Entwicklung von Enzymdesignalgorithmen mit flexibler Ligandenpositionierung

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

Dietmar Birzer

aus Hirschau

im Jahr 2012

Das Promotionsgesuch wurde eingereicht am:	09. Oktober 2012
Das Kolloquium fand statt am:	28. November 2012
Die Arbeit wurde angeleitet von:	apl. Prof. Dr. Rainer Merkl

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Christoph Oberprieler
Erstgutachter:	apl. Prof. Dr. Rainer Merkl
Zweitgutachter:	apl. Prof. Dr. Elmar Lang
Drittprüfer:	Prof. Dr. Dr. Hans Robert Kalbitzer

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

Abb.	Abbildung
Ala	Alanin
Asn	Asparagin
Asp	Aspartat
Arg	Arginin
CdRP	1-(o-Carboxyphenylamino)-1-desoxyribulose-5-phosphat
CSD	Cambridge Structural Database
Cys	Cystein
Gln	Glutamin
Glu	Glutamat
His	Histidin
HisA	ProFAR Isomerase
HisF	Synthase-Untereinheit der Imidazolglycerinphosphat-Synthase
IGP	Indol-3-glycerinphosphat
Ile	Isoleucin
Leu	Leucin
Lys	Lysin
Met	Methionin
MSA	Multiples Sequenzalignment
OMP	Orotidin-5'-Phosphat
*mt*PriA	PriA aus *Mycobacterium tuberculosis*
PDB	Protein Data Bank
Phe	Phenylalanin
PRA	Phosphoribosylanthranilat
PRFAR	N'-[(5'-Phosphoribulosyl)-formimino]-5-aminoimidazol-4-carboxamidribonukleotid
PriA	Phosphoribosylanthranilat Isomerase A
ProFAR	N'-[(5'-Phosphoribosyl)-formimino]-5-aminoimidazol-4-carboxamidribonukleotid
Pro	Prolin
rCdRP	reduzierte Form von 1-(o-carboxyphenylamino)-1-desoxyribulose-5-phosphat
Ser	Serin

sgn	Signumfunktion
Thr	Threonin
Trp	Tryptophan
*tm*HisA	HisA aus *Thermotoga maritima*
*tm*HisF	HisF aus *Thermotoga maritima*
*tm*TrpF	TrpF aus *Thermotoga maritima*
TrpA	α-Untereinheit der Tryptophansynthase
TrpC	Indolglycerolphosphat-Synthase
TrpF	Phosphoribosylanthranilat-Isomerase
Tyr	Tyrosin
Val	Valin

1 Kurzfassung

Enzyme sind hocheffiziente Biokatalysatoren, die ein breites Spektrum chemischer Reaktionen katalysieren. Diese Effizