson, M. P., Zhu, K., Zhao, S., and Friesner, R. A. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. *Proteins: Structure, Function, and Bioinformatics* **66**(4), 824–837 (2007). 57

[296] Bayden, A. S., Fornabaio, M., Scarsdale, J. N., and Kellogg, G. E. Web application for studying the free energy of binding and protonation states of protein–ligand complexes based on hint. *Journal of computer-aided molecular design* **23**(9), 621–632 (2009). 57

[297] Lippert, T. and Rarey, M. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *J Cheminform* **1**(1), 13 (2009). 57, 74

[298] Labute, P. Protonate3d: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **75**(1), 187–205, Apr (2009). 57

[299] Krieger, E., Dunbrack Jr, R. L., Hooft, R. W., and Krieger, B. Assignment of protonation states in proteins and ligands: Combining pka prediction with hydrogen bonding network optimization. In *Computational Drug Discovery and Design*, 405–421. Springer, New York (2012). 57

[300] Forrest, L. R. and Honig, B. An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. *Proteins: Structure, Function, and Bioinformatics* **61**(2), 296–309 (2005). 57

[301] Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **27**(3), 221–234, Mar (2013). 57

[302] Sotriffer, C. and Matter, H. *The Challenge of Affinity Prediction: Scoring Functions for Structure-Based Virtual Screening*, 177–221. Wiley-VCH Verlag GmbH & Co. KGaA (2011). 58

[303] Dong, X., Ebalunode, J. O., Yang, S.-Y., and Zheng, W. Receptor-based pharmacophore and pharmacophore key descriptors for virtual screening and qsar modeling. *Curr Comput Aided Drug Des* **7**(3), 181–189, Sep (2011). 58

[304] Muthas, D., Sabnis, Y. A., Lundborg, M., and Karlén, A. Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? an investigation of the advantages and pitfalls of post-filtering. *J Mol Graph Model* **26**(8), 1237–1251, Jun (2008). 58

[305] Peach, M. L. and Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. *J Cheminform* **1**(1), 6 (2009). 58

[306] Yang, J.-M. and Shen, T.-W. A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins* **59**(2), 205–220, May (2005). 59

[307] Fradera, X., Knegtel, R. M., and Mestres, J. Similarity-driven flexible ligand docking. *Proteins* **40**(4), 623–636, Sep (2000). 59

[308] Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T. M., Murray, C. W., Taylor, R. D., and Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J Chem Inf Comput Sci* **44**(3), 793–806 (2004). 59

[309] Hindle, S. A., Rarey, M., Buning, C., and Lengaue, T. Flexible docking under pharmacophore type constraints. *J Comput Aid Mol Des* **16**(2), 129–149, Feb (2002). 59, 94

[310] Kim, M., Nichols, S., Wang, Y., and McCammon, J. Effects of histidine protonation and rotameric states on virtual screening of m. tuberculosis rmlc. *Journal of Computer-Aided Molecular Design* **27**(3), 235–246 (2013). 60

[311] Zsoldos, Z., Reid, D., Simon, A., Sadjad, B. S., and Johnson, A. P. ehits: an innovative approach to the docking and scoring function problems. *Curr Protein Pept Sci* **7**(5), 421–435, Oct (2006). 60

[312] Urbaczek, S., Kolodzik, A., Groth, I., Heuser, S., and Rarey, M. Reading pdb: perception of molecules from 3d atomic coordinates. *J Chem Inf Model* **53**(1), 76–87, Jan (2013). 63, 67, 68

[313] Urbaczek, S., Kolodzik, A., and Rarey, M. The valence state combination model: a generic framework for handling tautomers and protonation states. *J Chem Inf Model* **54**(3), 756–766, Mar (2014). 63, 69, 70, 73, 82, 140

[314] Hilbig, M., Urbaczek, S., Groth, I., Heuser, S., and Rarey, M. Mona - interactive manipulation of molecule collections. *J Cheminform* **5**(1), 38, Aug (2013). 63, 81, 245, 249, 260

[315] Bietz, S., Urbaczek, S., Schulz, B., and Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J Cheminform* **6**, 12 (2014). 63, 74, 245

[316] Schellhammer, I. and Rarey, M. Flexx-scan: fast, structure-based virtual screening. *Proteins* **57**(3), 504–517, Nov (2004). 63, 85, 88, 245

[317] Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K., Grier, D. L., Leland, B. A., and Laufer, J. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J Chem Inf Comp Sci* **32**(3), 244–255 (1992). 64

[318] *TRIPOS Mol2 File Format*. 64

[319] *PDB File Formats*. 64

[320] Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comp Sci* **28**(1), 31–36 (1988). 64

[321] Ash, S., Cline, M. A., Homer, R. W., Hurst, T., and Smith, G. B. Sybyl line notation (sln): A versatile language for chemical structure representation. *J Chem Inf Comp Sci* **37**(1), 71–79 (1997). 64

[322] McNaught, A. The iupac international chemical identifier. *Chemistry international* **28**(6), 12–14 (2006). 64

[323] Lewis, G. N. The atom and the molecule. *J Am Chem Soc* **38**(4), 762–785 (1916). 64

[324] Gillespie, R. J. and Nyholm, R. S. Inorganic stereochemistry. *Q. Rev. Chem. Soc.* **11**, 339–380 (1957). 65

[325] Vismara, P. Union of all the minimum cycle bases of a graph. *Electron J Comb* **4**(1), 73–87 (1997). 67

[326] Hückel, E. Quantentheoretische Beiträge zum Problem der aromatischen und ungesättigten Verbindungen. III. *Zeitschrift fur Physik* **76**, 628–648, September (1932). 67

[327] *SMARTS - A Language for Describing Molecular Patterns.* 77, 85, 94

[328] Allen, F. H. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* **58**(Pt 3 Pt 1), 380–388, Jun (2002). 78

[329] Schulz-Gasch, T., Schärfer, C., Guba, W., and Rarey, M. Tfd: Torsion fingerprints as a new measure to compare small molecule conformations. *J Chem Inf Model* **52**(6), 1499–1512, Jun (2012). 81, 113

[330] Klebe, G. The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. *J Mol Biol* **237**(2), 212–235, Mar (1994). 87

[331] Wu, K. Fastbit: an efficient indexing technology for accelerating data-intensive science. In *Journal of Physics: Conference Series*, volume 16, 556. IOP Publishing, (2005). 95

[332] O'Neil, P. E. Model 204 architecture and performance. In *High Performance Transaction Systems*, 39–59. Springer (1989). 96

[333] Chan, C.-Y. and Ioannidis, Y. E. Bitmap index design and evaluation. *Sigmod Rec* **27**(2), 355–366 (1998). 97

[334] Wu, K., Otoo, E. J., and Shoshani, A. Compressing bitmap indexes for faster search operations. In *Scientific and Statistical Database Management, 2002. Proceedings. 14th International Conference on*, 99–108. IEEE, (2002). 97

[335] Wu, K., Otoo, E. J., Shoshani, A., and Nordberg, H. Notes on design and implementation of compressed bit vectors. Technical report, Technical Report LBNL/PUB-3161, Lawrence Berkeley National Laboratory, Berkeley, CA, (2001). 98

[336] Wu, K., Otoo, E. J., and Shoshani, A. An efficient compression scheme for bitmap indices. Technical report, Technical Report LBNL-49626, Lawrence Berkeley National Laboratory, Berkeley, CA, (2004). 98

[337] Kirchmair, J., Ristic, S., Eder, K., Markt, P., Wolber, G., Laggner, C., and Langer, T. Fast and efficient in silico 3d screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J Chem Inf Model* **47**(6), 2182–2196 (2007). 112, 148

[338] Connolly, M. L. Analytical molecular surface calculation. *Journal of Applied Crystallography* **16**(5), 548–558 (1983). 115

[339] Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans Math Softw* **22**(4), 469–483, December (1996). 115, 116

[340] Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., and Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov Today* **17**(23-24), 1270–1281, Dec (2012). 125

[341] Paul, N. and Rognan, D. Consdock: A new program for the consensus analysis of protein-ligand interactions. *Proteins* **47**(4), 521–533, Jun (2002). 146

[342] Vieth, M., Hirst, J. D., Kolinski, A., and Brooks, C. L. Assessing energy functions for flexible docking. *J Comput Chem* **19**(14), 1612–1622 (1998). 146

[343] Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., and Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **50**(4), 726–741, Feb (2007). 147

[344] Craig, I. R., Essex, J. W., and Spiegel, K. Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. *J Chem Inf Model* **50**(4), 511–524, Apr (2010). 149

[345] Neves, M. A. C., Totrov, M., and Abagyan, R. Docking and scoring with icm: the benchmarking results and strategies for improvement. *J Comput Aid Mol Des* **26**(6), 675–686, Jun (2012). 150, 170

[346] Spitzer, R. and Jain, A. N. Surflex-dock: Docking benchmarks and real-world application. *J Comput Aid Mol Des* **26**(6), 687–699, Jun (2012). 150

[347] Schneider, N., Hindle, S., Lange, G., Klein, R., Albrecht, J., Briem, H., Beyer, K., Claußen, H., Gastreich, M., Lemmen, C., and Rarey, M. Substantial improvements in large-scale redocking and screening using the novel hyde scoring function. *J Comput Aid Mol Des* **26**(6), 701–723, Jun (2012). 150, 157, 170, 171, 197, 233

[348] Novikov, F. N., Stroylov, V. S., Zeifman, A. A., Stroganov, O. V., Kulkov, V., and Chilov, G. G. Lead finder docking and virtual screening evaluation with astex and dud test sets. *J Comput Aid Mol Des* **26**(6), 725–735, Jun (2012). 150, 170

[349] Liebeschuetz, J. W., Cole, J. C., and Korb, O. Pose prediction and virtual screening performance of gold scoring functions in a standardized test. *J Comput Aid Mol Des* **26**(6), 737–748, Jun (2012). 150, 170

[350] Brozell, S. R., Mukherjee, S., Balius, T. E., Roe, D. R., Case, D. A., and Rizzo, R. C. Evaluation of dock 6 as a pose generation and database enrichment tool. *J Comput Aid Mol Des* **26**(6), 749–773, Jun (2012). 150, 170

[351] Corbeil, C. R., Williams, C. I., and Labute, P. Variability in docking success rates due to dataset preparation. *J Comput Aid Mol Des* **26**(6), 775–786, Jun (2012). 150

[352] Repasky, M. P., Murphy, R. B., Banks, J. L., Greenwood, J. R., Tubert-Brohman, I., Bhat, S., and Friesner, R. A. Docking performance of the glide program as evaluated on the astex and dud datasets: a complete set of glide sp results and selected results for a new scoring function integrating watermap and glide. *J Comput Aid Mol Des* **26**(6), 787–799, Jun (2012). 150, 170

[353] Zinc clean leads. `http://zinc.docking.org/subsets/clean-leads`. accessed Dec 7, 2012. 151

[354] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge, 2 edition, (1988). 153

[355] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *J Med Chem* **55**(14), 6582–6594, Jul (2012). 156

[356] Yan, X., Hollis, T., Svinth, M., Day, P., Monzingo, A. F., Milne, G. W., and Robertus, J. D. Structure-based identification of a ricin inhibitor. *J Mol Biol* **266**(5), 1043–1049, Mar (1997). 156, 193

[357] Malamas, M. S. and Hohman, T. C. N-substituted spirosuccinimide, spiropyridazine, spiroazetidine, and acetic acid aldose reductase inhibitors derived from isoquinoline-1,3-diones. 2. *J Med Chem* **37**(13), 2059–2070, Jun (1994). 156

[358] Negoro, T., Murata, M., Ueda, S., Fujitani, B., Ono, Y., Kuromiya, A., Komiya, M., Suzuki, K., and Matsumoto, J. Novel, highly potent aldose reductase inhibitors: (r)-(-)-2-(4-bromo-2-fluorobenzyl)-1,2,3,4- tetrahydropyrrolo[1,2-a]pyrazine -4-spiro-3'-pyrrolidine-1,2',3,5'-tetrone (as-3201) and its congeners. *J Med Chem* **41**(21), 4118–4129, Oct (1998). 156

[359] Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C., Moras, D., Schneider, T. R., Joachimiak, A., and Podjarny, A. Ultrahigh resolution drug design i: details of interactions in human aldose reductase-inhibitor complex at 0.66 a. *Proteins* **55**(4), 792–804, Jun (2004). 156

[360] El-Kabbani, O., Darmanin, C., Schneider, T. R., Hazemann, I., Ruiz, F., Oka, M., Joachimiak, A., Schulze-Briese, C., Tomizaki, T., Mitschler, A., and Podjarny, A. Ultrahigh resolution drug design. ii. atomic resolution structures of human aldose reductase holoenzyme complexed with fidarestat and minalrestat: implications for the binding of cyclic imide inhibitors. *Proteins* **55**(4), 805–813, Jun (2004). 157, 197

[361] Schneider, N., Lange, G., Hindle, S., Klein, R., and Rarey, M. A consistent description of hydrogen bond and dehydration energies in protein-ligand complexes: methods behind the hyde scoring function. *J Comput Aid Mol Des* **27**(1), 15–29, Jan (2013). 157, 197

[362] Dönnecke, D., Schweinitz, A., Stürzebecher, A., Steinmetzer, P., Schuster, M., Stürzebecher, U., Nicklisch, S., Stürzebecher, J., and Steinmetzer, T. From selective substrate analogue factor xa inhibitors to dual inhibitors of thrombin and factor xa. part 3. *Bioorg Med Chem Lett* **17**(12), 3322–3329, Jun (2007). 173

[363] Nar, H., Bauer, M., Schmid, A., Stassen, J. M., Wienen, W., Priepke, H. W., Kauffmann, I. K., Ries, U. J., and Hauel, N. H. Structural basis for inhibition promiscuity of dual specific thrombin and factor xa blood coagulation inhibitors. *Structure* **9**(1), 29–37, Jan (2001). 173

[364] Wiley, M. R., Weir, L. C., Briggs, S., Bryan, N. A., Buben, J., Campbell, C., Chirgadze, N. Y., Conrad, R. C., Craft, T. J., Ficorilli, J. V., Franciskovich, J. B., Froelich, L. L., Gifford-Moore, D. S., Goodson, Jr, T., Herron, D. K., Klimkowski, V. J., Kurz, K. D., Kyle, J. A., Masters, J. J., Ratz, A. M., Milot, G., Shuman, R. T., Smith, T., Smith, G. F., Tebbe, A. L., and Tinsley, J. M. Structure-based design of potent, amidine-derived inhibitors of factor xa: evaluation of selectivity, anticoagulant activity, and antithrombotic activity. *J Med Chem* **43**(5), 883–899, Mar (2000). 173

[365] Breault, G. A., Ellston, R. P. A., Green, S., James, S. R., Jewsbury, P. J., Midgley, C. J., Pauptit, R. A., Minshull, C. A., Tucker, J. A., and Pease, J. E. Cyclin-dependent kinase 4 inhibitors as a treatment for cancer. part 2: identification and optimisation of substituted 2,4-bis anilino pyrimidines. *Bioorg Med Chem Lett* **13**(18), 2961–2966, Sep (2003). 173

[366] Gassel, M., Breitenlechner, C. B., König, N., Huber, R., Engh, R. A., and Bossemeyer, D. The protein kinase c inhibitor bisindolyl maleimide 2 binds with reversed orientations to different conformations of protein kinase a. *J Biol Chem* **279**(22), 23679–23690, May (2004). 173

[367] Wildman, S. A. and Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences* **39**(5), 868–873 (1999). 206

[368] Ertl, P., Rohde, B., and Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem* **43**(20), 3714–3717, Oct (2000). 206

[369] Schomburg, K. T., Bietz, S., Briem, H., Henzler, A. M., Urbaczek, S., and Rarey, M. Facing the challenges of structure-based target prediction by inverse virtual screening. *J Chem Inf Model* **54**(6), 1676–1686, Jun (2014). 207

[370] von Behren, M. M., Volkamer, A., Henzler, A. M., Schomburg, K. T., Urbaczek, S., and Rarey, M. Fast protein binding site comparison via an index-based screening technology. *J Chem Inf Model* **53**(2), 411–422, Feb (2013). 208

[371] Lippert, T., Schulz-Gasch, T., Roche, O., Guba, W., and Rarey, M. De novo design by pharmacophore-based searches in fragment spaces. *J Comput Aided Mol Des* **25**(10), 931–945, Oct (2011). 209

[372] Meslamani, J., Li, J., Sutter, J., Stevens, A., Bertrand, H.-O., and Rognan, D. Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *J Chem Inf Model* **52**(4), 943–955, Apr (2012). 210

[373] Henrich, S., Salo-Ahen, O. M. H., Huang, B., Rippmann, F. F., Cruciani, G., and Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit* **23**(2), 209–219 (2010). 210

[374] Henzler, A. M., Urbaczek, S., Hilbig, M., and Rarey, M. An integrated approach to knowledge-driven structure-based virtual screening. *J Comput Aided Mol Des* **28**(9), 927–939, Sep (2014). 245

A Daten und Ausführliche Resultate

A.1 Interventionen am Astex$_{ACS}$-Datensatz

Um die Leistung von cRAISE bei möglichst optimaler Datenlage zu bewerten, wurden die Orginalstrukturen des Astex$_{ACS}$-Datensatz präpariert, adäquate Seitenkettenorientierungen gewählt und die Zustände der Proteine und Startliganden korrigiert. Zudem wurden die unvollständigen Bindetaschen, die die Vorhersage zu drastisch beeinflussten ausgeschlossen. Die Modifizierungen die die 146 Bindetaschen des Astex$_{ACS}$-Datensatz definieren, sind in Tabelle A.1 zusammengefasst.

Tabelle A.1: Modifikationen am Astex$_{ACS}$

Code	Fehler	Korrektur
1gkc	Suboptimaler Ligandzustand	Hydroximate protoniert
1ia1	Korrupter Cofaktor	NDP durch Originalcofaktor ersetzt
1ia1	Suboptimaler Ligandzustand	Pyrimidine protoniert
1j3j	Suboptimaler Ligandzustand	Pyrimidine protoniert
1kzk	Suboptimaler Proteinzustand	Asp25 protoniert
1l2s	Suboptimaler Ligandzustand	Sulfonamidestickstoff protoniert
1mmv	Suboptimaler Proteinzustand	Asp597 protoniert
1n46	Suboptimaler Ligandzustand	Uracil deprotoniert
1r1h	Suboptimaler Proteinzustand	Glu584 protoniert
1r58	Suboptimaler Ligandzustand	Ligand neutralisiert
1s3v	Suboptimaler Ligandzustand	Pyrimidine protoniert
1tz8	Unvollständige Bindetasche	Exklusion der Taschen 2–5
1vcj	Alternative Seitenkette	Austausch von AGlu275 durch BGlu275
1u1c	Unvollständige Bindetasche	Exklusion von Tasche 3
1uml	Suboptimaler Ligandzustand	Imidazol protoniert
1y6b	Suboptimaler Ligandzustand	Pyridine protoniert

Zudem wurden den Strukturen essentielle Wassermoleküle hinzugefügt, um den Ein-

fluss dieser auf das Docking-Ergebnis untersuchen zu können. Die vorgenommenen Korrekturen an den 146 Bindetaschen des $\text{Astex}_{\text{ACS}}$-Datensatzes, die im $\text{Astex}_{\text{h2o}}$-Datensatz resultierten sind in Tabelle A.2 zusammengestellt.

Tabelle A.2: Modifikationen am $\text{Astex}_{\text{revised}}$, um den $\text{Astex}_{\text{h2o}}$-Datensatz zu erhalten

Code	Inklusion essentieller Wassermoleküle
1gm8	O(6754), O(6755), O(6756)
1gpk	O(4592)
1hvy	O(9566), O(9584), O(9664), O(9699), O(9735), O(9747), O(9864), O(9867), O(9891), O(10017), O(10025), O(10038), O(10064)
1l2s	O(5567), O(5676), O(5737)
1l7f	O(3289)
1lpz	O(2502)
1r55	O(1678)
1sj0	O(1005)
1sq5	O(9853), O(9986), O(10117), O(10123), O(10316), O(10320), O(10598), O(10606), O(10610)
1sqn	O(4036), O(4157)
1t9b	O(10196)
1tow	O(1112)
1u4d	O(4259), O(4347)
1uml	O(2940)
1w2g	O(3026), O(3088)
1x8x	O(2616), O(2672)
1xm6	O(5542), O(5686)
1xoq	O(5492), O(5667)
1y6b	O(2214), O(2241), O(2296)
1ygc	O(2493), O(2550)
1yqy	O(4284)
1yv3	O(5818)
1yvf	O(4456), O(4460)

A.2 Beispielhafte Pharmakophordefinition

Ein Beispiel einer cRAISE Pharmakophordefinition ist in Listing A.1 gegeben. Es umfasst alle Merkmalstypen, die von cRAISE unterstützt werden.

Listing A.1: Beispielhafte Pharmakophordefinition

```
<!DOCTYPE cRAISE_pharm_definition>
<pharm_definition>
```

```
<inclusion_volumes>
 <volume>
  <sphere type="donor" radius="1.0" \
       center_x="31.3" center_y="-5.3" center_z="-6.4"/>
  <direction direction_x="-2.5" direction_y="-0.9" direction_z="0.5"/>
 </volume>
 <volume>
  <sphere type="acceptor" radius="1.65" \
       center_x="30.4" center_y="-8.7" center_z="-7.5"/>
  <direction direction_x="-1.6" direction_y="-1.9" direction_z="1.1"/>
 </volume>
 <volume>
  <sphere type="hydrophil" radius="1.6" \
       center_x="32.6" center_y="-3.5" center_z="-5.3"/>
  <direction direction_x="0.2" direction_y="2.7" direction_z="-0.6"/>
 </volume>
 <volume>
  <sphere type="hydrophob" radius="1.0" \
       center_x="34.6" center_y="-6.6" center_z="-13.1"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="any" radius="1.0" \
       center_x="33.4" center_y="-4.2" center_z="-13.1"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
</inclusion_volumes>
<exclusion_volumes>
 <volume>
  <sphere type="donor" radius="1.0" \
       center_x="31.2" center_y="-8.9" center_z="-8.5"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="acceptor" radius="1.0" \
       center_x="30.5" center_y="-9.5" center_z="-7.5"/>
```

```
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="hydrophil" radius="1.6" \
       center_x="33.3" center_y="-3.4" center_z="-4.6"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="hydrophob" radius="1.0" \
       center_x="33.4" center_y="-10.2" center_z="-11.9"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="any" radius="1.0" \
       center_x="32.2" center_y="-3.0" center_z="-13.1"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
</exclusion_volumes>
<nof_essentials>3</nof_essentials>
</pharm_definition>
```

A.3 Beispielhafte Molekülprofildefinition

Ein Beispiel eines Molekülprofils, das ein weiter beschränktes Leitstrukturkriterium definiert, ist in Listing A.2 gegeben. Der Filter folgt der XML-Formatspezifikation von Mona und wurde in den großangelegten VS-Studien dieser Arbeit genutzt.

Listing A.2: Weiter beschränkter Leitstrukturfilter

```
<!DOCTYPE MonaFilterPresets>
<presetlist version="6">
 <preset name="restrictedleadlike">
  <filterchain tolerance="0">
   <filter inverted="0" type="0">
    <property id="0" min="250" max="300"/>
   </filter>
   <filter inverted="0" type="0">
```

```
   <property id="14" min="0" max="5"/>
  </filter>
  <filter inverted="0" type="0">
   <property id="3" min="0" max="3.5"/>
  </filter>
 </filterchain>
</preset>
</presetlist>
```

A.4 ALDR-Ensemble

ALDR-Ensemble mit Fidarestat-ähnlichen (Listing A.3) und IDD594-ähnlichen (Listing A.4) Liganden.

Listing A.3: Fidarestat-ähnliche Protein-Ligand-Komplexe

```
1ef3 1pwm 1x97 2agt 2pd9 2pdw 2pdy
```

Listing A.4: IDD594-ähnliche Protein-Ligand-Komplexe

```
1ah3 1eko 1el3 1iei 1t40 1t41 1us0 1x98 1z3n 2fzb 2fzd 2hv5 2i16 2i17
2ikj 2pdc 2pdg 2pdh 2pdj 2pdl 2pdn 2pdp 2pdq 2pdu 2pev 2pf8 2pfh 2pzn
2qxw 3bcj 3ghr 3ghs 3ght 3ghu 3lbo 3ld5 3lql 3lz5 3m64 3mc5 3onb 3onc
4jir
```

A.5 Ausführliche Screening-Ergebnisse

Tabelle A.3 zeigt die grundlegende Anreicherungsleistung auf dem $\mathrm{DUD}_{\mathrm{ACS}}$-Datensatz mittels der ROC-Metriken, die in Abschnitt 8.2 diskutiert sind. Die Klassifizierung der Zielstruktur entsprechend ihrer Qualität Q1 – Q4 folgt der in [347] definierten Qualitätskriterien.

A.6 Pharmakophorgeleitete Screening-Ergebnisse

Abbildung A.6 und A.6 zeigen alle ROC-Kurven, die aus VS-Läufen des $\mathrm{DUD}_{\mathrm{ACS}}$-Datensatz resultierten. Die roten Kurven resultierten aus gewöhnlichen VS-Läufen ohne Verwendung eine Pharmakophormodells. Die blauen und gelben Kurven wurden bei Verwendung eines Pharmakophormodells erhalten. Die blauen Kurven nutzten N_e^g-Modelle.

Tabelle A.3: Anreicherungsleistung auf dem DUD_{ACS}

Zielstruktur	Qualität	tAUC	$ROC_{5\%}$	$ROC_{2\%}$	$ROC_{1\%}$	Anti-Zielstruktur	ntAUC	ΔAUC	$A_t \cap A_{nt}$	$D_t \cap A_{nt}$
				Metalloenzyme						
pde5	Q2	0,76	0,28	0,16	0,10	comt	0.51	0,25	0	0
comt	Q4	0,50	0,00	0,00	0,00	pde5	0,41	0,09	0	0
ace	Q2	0,62	0,06	0,04	0,02	ada	0.31	0,31	0	0
ada	Q1	0,59	0,13	0,09	0,00	ace	0.54	0,04	0	0
Durchschnittlich		0,62	0,12	0,07	0,03		0,44	0,17		
				Kernrezeptoren						
er_agonist	Q1	0,91	0,63	0,61	0,45	mr	0,86	0,05	1	0
mr	Q1	0,84	0,47	0,20	0,13	er_agonist	0,81	0,03	1	0
er_antagonist	Q1	0,87	0,56	0,44	0,18	ppar	0,71	0,16	0	0
ppar	Q1	0,66	0,15	0,09	0,04	er_antagonist	0,56	0,10	0	0
rxr	Q1	0,83	0,65	0,35	0,20	ar	0,42	0,41	0	0
ar	Q1	0,80	0,43	0,16	0,08	rxr	0,74	0,06	0	0
pr	Q1	0,74	0,15	0,07	0,07	gr	0,75	-0,01	6	0
gr	Q1	0,49	0,09	0,05	0,04	pr	0,49	0,01	6	0
Durchschnittlich		0,77	0,39	0,25	0,15		0,67	0,10		
				Kinasen						
src	Q3	0,69	0,20	0,13	0,05	fgfr1	0,53	0,16	84	0
fgfr1	Q4	0,46	0,09	0,03	0,02	src	0,59	-0,13	84	0
hsp90	Q1	0,66	0,21	0,08	0,08	egfr	0,54	0,11	0	0
egfr	Q4	0,60	0,13	0,08	0,06	hsp90	0,57	0,03	0	0
vegfr2	Q4	0,60	0,23	0,10	0,05	p38	0,53	0,07	0	0
p38	Q1	0,56	0,23	0,11	0,07	vegfr2	0,43	0,13	0	0
cdk2	Q3	0,60	0,26	0,14	0,10	pdgfrb	0,37	0,23	0	0
pdgfrb	Q4	0,28	0,01	0,01	0,01	cdk2	0,57	-0,29	0	0
tk	Q3	0,59	0,09	0,09	0,05	pnp	0,62	-0,03	0	0

Tabelle A.3: Fortsetzung Anreicherungsleistung auf dem DUD_{ACS}

Zielstruktur	Qualität	tAUC	$ROC_{5\%}$	$ROC_{2\%}$	$ROC_{1\%}$	Anti-Zielstruktur	ntAUC	ΔAUC	$A_t \cap A_{nt}$	$D_t \cap A_{nt}$
Durchschnittlich		0,56	0,16	0,09	0,05		0,53	0,03		
				Serinproteasen						
thrombin	Q4	0,69	0,12	0,03	0,02	fxa	0,57	0,12	0	0
fxa	Q1	0,64	0,17	0,11	0,08	thrombin	0,70	-0,06	0	0
trypsin	Q3	0,50	0,14	0,09	0,05	hivpr	0,75	-0,25	0	0
Durchschnittlich		0,61	0,14	0,08	0,05		0,67	-0.06		
				Folatenzyme						
dhfr	Q3	0,71	0,09	0,03	0,01	gart	0,69	0,02	0	0
gart	Q3	0,59	0,00	0,00	0,00	dhfr	0,19	0,30	0	0
Durchschnittlich		0,65	0,04	0,02	0,01		0,44	0,16		
				Andere						
sahh	Q1	0,95	0,82	0,70	0,06	cox1	0,74	0,21	0	0
cox1	Q1	0,54	0,00	0,00	0,00	sahh	0,61	-0,07	0	0
cox2	Q3	0,90	0,73	0,62	0,51	na	0,89	0,01	0	0
na	Q1	0,59	0,02	0,00	0,00	cox2	0,35	0,24	0	0
gpb	Q1	0,85	0,27	0,14	0,12	hivrt	0,33	0,52	0	0
hivrt	Q1	0,53	0,03	0,00	0,00	gpb	0,37	0,16	0	0
parp	Q1	0,77	0,52	0,52	0,52	inha	0,54	0,23	0	0
inha	Q4	0,59	0,27	0,24	0,20	parp	0,47	0,12	0	0
ampc	Q3	0,73	0,05	0,00	0,00	alr2	0,65	0,08	0	0
alr2	Q2	0,51	0,19	0,12	0,08	ampc	0,46	0,05	0	0
hmga	Q1	0,52	0,00	0,00	0,00	ache	0,39	0,13	0	0
ache	Q2	0,47	0,06	0,02	0,01	hmga	0,40	0,07	0	0
pnp	Q3	0,71	0,12	0,00	0,00	tk	0,54	0,16	0	0
hivpr	Q1	0,60	0,11	0,06	0,02	trypsin	0,57	0,03	0	0
Durchschnittlich		0,66	0,23	0,17	0,11		0,52	0,14		

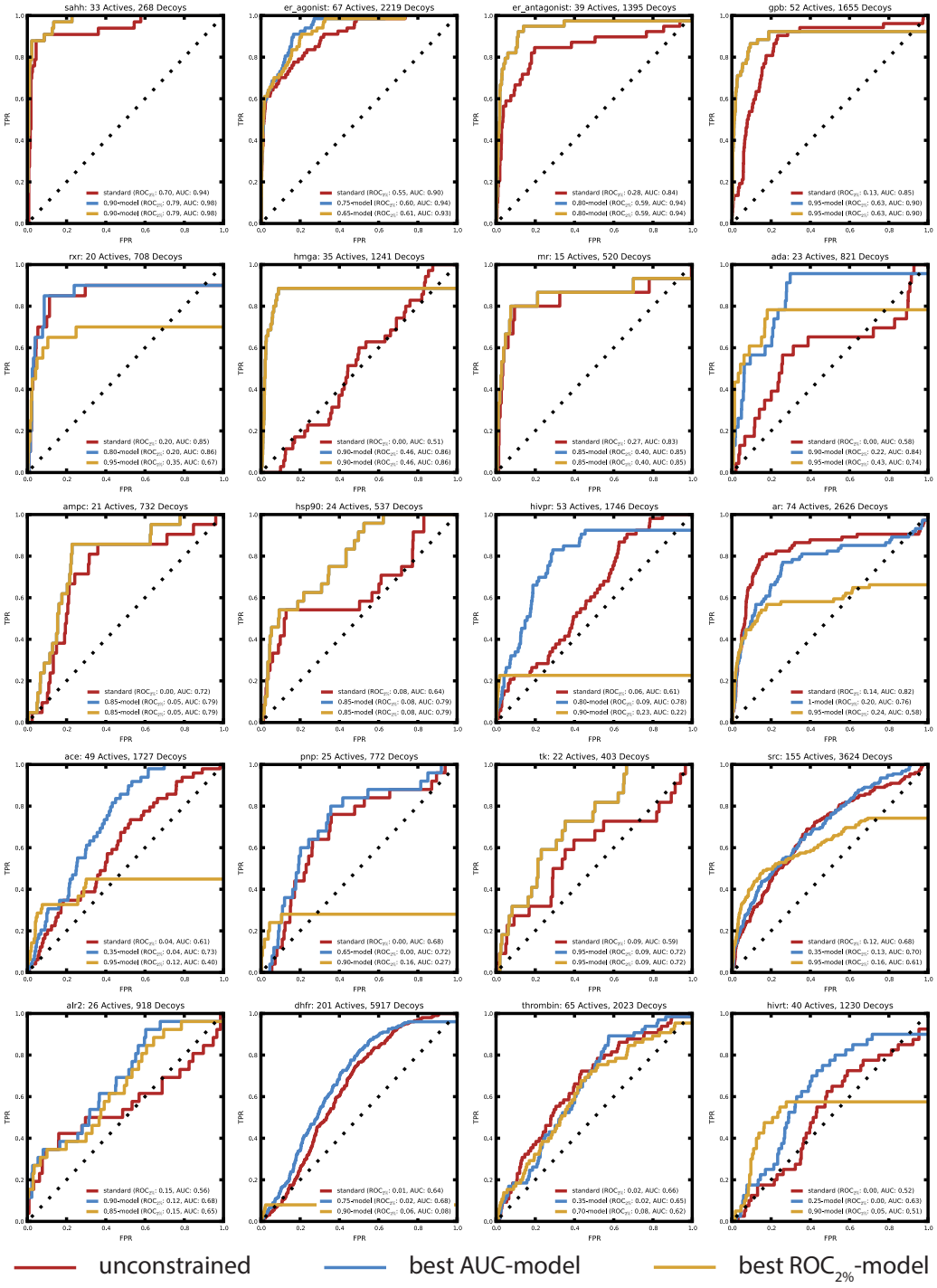


Abbildung A.1: ROC-Kurven beim VS des $\mathrm{DUD}_{\mathrm{ACS}}$ ohne (rot) und mit Pharmakophormodell (blau mit N_e^g-Modellen, gelb mit strikteren Modellen)

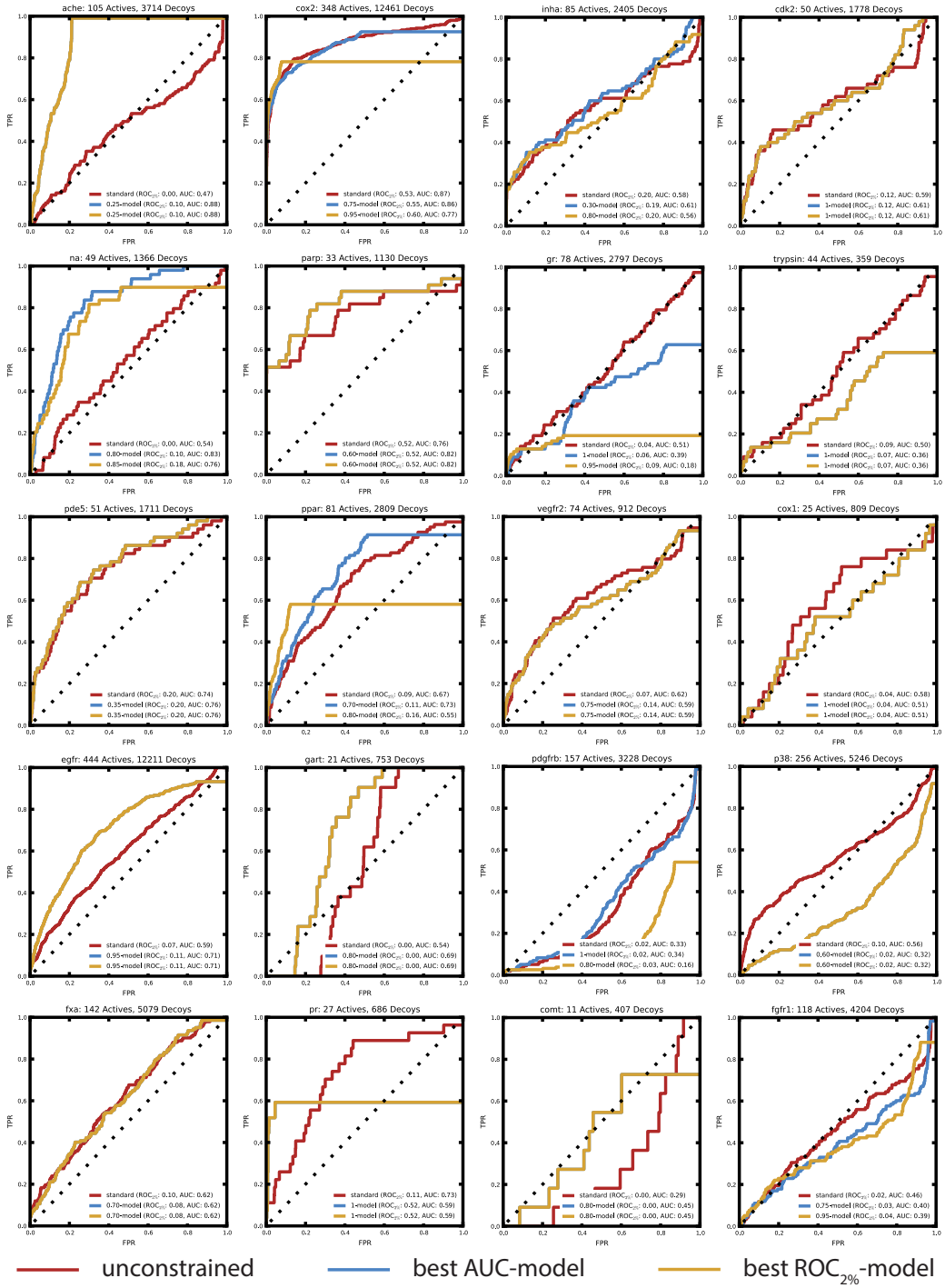


Abbildung A.2: ROC-Kurven beim VS des DUD_{ACS}: Fortsetzung

B Dokumentation der Implementierung

Dieser Abschnitt bietet einen Überblick über die Implementierung von cRAISE, das im Zuge dieser Arbeit entstand. Es ist ein Softwarepaket, das in C++ realisiert wurde und aus den Werkzeugen

- zur Bibliothekspräparierung (*craiseIndexer*)
- zum strukturbasierten virtuellen Screening (*craiseScreener*)
- zur Durchführung einzelner Docking-Läufe (*craiseDocker*)
- und zur Aufbereitung eines Pharmakophormodells (*craisePharmGenerator*)

besteht. Zudem wurden als Teil der Naomi-Softwarebibliothek

- das *RAISE*-Modul und
- das *FastBitIndex*-Modul

entwickelt, um generische Funktionalitäten auch zur Verwendung in RAISE-Derivaten wie iRAISE und TrixP zur Verfügung zu stellen. Einige der entwickelten Algorithmen, die allgemeine Probleme lösen und in anderem Kontext Verwendung finden können, wurden in die entsprechenden Basismodule der Naomi-Bibliothek überführt. Neben den Abhängigkeiten zu Modulen, der am Zentrum für Bioinformatik in der Arbeitsgruppe für Algorithmisches Molekulares Design intern entwickelten Naomi-Bibliothek, besitzt cRAISE auch Abhängigkeiten zu extern entwickelten Softwarebibliotheken. Alle internen und externen Abhängigkeiten sind in Abbildung B dargestellt. Im Folgenden werden die von cRAISE genutzten Funktionalitäten aus den internen und externen Bibliotheken und die neu entwickelten Bestandteile aufgeführt.

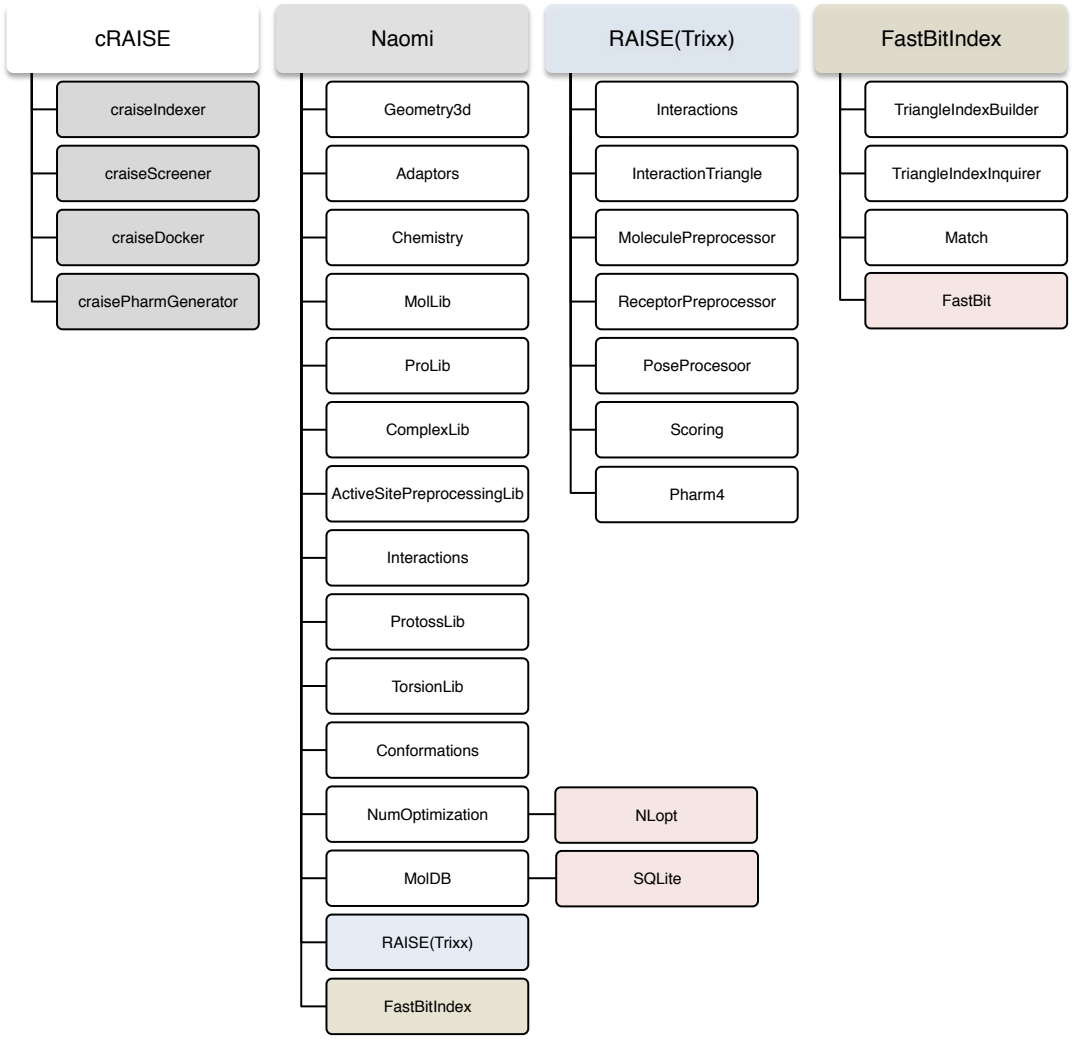


Abbildung B.1: Übersicht über die cRAISE-Entwicklung. Interne Abhängigkeiten (grau) und externe Abhängigkeiten (rot). Neben den cRAISE-Werkzeugen, wurden das *RAISE(Trixx)*- und das *FastBitIndex*-Modul entwickelt

Abhängigkeiten zur Naomi-Bibliothek

Die cRAISE-Werkzeuge verbinden Funktionalitäten aus unterschiedlichen Modulen der Naomi-Bibliothek:

- *Geometry3d* enthält Datenstrukturen zur Repräsentation dreidimensionaler Objekte (z. B. Punkte, Vektoren, Sphären, Ikosaeder, etc.) und Algorithmen, um geometrische Probleme im Raum zu lösen, z. B. Berechnung affiner Transforma-

tionen. Die Berechnung der konvexen Hülle mittels des Quick-Hull-Algorithmus wird verwendet, um das aktive Volumen der Bindetasche zu bestimmen.

- *Adaptors* enthält die Anbindung zur graphischen Darstellung dreidimensionaler Objekte über FlexV, die für den *craiseDocker* und *craisePharmGenerator* verwendet wird. Zudem findet sich hier auch die Anbindung zur Oberflächenberechnung, die zur Bestimmung zugänglicher Atome der Bindetasche genutzt wird.
- *Chemistry* enthält die Repräsentation und statische Information von Element, Valenzzustand und Atomtyp nach dem Naomi-Modell.
- *MolLib* enthält die Datenstruktur des Naomi-Moleküls und Funktionalitäten zum Aufbau, Lesen und Schreiben, zur Kanonisierung, zu Hinzufügen von Wasserstoffen und zur Überlagerung. Zudem findet sich hier die Implementierung des VSC-Modells, die die Identifizierung von Multizustandszonen und -atomen, aber auch die Normalisierung und die Enumeration alternativer Zustände ermöglicht.
- *ProLib* enthält die Datenstruktur des Naomi-Proteins und Funktionalitäten zum Lesen, Aufbau und Schreiben von Naomi-Proteinen.
- *ComplexLib* enthält die Datenstruktur des Naomi-Komplex und Funktionalitäten zum Lesen, Aufbau und Schreiben von Naomi-Komplexen.
- *ActiveSitePreprocessingLib* enthält die Datenstruktur der aktiven Bindetasche, Funktionalitäten zum internen Aufbau und unterschiedliche Präparierungsmöglichkeiten, z. B. welche Wassermoleküle als Teil der Bindetasche betrachtet und in welchem Zustand Residuen initialisiert werden sollen.
- *Interactions* baut gerichtete Interaktionsgeometrien für individuelle Atome. Es interpretiert die Naomi-Atominformation und die direkte Atomumgebung und entschiedet ob, wie viele und in welche Richtungen potentiell Interaktionen gebildet werden können. Es enthält auch Funktionalitäten zur paarweisen Bewertung von Interaktionsgeometrien, die cRAISE jedoch nicht verwendet.
- *ProtossLib* enthält die Implementierung von Protoss und Funktionalitäten zur individuellen Wasserstoffbrückennetzwerkoptimierung.
- *TorsionLib* enthält die Implementierung, der von Confect genutzten Torsionsbibliothek.
- *Conformations* enthält die Implementierung von Confect und die Definition der Qualitätsstufen. Zudem sind auch die cRAISE-spezifischen Modifikationen enthalten, die bei Angabe der cRAISE-Qualitätsstufe ausgeführt werden.

- *NumOptimization* enthält die Implementierung einer numerischen Optimierungsroutine zur Optimierung von Ligandkoordinaten, die optional zur Nachbewertung von Posen ausgeführt werden können.
- *MolDB* enthält Funktionalitäten zum Aufbau und Zugriff auf die Molekülbibliothek und zur Filterung anhand molekularer Eigenschaften.

Abhängigkeiten zum RAISE(Trixx)-Modul

Das *RAISE(Trixx)*-Modul stellt Datenstrukturen und Funktionalitäten für deskriptorbasierte Screening-Anwendungen bereit:

- *Interactions* enthält Funktionalitäten, um Interaktionsstellen an ein Molekül bzw. Rezeptor zu annotieren. Es baut auf dem *Interactions*-Modul der NAOMI-Bibliothek zur Darstellung von gerichteten Interaktionen auf, schränkt sie aber auf plausible ein. Zudem werden hier ungerichtete hydrophobe Interaktionsstellen und gegebenenfalls Multizustandsinteraktionsstellen bestimmt.
- *InteractionTriangle* enthält Funktionalitäten, um anhand von Interaktionsstellen, molekül- und rezeptorseitig Deskriptoren zu enumerieren. Zudem können Deskriptoren durch ein Pharmakophor beschränkt und Multizustandsdeskriptoren berechnet werden.
- *MoleculePreprocessor* baut ein RAISE-Molekül auf, das Molekülinformation, Interaktionsstellen und Moleküldeskriptoren enthält.
- *ReceptorPreprocessor* baut ein RAISE-Rezeptorstruktur auf, die Bulk-Atome, aktive Atome und zugängliche Atome klassifiziert, Interaktionstellen, Deskriptoren und ggf. vorberechnete Bewertungsinformation enthält.
- *PoseProcessor* initialisiert anhand einer Transformation und einer Konformation eine Pose und enthält Funktionalitäten, um die Pose auf ein Molekül anzuwenden und sie mit einem Rezeptor zu bewerten.
- *Scoring* enthält die Implementierung der gitter- und atombasierten cRAISE-Bewertungfunktion, die Vorberechnung von Überlappinformation und Annotation von Exklusionsmerkmalen.
- *Pharm4* enthält die Datenstruktur des cRAISE-Pharmakophors und Funktionalitäten zur Erstellung, zum Schreiben und Lesen von Pharmakophormodellen und zur Überprüfung von Posen auf die Einhaltung von Inklusionsmerkmalen.

Abhängigkeiten zum FastBitIndex-Modul

Das *FastBitIndex*-Modul bildet die Anbindung an FastBit, um von RAISE-Deskriptoren einen Index aufzubauen. Zudem sind hier Funktionalitäten implementiert, um unterschiedliche Arten von Anfragen an einen Deskriptorindex zu stellen.

- *TriangleIndexBuilder* Funktionalitäten zum Aufbau einer Indexstruktur anhand von RAISE-Deskriptoren.
- *TriangleIndexInquirer* Funktionalitäten zur Anfrage an eine Indexstruktur und zum sukzessiven, speicherschonenden Iterieren über Deskriptortreffer.
- *Match* Minimalrepräsentation eines Deskriptortreffers.

Abhängigkeiten zu externen Bibliotheken

Externe Abhängigkeiten finden sich vor allem innerhalb der NAOMI-Bibliothek wieder:

- *FastBit* 1.2.4 (`https://codeforge.lbl.gov/projects/fastbit`) zur Erstellung eines komprimierten Bitmap-Index.
- *NLOpt* 2.3 (`http://ab-initio.mit.edu/wiki/index.php/NLopt`) eine frei verfügbare Open-Source-Bibliothek zur nicht-linearen Optimierung, die unterschiedliche Optimierungsroutinen bereitstellt.
- *SQLite* (`https://www.sqlite.org`) eine öffentliche Softwarebibliothek, die eine serverfreie SQL-Datenbank implementiert.
- *Qt* 5.1.1 (`http://qt-project.org/downloads`) C++-Klassenbibliothek zur plattformübergreifenden Programmierung, hauptsächlich von graphischen Oberflächen. Für cRAISE sind vor allem, die Datenbank- und XML-Funktionalitäten und die Funktionalitäten zum Datei-Handling relevant.
- *Boost* (`http://www.boost.org`) C++-Klassenbibliothek, die typische Datenstrukturen und Algorithmen standardisiert und frei zur Verfügung stellt.

C Benutzung der Software

C.1 Einführung

C.1.1 Über cRAISE

cRAISE[374] entstand auf Basis von Forschungsarbeiten, die in der Gruppe für Algorithmisches Molekulares Design am Zentrum für Bioinformatik der Universität Hamburg durchgeführt wurden. Es ist ein Softwarepaket zum Screening einer virtuellen Molekülbibliothek gegen die aktive Bindetasche eines Proteins. Die Bindetasche muss durch den Anwender spezifiziert und die Molekülbibliothek bereitgestellt werden. Sind zusätzliche Informationen über Liganden oder über typische Bindungsmuster bekannt, so kann dieses Wissen im Screening-Prozess genutzt werden. Dafür können Pharmakophormodelle und/oder Molekülprofile durch den Nutzer definiert werden. Die Zusatzinformation wird während des Screening-Prozesses verarbeitet, um eine gerichtete Suche durch den Molekülraum zu unterstützen. cRAISE erweitert optional den Molekülraum mit den Konformeren flexibler Liganden. Zudem kann optional der Suchraum mit Protomere des Zielproteins und der Moleküle erweitert werden, um mögliche Protonierungszustände und Tautomere zu berücksichtigen. Um die Ergebnisse, die mit cRAISE erzeugt werden, verstehen und interpretieren zu können, ist ein Grundwissen über die zugrundeliegenden Modelle und Algorithmen notwendig. cRAISE[374] basiert auf den Konzepten seiner Vorgängerversionen[316, 1, 2], integriert das chemische Model und Funktionalitäten von Naomi[3] und Funktionalitäten von Confect[254], Mona[314] und Protoss[315]. Die Erläuterung der Konzepte ist in den entsprechenden Publikationen zu finden und nicht Gegenstand dieses Leitfadens. cRAISE ist eine Software, die stetig weiterentwickelt wird. Die Programme sind und werden mit einer kontinuierlich wachsenden Anzahl von Protein- und Moleküldaten getestet, dennoch sind wir sicher, dass cRAISE nicht frei von Fehlern ist. Zudem hängt die Leistungsfähigkeit von cRAISE entscheidend von der

Qualität der gegebenen Daten ab. Daher kann generell keine Gewährleistung für die erzielten Ergebnisse übernommen werden.

C.1.2 Über diesen Leitfaden

cRAISE ist ein reines Kommandozeilenprogramm. Es unterstützt direkt keine interaktive Verwendung oder graphische Oberfläche. In diesem Leitfaden sind alle notwendigen Anweisungen und Programmoptionen beschrieben, die ein virtuelles Screening mit den cRAISE-Programmen ermöglichen und den kompletten Funktionsumfang beschreiben. Jede Anweisung an cRAISE ist ein Programmaufruf, dem eine Liste von Programmparametern folgt:

```
programm <parameter 1> <parameter 2> [<parameter 3>] ...
```

Die Reihenfolge der Parameter ist dabei irrelevant. Optionale Parameter sind durch eckige Klammern gekennzeichnet: [<optionaler Parameter>]. Endet eine Textzeile mit einem \-Zeichen, dann wird die Anweisungszeile in der nächsten Textzeile fortgesetzt. Endet eine Textzeile nicht mit einem \-Zeichen, dann ist die Anweisung beendet. Programmparameter können immer in einer ausführliche Form oder in einer Kurzform angegeben werden. Der Typ eines Programmparameters wird in seiner Kurzform mit einem einzelnen Buchstaben angegeben und mit „-“ eingeleitet. In der ausführlichen Form ist der Parameter durch mehrere Buchstaben charakterisiert und wird mit „--“ eingeleitet. Handelt es sich bei der Programmoption um einen Schalter, so steht der Parameter allein. Spezifiziert der Parameter einen Wert, so folgt dieser dem Parametertyp getrennt durch ein Leerzeichen. Werte können entweder eine Datei, ein Pfad oder ein numerischer Wert sein. Essentielle Parameter sind in der Regel durch Großbuchstaben in der Kurzform gekennzeichnet, wohingegen optionale Parameter mit Kleinbuchstaben gekennzeichnet sind. Beispiele für Programmanweisungen werden in diesem Leitfaden immer wie folgt hervorgehoben:

```
craiseScreener -v
```

Beispiele für Standardausgaben von cRAISE sind grau hinterlegt:

Detaillierte Aufführung C.1: Beispielausgabe

```
user@host:/local/user/craise> craiseScreener -v
cRAISE (alpha) (Built: May 31 2015 12:09:20)
+ Version: 1.0.0
```

C.2 Installation

C.2.1 Bestandteile von cRAISE

Das cRAISE-Softwarepaket ist als komprimiertes tar-Archiv craise.tar.gz bereitgestellt. Nach entpacken (`tar -xvzf craise.tar.gz`) enthält das Paket, die in Tabelle C.1 gelisteten Bestandteile. craiseIndexer, craiseScreener, craisePharmGenerator und craiseDocker sind ausführbare Binärdateien. Sollten diese nicht das „x“-Flag gesetzt haben, muss dieses mit `chmod +x <programm>` geändert werden.

Tabelle C.1: Bestandteile des cRAISE-Softwarepakets

Dateiname	Beschreibung
craiseIndexer	Werkzeug zur Präparierung von Molekülbibliotheken
craiseScreener	Werkzeug zum virtuellen Screening
craisePharmGenerator	Werkzeug zur Pharmakophorgenerierung
craiseDocker	Werkzeug zum Protein-Ligand-Docking
lib/	Dynamische Bibliotheken
bin/	FlexV
userGuide.pdf	Dieser Leitfaden
examples/	Bsp. Pharmakophordefinition und Molekülprofil

C.2.2 Lizenzierung

Die cRAISE-Werkzeuge stehen für den akademischen Gebrauch frei zur Verfügung. Sie sind jedoch durch einen Lizenzschlüssel geschützt. Bei Ausgabe von

Detaillierte Aufführung C.2: Lizenz notwendig

```
Your license for craiseIndexer has expired. \
Please provide a new one with the -license option.
```

muss für das Werkzeug *craiseIndexer* eine neue Lizenz bereitgestellt werden. Der Schlüssel `<key>` kann unter Angabe des Namen des Werkzeugs auf Anfrage am Zentrum für Bioinformatik der Universität Hamburg erhalten werden. Mit

```
craiseIndexer --license <key>
```

kann man dann die Lizenz verlängern und erhält erneut den vollen Funktionsumfang.

C.2.3 Installationsanleitung

Nach Wechseln in das `craise/`-Verzeichnis sind alle Funktionalitäten von cRAISE verfügbar. Es ist notwendig, dass die binär zur Verfügung gestellten, dynamisch genutzten Bibliotheken des `lib/`-Verzeichnisses während der Ausführung relativ zu den cRAISE-Binärdateien vorliegen.

C.2.4 Notwendige Bibliotheken

cRAISE nutzt die auf dem System installierte glibc (2.11.2) und libstd++ (6.0.14). Sind diese Bibliotheken auf dem genutzten System nicht verfügbar, so müssen sie bereitgestellt werden. Die FastBit- (1.2.4) und Qt5-Bibliotheken (5.1.1), die den Indexierungsprozess unterstützen, aber auch die Schnittstelle zu den von cRAISE genutzten/erzeugten SQLite-Datenbanken bieten und das Prozessieren von XML-Dateien ermöglichen, sind im lib/-Verzeichnis bereitgestellt. Sind diese Bibliotheken nicht nutzbar, so müssen deren Quellen erneut kompiliert und die resultierenden Bibliotheken im lib/-Verzeichnis entsprechend bereitgestellt werden. Die Quellen sind unter `https://codeforge.lbl.gov/projects/fastbit` bzw. `http://qt-project.org/downloads` verfügbar. Zur Verwendung der Optimierungsfunktion muss zusätzlich die NLopt-Bibliothek (2.3) (`http://ab-initio.mit.edu/wiki/index.php/NLopt`) zur Verfügung gestellt werden.

C.2.5 Externe Programme

Einige Funktionen von cRAISE basieren auf externer Software. Auch wenn die cRAISE-Tools mit den notwendigen Bibliotheken (siehe C.2.4) bereits eigenständig genutzt werden können, ist es ratsam die folgenden Hilfsmittel verfügbar zu halten, um den vollen Funktionsumfang von cRAISE nutzen zu können.

Graphische Darstellung

Generell besitzen die craiseIndexer- und craiseScreener-Werkzeuge keine graphische Benutzerschnittstelle. Sollen Screening-Ergebnisse visualisiert werden, so können die von cRAISE erzeugten Posen exportiert werden und extern in einen gängigen Molekülbetrachter geladen werden. Das craiseDocker- und craisePharmGenerator-Programm bietet darüber hinaus eine Schnittstelle zu FlexV, einem Visualisierungswerkzeug, das auf OpenGL basiert. FlexV ist im bin/-Verzeichnis bereitgestellt, das dort relativ zu den cRAISE-Binärdateien vorliegen muss.

Mona

Zur Filterung von Molekülen während des Screening-Prozesses müssen zuvor beschränkende molekularen Eigenschaften definiert werden. Dies setzt voraus, dass die Verteilung der molekularen Eigenschaften innerhalb Molekülbibliothek bekannt ist. Mona[314] ist ein Werkzeug, das Möglichkeiten zur Analyse und Visualisierung von Molekülbibliotheken bietet. Zur akademischen Nutzung ist Mona derzeit unter `http://www.zbh.uni-hamburg.de/mona` frei verfügbar. Mit Mona können unter anderem auch Molekülprofile definiert und in einem XML-Format exportiert werden. cRAISE unterstützt die von Mona exportierten Filterregeln und ist in der Lage diese zu prozessieren.

C.3 Arbeiten mit cRAISE

Dieser Abschnitt stellt eine Einführung in das Arbeiten mit cRAISE dar. Zuerst wird gezeigt, wie mit cRAISE eine gegebene Molekülbibliothek präpariert und ein einfaches strukturbasiertes virtuelles Screening durchgeführt werden kann. In den darauffolgenden Abschnitten wird erläutert, wie der Anwender den Screening-Prozess mit bereitgestellten Pharmakophordefinitionen und Molekülprofilen beeinflussen kann. Zudem wird gezeigt wie ein Pharmakophormodell für die Anwendung in cRAISE erzeugt und für einzelne Moleküle selektiv ein Docking durchgeführt werden kann.

C.3.1 Präparierung eines virtuellen Screenings

cRAISE ist zum strukturbasierten virtuellen Screening großer Molekülbibliotheken gedacht, die sich inhaltlich kaum ändern. Dafür wird die Molekülbibliothek einmalig präpariert indem Moleküldeskriptoren berechnet und persistent in den cRAISE-Index überführt werden. Die Präparierung der Molekülbibliothek führt man mit dem Indexierungswerkzeug durch. Gibt man in der Eingabeaufforderung unter dem `craise/`-Verzeichnis

```
craiseIndexer
```

ein, dann erklärt das Programm unter „Usage“ die Verwendung des Programms mit den essentiellen Parametern und listet zudem alle optionalen Parameter, die vom Indexierungswerkzeug berücksichtigt werden können:

Detaillierte Aufführung C.3: Indexer Optionen

```
user@host:/local/craise> craiseIndexer
cRAISE (alpha) (Built: May 31 2015 12:09:20)
+ Version 1.0.0
```

```
Usage:
craiseIndexer -L <arg> -X <arg> [options]

craiseIndexer options:
-L [ --library ] arg   Input library file (suffix required)
-X [ --craiselib ] arg cRAISE prepared library output path
-c [ --confgen ] arg (=0) Generate conformers. Quality: 0, 1, 2, 3, 100
-s [ --states ]        Generate protonation and tautomeric states
-f [ --from ] arg (=0) Start indexing from entry
-t [ --to ] arg (=0)   End indexing at entry (off end)
-x [ --suffix ] arg    Add suffix to molecule names
-l [ --log ] arg (=0)  Log level: 0 (quiet) - 5 (steps)
--license arg          Provide new license
```

Mit

```
craiseIndexer -L <library> -X <craiselib>
```

wird die Indexierung der Molekülbibliothek `<library>` gestartet und das Ergebnis (als cRAISE-*Bibliothek* bezeichnet) in das Verzeichnis `<craiselib>` geschrieben. Die Molekülbibliothek kann dafür in Form einer Multi-SDF- oder Multi-MOL2-Datei bereitgestellt werden. Es ist notwendig, dass die Dateiendung explizit angegeben wird, damit das entsprechende Format gelesen werden kann. `<craiselib>` steht für einen durch den Anwender spezifizierten Pfad. Existiert dieser zuvor nicht, so wird er während der Programmausführung angelegt. Nachdem die Molekülbibliothek präpariert und die Indexierung erfolgreich beendet wurde, sind im `<craiselib>`-Verzeichnis die in Tabelle C.2 gelisteten Inhalte vorzufinden. Da dieses Verzeichnis eine Eingabe des Screening-Prozesses darstellt, ist es wichtig dessen Struktur und Inhalt nach Erstellung zu erhalten.

Tabelle C.2: Bestandteile der cRAISE-Bibliothek

Dateiname	Beschreibung
<library>_indexer.err	Fehlermeldungen
<library>_indexer.log	Statusmeldungen
idx/	cRAISE-Index
idx_config.ini	cRAISE-Index Konfiguration
mol.db	Index-assoziierte Molekül datenbank
uuid	ID der cRAISE-Bibliothek

`<library>_indexer.err` sollte nach erfolgreicher Ausführung eine leere Datei sein. Schlägt die Ausführung fehl, so sind hier Fehlermeldungen vorzufinden. `<library>_indexer.log` dokumentiert die durch den Anwender vorgenommen Einstellungen und hält Informationen zum Ablauf des Programms fest:

Detaillierte Aufführung C.4: Indexer Log-Datei

```
cRAISE (alpha) (Built: Oct 5 2014 10:23:02)
+ Version: 1.0.0

Indexing options:
+ Input library: astex_ligs/1g9v_1.mol2
+ From entry: 0
+ To entry (off the end): 0
+ Suffix added to molecule database entries:
+ Molecule database and Index located in: test_idx
+ Conformer generation disabled
+ Library protonation and tautomeric states disabled
+ Logging level: 5

> Determininig directed molecule interactions ...
> Nof directed interactions found: 8
> Determininig undirected molecule interactions ...
> Calculating molecule index triangles ...

Index:
+ cRAISE version: 1.0.0
+ Molecules: 1
+ Discarded molecules: 0
+ Conformers: 0
+ Conformer level: 0
+ Max. number of conformers: 250
+ Descriptors: 69
+ States: no

Times:
+ db/index buildup: 0.0s wall, 0.0s user + 0.0s system = 0.1s CPU (81.7%)
```

Die wichtigsten Bestandteile der cRAISE-Bibliothek sind der in `idx/` gespeicherte Deskriptor-Index und dessen assoziierte Molekül datenbank (`mol.db`). `idx_config.ini` ist eine Konfigurationsdatei, die die Art des erstellten Index beschreibt und im Screening-Prozess verarbeitet wird. `uuid` versieht die cRAISE-Bibliothek mit einer eindeutigen ID. Sie wird für eine konsistente Auswertung von Screening-Resultaten bei paralleler Ausführung benötigt.

C.3.2 Konfiguration der Präparierung

Konformergenerierung (-c) Ohne jegliche Konfiguration nimmt cRAISE an, dass die durch den Anwender bereitgestellte Molekülbibliothek bereits mit Konformeren angereichert ist. Ist dies nicht der Fall, so ist es möglich durch Angabe von `-c <arg>` während des Indexierungsprozesses Konformere zu erzeugen. `<arg>` spezifiziert dabei die Qualität der erzeugten Konformere. Folgende Qualitätsstufen stehen derzeit zur Verfügung:

Tabelle C.3: Qualitätsstufen der Konformergenerierung

Qualitätsstufe	Beschreibung
0	Keine Konformergenerierung
1	Confect-Qualitätsstufe 1
2	Confect-Qualitätsstufe 2
3	Confect-Qualitätsstufe 3
100	cRAISE-Qualitätsstufe

Zur Erläuterung der Qualitätsstufen sei hier auf die Originalpublikation von Confect[254] verwiesen. Prinzipiell führt eine Erhöhung der Qualitätsstufe zu einer erhöhten Anzahl an Konformeren. Dies kann den Indexierung- und den Screening-Prozess deutlich verlangsamen und zu erhöhtem Speicherbedarf führen. Deshalb ist die Anzahl erzeugter Konformere auf 250 beschränkt. Die cRAISE-Qualitätsstufe ist eigens auf cRAISE abgestimmt und nutzt dieses Limit erschöpfend aus.

Zustandsgenerierung (-s) Durch Einstellen des `-s`-Schalters werden intern Protonierungszustände und Tautomere der prozessierten Moleküle generiert und der Index durch Multizustandsdeskriptoren erweitert.

Spezifizierung eines Bereichs der Eingabebibliothek (-f, -t) Mit `-f <arg>` und `-t <arg>` kann die Eingabebibliothek partitioniert werden.

```
craiseIndexer -L <library> -X <craiselib> -f 0 -t 1000
```

prozessiert die ersten 1000 Moleküle der Eingabebibliothek.

Spezifizierung eines Molekülsuffix (-x) Mit `-x <arg>` kann ein Suffix spezifiziert werden, das an die Molekülnamen der prozessierten Eingabemoleküle angehängt wird.

```
craiseIndexer -L <library> -X <craiselib> -x decoy
```

erweitert alle Molekülnamen der Moleküle in `<library>` mit „decoy“.

Spezifizierung des Log-Levels (-l) Für eine großangelegte Indexierung sind die Statusmeldungen grundsätzlich auf ein Minimum eingestellt (`-l 1`). Zur Fehlerdiagnose kann die Ausgabe bis zu einem Wert von 5 erhöht werden. Alternativ kann die Ausgabe komplett ausgestellt werden (`-l 0`).

C.3.3 Durchführung eines virtuellen Screenings

Das virtuelle Screening startet man mit dem Screening-Werkzeug. Tippt man in der Eingabeaufforderung unter dem `craise/`-Verzeichnis

```
craiseScreener
```

dann listet das Programm alle möglichen Parameter, die vom Screening-Werkzeug berücksichtigt werden können:

Detaillierte Aufführung C.5: Screener Optionen

```
user@host:/local/craise> craiseScreener
cRAISE (alpha) (Built: May 31 2015 12:09:20)

Usage:
craiseScreener -S -X <arg> -T <arg> -R <arg> -O <arg> [options]
craiseScreener -E [-h|d|j <arg>] -O <arg> [options]

Screening options:
-S [ --screen ]              Start screening
-T [ --target ] arg          Target protein (suffix required)
-X [ --craiselib ] arg       cRAISE library (moldb/index) path
-R [ --reflig ] arg          Reference ligand (suffix required)
-O [ --outdir ] arg          Write solution database to
```

```
-a [ --activeradius ] arg (=6.5) Active site radius
-p [ --property ] arg         Mona property xml file
-c [ --pharm ] arg            Constrain queries with pharmacophore
-o [ --optimize ]             Optimization and re-scoring of best poses

Evaluate options:
-E [ --evaluate ]       Evaluate screening results
-O [ --outdir ] arg     Write hitlist/poses/solutions to
-h [ --hitlist ] arg    Write hitlist from solution database
-d [ --draw ] arg       Export hits to sdf from solution database
-j [ --join ] arg       Join *_sol.db in directory
-n [ --hits ] arg (=1000) Max number of hits written/drawn
-b [ --basename ] arg   Base name of output

General options:
-l [ --log ] arg (=0)   Log level: 0 (quiet) - 5 (steps)
--license arg           Provide new license
```

Das Screening-Werkzeug ist in zwei Modi ausführbar. Mit der Option `-S` wird der Screening-Modus gestartet. Dafür wird mit

```
craiseScreener -S -X <craiselib> -T <target> -R <reference> -O <out>
```

ein Screening des Zielproteins `<target>` gegen die cRAISE-Bibliothek `<craiselib>` gestartet. Es setzt voraus, dass die cRAISE-Bibliothek zuvor erfolgreich mit dem Indexierungs-Werkzeug erstellt wurde. Das Zielprotein `<target>` kann als PDB- oder MOL2-Datei vorliegen. `<reference>` bezeichnet ein Referenzmolekül, das zur Bestimmung der aktiven Bindetasche herangezogen wird. Dafür definiert cRAISE standardmäßig die Proteinatome als Teil der Bindetasche, die nicht weiter als 6,5 Å von irgendeinem Schweratom der Referenzmoleküls entfernt liegen. Die Referenz kann als SDF- oder MOL2-Datei bereitgestellt werden. Auch hier ist es notwendig das explizit Dateiendungen angegeben werden, damit das Moleküldateiformat entsprechen prozessiert werden kann. `<out>` spezifiziert einen Pfad, der nach erfolgreicher Ausführung das Screening-Resultat enthält. Existiert dieser zuvor nicht, so wird er während der Programmausführung angelegt. Nachdem das Screening erfolgreich beendet wurde, sind im `<out>`-Verzeichnis die in Tabelle C.4 gelisteten Inhalte vorzufinden.

Analog zum Indexierungswerkzeug, dokumentiert das Screening-Werkzeug in der Datei `<craiselib>_screener.err` bzw. `<craiselib>_screener.log` eventuelle Fehler

Tabelle C.4: Screening-Ausgabe

Dateiname	Beschreibung
<craiselib>_screener.err	Fehlermeldungen
<craiselib>_screener.log	Statusmeldungen
<craiselib>_sol.db	Lösungsdatenbank

bzw. die durch den Anwender vorgenommen Einstellungen und hält Informationen zum Ablauf des Programms fest:

Detaillierte Aufführung C.6: Screener Log-Datei

```
cRAISE (alpha) (Built: Oct 5 2014 10:23:02)
+ Version: 1.0.0

Screening options:
+ Target: target.pdb
+ Reference ligand: reference.mol2
+ Active site radius: 6.5
+ MolDB/Index: craiselib
+ Solutions: out
+ No pharmacophoric constraints used
+ Protonation and tautomeric states disabled
+ Logging level: 1

Receptor:
+ Active atoms: 122
+ Accessible atoms: 78
+ Directed interactions: 53
+ Undirected interactions: 40
+ Query descriptors: 44344

> Loading index ...

Index:
+ cRAISE version: 1.1.0
+ Molecules: 2500
+ Discarded molecules: 0
```

```
+ Conformers: 601133
+ Conformer level: 100
+ Max. number of conformers: 250
+ Descriptors: 64812326
+ States: no

> Screening ...

Times:
+ Screening: 13006s wall, 12537s user + 92s system = 12628s CPU (97.1%)
+ Receptor preparation: 14s wall, 14s user + 0s system = 14s CPU (99.6%)
+ Total: 13020s wall, 12551s user + 92s system = 12642s CPU (97.1%)

Poses:
+ Property fulfilling matches: 3218351
+ Non-clashing poses: 197060
+ In convex hull poses: 190705
+ Pharm fulfilling poses: 190705
+ Atom scored poses: 178891
```

C.3.4 Konfiguration des virtuellen Screenings

Zustandsgenerierung Sollen Protonierungszustände und Tautomere während des virtuellen Screenings berücksichtigt werden, muss die zuvor erzeugte cRAISE-Bibliothek wie in Abschnitt C.3.2 beschrieben mit Zuständen angereichert worden sein. In diesem Fall berücksichtigt das Screening zusätzlich auch mögliche Zustandsänderungen des Proteins. Sollen keine Zustände während des Screenings berücksichtigt werden, so darf die zuvor erzeugte cRAISE-Bibliothek nicht mit Zuständen angereichert worden sein.

Definition der aktiven Bindetasche (-a) Mit der Option `-a <arg>` wird der aktive Radius zur Bindetaschendefinition spezifiziert. Anhand des Radius werden Proteinschweratome, deren Zentren nicht weiter als `<arg>` von einem Schweratomzentrum des Referenzliganden entfernt liegen, als Bestandteil der aktiven Bindetasche betrachtet. Standardmäßig ist ein Radius von 6,5 Å voreingestellt.

Filtern von Molekülen (-p) Mit der Option `-p <arg>` wird ein zuvor durch Mona definiertes Molekülprofil spezifiziert. Es wird während des Screenings dazu verwendet entsprechende Moleküle auszusortieren. Vergleiche hierzu Abschnitt C.3.7.

Screening mit einem Pharmakophormodell (-c) Mit der Option `-c <arg>` wird ein zuvor definiertes Pharmakophormodell spezifiziert. Es wird während des Screenings dazu verwendet entsprechende Posen auf Erfüllung zu überprüfen und bei Verletzung zu verwerfen. Vergleiche hierzu Abschnitt C.3.8.

Postoptimierung und Re-Scoring (-o) Mit dem Schalter `-o` wird cRAISE mitgeteilt, die besten Posen der Moleküle weiter zu optimieren und erneut zu bewerten. *Per se* werden die besten 32 Posen für jedes Molekül optimiert und bewertet. Diese Option führt zu einer erhöhten Laufzeit des Screening-Prozesses.

Spezifizierung des Log-Levels (-l) Analog zu C.3.2.

C.3.5 Auswertung eines virtuellen Screenings

Nach erfolgreich durchgeführtem Screening sind alle relevanten Ergebnisse (die beste Pose eines Hits und zugehörige Bewertungen) in der Lösungsdatenbank registriert. Der Auswertungsmodus (`-E`) des Screening-Werkzeugs bietet verschiedene Möglichkeiten auf Informationen aus der Lösungsdatenbank(en) `<in>` zuzugreifen. Mit

```
craiseScreener -E  [-h|d|j <in>] -O <out> [options]
```

startet man den Auswertungsmodus von cRAISE. Dabei muss die Option zur Generierung der Hitliste (-h), zum Exportieren von Posen (-d) oder zum Vereinigen von Teillösungen (-j) gewählt sein.

Generierung einer Hitliste (-h) Im Normalfall möchte man zunächst eine Hitliste erstellen, die die besten Hits entsprechend der Bewertungsfunktion sortiert angibt. Die Hitliste für eine Lösungsdatenbank `<sol.db>` wird mit

```
craiseScreener -E -h <sol.db> -O <out>
```

erzeugt. Die Hitliste ist im Anschluss in Form einer kommaseparierten Datei namens `<basename>.csv` im angegeben Ausgabeverzeichnis `<out>` zu finden:

Detaillierte Aufführung C.7: Beispiel einer Hitliste

```
craiseAtom, craiseGrid, MolName, MolID, ConfID, SplitID
-34.2402, -21.61, ZINC65553284, 435, 107553, {a289bcd7-ee74-4a25-bf0 ...
-33.2108, -18.6173, ZINC78490416, 2111, 509992, {a289bcd7-ee74-4a25- ...
-32.2175, -18.606, ZINC16649065, 884, 214927, {a289bcd7-ee74-4a25-bf ...
-32.159, -17.6293, ZINC40378879, 1197, 290573, {a289bcd7-ee74-4a25-b ...
-32.0826, -17.3311, ZINC70817879, 2261, 545894, {a289bcd7-ee74-4a25- ...
-31.9839, -17.2316, ZINC76363418, 970, 235279, {a289bcd7-ee74-4a25-b ...
-31.6003, -16.9997, ZINC48468882, 1327, 321355, {a289bcd7-ee74-4a25- ...
-31.5223, -17.7572, ZINC36016349, 937, 227712, {a289bcd7-ee74-4a25-b ...
...
```

Die Einträge in der Hitliste sind in Tabelle C.5 beschrieben.

Tabelle C.5: Einträge der Hitliste

Eintrag	Beschreibung
craiseAtom	Atombasierte Bewertung des Hits, die das Ranking bestimmt
craiseGrid	Frühe, gitterbasierte Bewertung der Pose
MolName	Molekülname
MolID	Molekül-ID in der Moleküldatenbank
ConfID	Konformer-ID in der Moleküldatenbank
SplitID	ID der cRAISE-Bibliothek

Exportieren von Posen (-d) Aus der Lösungsdatenbank lassen sich der vorhergesagte Bindungsmodus der Hits exportieren. Sie werden als SDF-Datei `<basename>.sdf` mit

```
craiseScreener -E -d <sol.db> -O <out>
```

im Ausgabeverzeichnis `<out>` erzeugt.

Vereinigen von Screening-Teillösungen (-j) Wird ein Screening parallelisiert auf einer partitionierten cRAISE-Bibliothek durchgeführt, so ist es notwendig, die erhaltenen Teilergebnisse (in Form mehrerer Lösungsdatenbanken) zu vereinen, um objektiv die besten Lösungen für den gesamten Screening-Lauf zu ermitteln. Dafür müssen die Lösungsdatenbanken gemeinsam in einem Verzeichnis `<solutions>` gesammelt oder verlinkt vorliegen.

```
craiseScreener -E -j <solutions> -O <out>
```

akkumuliert die Screening-Lösungen mit der Dateiendung `*sol.db` aus dem Verzeichnis `<solutions>` und erzeugt eine neue Lösungsdatenbank in `<out>`, die die vereinte Lösungsmenge enthält.

C.3.6 Konfiguration der Auswertung

Beschränkung der Lösungen (-n) Mit der Option `-n <arg>` wird die Anzahl der besten Hits, die in die Hitliste geschrieben bzw. die in das SDF-Format exportiert werden sollen beschränkt. Standardmäßig listet bzw. exportiert cRAISE die besten 1000 Hits.

Spezifizierung des Basisnamens von Ausgabedateien (-b) Mit der Option `-b <arg>` kann ein Basisnamen für erzeugte Hitlisten, exportierte Posen bzw. akkumulierte Lösungsdatenbanken spezifiziert werden.

Spezifizierung des Log-Levels (-l) Analog zu C.3.2.

C.3.7 Definition von Molekülprofilen

Molekülprofile ermöglichen den von cRAISE berechneten Index statisch zu verwalten und ohne erneute Präparierung, ein virtuelles Screening auf einer eingeschränkten Molekülbibliothek durchzuführen. cRAISE unterstützt die von Mona erzeugten Filterregeln und das darüber exportierte XML-Format.

Detaillierte Aufführung C.8: Beispiel für die Definition molekularer Filter

```
<!DOCTYPE MonaFilterPresets>
<presetlist version="5">
 <preset name="new preset">
  <filterchain tolerance="0">
   <filter inverted="0" type="0">
    <property id="2" min="1" max="2"/>
   </filter>
   <filter inverted="0" type="0">
    <property id="1" min="4" max="5"/>
   </filter>
  </filterchain>
 </preset>
</presetlist>
```

Mona ist derzeit unter `http://www.zbh.uni-hamburg.de/mona` für den akademischen Gebrauch frei verfügbar. Für die durch Mona und cRAISE unterstützten Filterarten sei an dieser Stelle auf die Originalpublikation[314] verwiesen.

C.3.8 Generierung eines Pharmakophormodells

Zur Definition von Pharmakophormodellen stellt das cRAISE-Software-Paket den *craisePharmGenerator* bereit. Tippt man in der Eingabeaufforderung unter dem `craise/`-Verzeichnis

```
craisePharmGenerator
```

dann listet das Programm alle möglichen Parameter auf, die vom Pharmakophorgenerierungs-Werkzeug berücksichtigt werden können:

Detaillierte Aufführung C.9: PharmGenerator Optionen

```
user@host:/local/user/craise> craisePharmGenerator
cRAISE (alpha) (Built: Sep 1 2013 19:04:12)
+ Version: 1.0.0

Usage:
craisePharmGenerator -T <arg> -R <arg> -P <arg> [options]

Generator options:
-T [ --target ] arg          Target protein (suffix required)
-R [ --refligs ] arg         Reference ligand(s) (suffix required)
-P [ --pharmfile ] arg       Pharm xml file
-a [ --activeradius ] arg (=6.5) Active site radius
-s [ --states ]              Recognize protonation and tautomeric states
-r [ --read ]                Read and draw pharm only
-e [ --essentials ] arg (=1) Absolute number of essential inclusions
-f [ --fraction ] arg        Fraction of essential inclusions (0.0 - 1.0)
-v [ --verbosity ] arg (=0)  Verbosity: 0 (quiet) - 5 (steps)
--license arg                Provide new license
```

Dieses Werkzeug dient nicht zur Pharmakophormodellierung, vielmehr bietet es dem Anwender ein Hilfsmittel zur Erstellung eines Modells für den cRAISE-Gebrauch. Anhand ein oder mehrerer gegebener Referenzmoleküle `<references>`, die zuvor vom Anwender am Zielprotein `<target>` entsprechend ausgerichtet worden sind, bestimmt das Werkzeug ein strukturbasiertes Pharmakophormodell <pharm>:

```
craisePharmGenerator -T <target> -R <references> -P <pharm>
```

Die erzeugte Pharmakophordefinition in Form einer XML-Datei kann durch den Anwender weiter editiert werden.

Detaillierte Aufführung C.10: Beispiel einer Pharmakophordefinition

```
<!DOCTYPE cRAISE_pharm_definition>
<pharm_definition>
 <inclusion_volumes>
  <volume>
   <sphere type="donor" radius="1.0" \
        center_x="31.3" center_y="-5.3" center_z="-6.4"/>
   <direction direction_x="-2.5" direction_y="-0.9" direction_z="0.5"/>
  </volume>
  <volume>
   <sphere type="acceptor" radius="1.65" \
        center_x="30.4" center_y="-8.7" center_z="-7.5"/>
   <direction direction_x="-1.6" direction_y="-1.9" direction_z="1.1"/>
  </volume>
  <volume>
   <sphere type="hydrophil" radius="1.6" \
        center_x="32.6" center_y="-3.5" center_z="-5.3"/>
   <direction direction_x="0.2" direction_y="2.7" direction_z="-0.6"/>
  </volume>
  <volume>
   <sphere type="hydrophob" radius="1.0" \
        center_x="34.6" center_y="-6.6" center_z="-13.1"/>
   <direction direction_x="0" direction_y="0" direction_z="0"/>
  </volume>
  <volume>
   <sphere type="any" radius="1.0" \
        center_x="33.4" center_y="-4.2" center_z="-13.1"/>
   <direction direction_x="0" direction_y="0" direction_z="0"/>
  </volume>
 </inclusion_volumes>
 <exclusion_volumes>
  <volume>
```

```
  <sphere type="donor" radius="1.0" \
       center_x="31.2" center_y="-8.9" center_z="-8.5"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="acceptor" radius="1.0" \
       center_x="30.5" center_y="-9.5" center_z="-7.5"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="hydrophil" radius="1.6" \
       center_x="33.3" center_y="-3.4" center_z="-4.6"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="hydrophob" radius="1.0" \
       center_x="33.4" center_y="-10.2" center_z="-11.9"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
 <volume>
  <sphere type="any" radius="1.0" \
       center_x="32.2" center_y="-3.0" center_z="-13.1"/>
  <direction direction_x="0" direction_y="0" direction_z="0"/>
 </volume>
</exclusion_volumes>
<nof_essentials>3</nof_essentials>
</pharm_definition>
```

C.3.9 Konfiguration der Pharmakophorgenerierung

Definition der aktiven Bindetasche (-a) Analog zu C.3.4. Die Bindetasche wird anhand des ersten gegebenen Referenzmoleküls bestimmt.

Berücksichtigung von Zuständen (-s) Mit der Option `-s` werden mögliche Protonierungszustände und Tautomere der prozessierten Moleküle berücksichtigt und für ihre Multizustandsatome hydrophile Inklusionen erzeugt.

Visualisierung eines Pharmakophormodells (-r) Mit der Option `-r` wird das angegebene Pharmakophor gelesen und durch FlexV dargestellt. Es wird keine Pharmakophordefinition erzeugt.

Spezifizierung der Anzahl essentieller Inklusionen (-e, -f) Mit der Option `-e <arg>` wird die Anzahl der Inklusionen gesetzt, die zur Erfüllung des gesamten Pharmakophormodells zeitgleich erfüllt werden müssen. Mit `-f <arg>` kann alternativ der Anteil der Inklusionen, die erfüllt werden müssen spezifiziert werden.

Spezifizierung des Ausgabelevels (-v) Grundsätzlich wird die Ausgabe des *craisePharmGenerator*s auf die Standardausgabe umgeleitet. Zur Fehlerdiagnose kann die Ausgabe bis zu einem Wert von 5 erhöht werden. Zudem weist ein Wert von 5 an, dass die Eingabemoleküle und das erzeugte/gelesene Pharmakophor mit FlexV visualisiert werden sollen.

C.3.10 Durchführung eines Protein-Ligand-Dockings

Für interessante Hits, die im virtuellen Screening erhalten wurden, kann ein Protein-Ligand-Docking durchgeführt werden. Dies ermöglicht den von cRAISE vorhergesagten Bindungsmodus weiter zu analysieren und zusätzlich zur besten Pose, die im Screening bestimmt wird, alternativ vorhergesagte Bindungsmodi zu betrachten. Tippt man in der Eingabeaufforderung unter dem `craise/`-Verzeichnis

```
craiseDocker
```

dann listet das Programm alle möglichen Parameter, die vom Docking-Werkzeug berücksichtigt werden können:

Detaillierte Aufführung C.11: Docker Optionen

```
user@host:/local/user/craise> craiseDocker
cRAISE (alpha) (Built: May 31 2015 12:09:20)
+ Version 1.0.0

Usage:
craiseDocker -T <arg> -R <arg> -M <arg> [options]

Docking options:
-T [ --target ] arg         Target protein (suffix required)
-R [ --reflig ] arg         Reference ligand (suffix required)
```

```
-M [ --molecule ] arg       Molecule to dock (suffix required)
-a [ --activeradius ] arg (=6.5) Active site radius
-C [ --pharm ] arg          Constrain queries with pharmacophore
-p [ --poses ] arg          Poses out basename
-c [ --confgen ] arg (=0)   Generate conformers. Quality: 0, 1, 2, 3, 100
-s [ --states ]             Generate tautomeric and protonation states
-o [ --optimize ]           Optimization and re-scoring of best poses
-r [ --rmsd ]               Calculate rmsd
-v [ --verbosity ] arg (=0) Verbosity: 0 (quiet) - 5 (steps)
--license arg               Provide new license
```

Mittels

```
craiseDocker -T <target> -R <reference> -M <molecule>
```

wird das durch `<molecule>` spezifizierte Molekül in die aktive Bindetasche, die im Protein `<target>` über den Referenzligand `<reference>` definiert ist, platziert (vergl. C.3.3). Auf der Standardausgabe werden daraufhin die Bewertung der Posen auf Basis der verwendeten Bewertungsfunktion gelistet:

Detaillierte Aufführung C.12: Docker Ausgabe

```
Rank,   Name,   ConfID,    Score, craiseAtom, craiseGrid, craiseOpt, Rmsd
 1.,    1g9v,       0,   -29.787,    -29.787,   -20.613,   0.000, 1.140
 2.,    1g9v,       0,   -29.773,    -29.773,   -20.872,   0.000, 1.045
 3.,    1g9v,       0,   -29.330,    -29.330,   -19.262,   0.000, 1.290
 4.,    1g9v,       0,   -28.864,    -28.864,   -18.806,   0.000, 1.266
 5.,    1g9v,       0,   -28.853,    -28.853,   -19.756,   0.000, 1.090
 6.,    1g9v,       0,   -28.396,    -28.396,   -19.166,   0.000, 0.752
 7.,    1g9v,       0,   -28.191,    -28.191,   -18.669,   0.000, 0.455
 8.,    1g9v,       0,   -28.183,    -28.183,   -18.913,   0.000, 1.126
 9.,    1g9v,       0,   -27.851,    -27.851,   -18.737,   0.000, 0.886
10.,    1g9v,       0,   -27.822,    -27.822,   -18.753,   0.000, 0.637
...
```

Die Einträge in der Vorhersage sind in Tabelle C.6 beschrieben.

Konfiguration des Protein-Ligand-Dockings

Docking mit Pharmakophoren (-C) Mit der Option `-C <arg>` wird ein zuvor definiertes Pharmakophor spezifiziert. Dieser wird während des Dockings dazu verwen-

Tabelle C.6: Einträge des Docking-Resultats

Eintrag	Beschreibung
Rank	Rang der Pose
ConfID	Konformer-ID in der Molekül datenbank
Score	Finale Bewertung der Pose, die das Ranking bestimmt
craiseAtom	Atombasierte Bewertung der Pose
craiseGrid	Frühe, gitterbasierte Bewertung der Pose
craiseOpt	Bewertung der Pose nach Optimierung (optional)
Rmsd	RMSD der Pose zum Referenzmolekül

det entsprechende Posen auf Erfüllung zu überprüfen und bei Verletzung zu verwerfen. Vergleiche hierzu C.3.8.

Konformergenerierung (-c) Durch Angabe von `-c <arg>` werden Konformere des Eingabemoleküls erzeugt und platziert. `<arg>` spezifiziert dabei die Qualität der erzeugten Konformere (vergl. C.3.2).

Zustandsgenerierung (-s) Durch Einstellen des `-s`-Schalters werden intern Protonierungszustände und Tautomere des Proteins und der prozessierten Moleküle generiert und dem platziertem Molekül ermöglicht seinen Zustand zu ändern (vergl. C.3.2 und C.3.4).

Postoptimierung und Re-Scoring (-o) Mit der Option `-o` werden die besten Posen optimiert und erneut bewertet (vergl. C.3.4).

RMSD-Berechnung (-t) Soll ein Re-Docking durchgeführt werden, so müssen sowohl `<reference>` als auch `<molecule>` das selbe Molekül repräsentieren. In diesem Fall kann mit der `-t`-Option für jede Pose ein RMSD bezüglich der Referenz berechnet werden.

Exportieren von Posen (-p) Mit der Option `-p <basename>` werden die von cRAISE erzeugten Posen ins SDF-Format exportiert.

Spezifizierung des Ausgabelevels (-v) Analog zu C.3.9.

D Veröffentlichungen

D.1 Peer-Review Publikationen

1. Henzler A M, Urbaczek S, Bietz S, Rarey M. Consistent handling of tautomers and protonation states in structure-based virtual screening. (2015) *in Vorbereitung*

2. Henzler A M, Urbaczek S, Hilbig M, Rarey M. An integrated approach to knowledge-driven structure-based virtual screening. *Journal of Computer-Aided Molecular Design*, 28(9):927–939, Sep (2014).

3. Schomburg K T, Bietz S, Briem H, Henzler A M, Urbaczek S, Rarey M. Facing the challenges of structure-based target prediction by inverse virtual screening. *Journal of Chemical Information and Modeling*, 54(6):1676-1686, Jun (2014).

4. Ehrlich H C, Henzler A M, Rarey M. Searching for recursively defined generic chemical patterns in non-enumerated fragment spaces. *Journal of Chemical Information and Modeling*, 53(7):1676–1688, Jul (2013).

5. von Behren M, Volkamer A, Henzler A M, Schomburg K T, Urbaczek S, Rarey M. Fast protein binding site comparison via an index-based screening technology. *Journal of Chemical Information and Modeling*, 53(2):411–422, Feb (2013).

D.2 Buchkapitel und systematische Übersichtsarbeiten

1. Henzler A M, Rarey M. 8. Protein flexibility in structure-based virtual screening: from models to algorithms. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines* (Sotriffer C, ed.), pp. 223–244. Wiley-VCH, May (2011).

2. Henzler A M, Rarey M. In Pursuit of fully flexible protein-ligand docking: modeling the bilateral mechanism of binding. *Molecular Informatics*, 29(3):164–173, Mar (2010).

D.3 (Inter-)nationale Konferenzbeiträge

D.3.1 Vorträge

1. Henzler A M, Urbaczek S, Rarey M. Consistent handling of flexible interaction sites for efficient structure-based virtual screening. *19th EuroQSAR Knowledge Enabled Ligand Design, Wien, Österreich*, Aug (2012).

2. Henzler A M, Urbaczek S, Rarey M. A structure-based virtual screening approach considering varying molecular hydrogen distributions. *244th American Chemical Society National Meeting, Philadelphia, PA, USA*, Aug (2012).

D.3.2 Poster

1. Henzler A M, Urbaczek S, Schulz B, Rarey M. A flexible-hydrogen interaction model for protein-ligand docking. *7th German Conference on Cheminformatics, Goslar, Deutschland*, Nov (2011), *Journal of Cheminformatics*, 4(Suppl 1):P14, May (2012). *mit Posterpreis ausgezeichnet.*

2. Henzler A M, Rarey M. Introducing protein flexibility to index-driven structure-based screening. *CHI's 10th Annual Structure-based Drug Design, Cambridge, MA, USA*, Jun (2010).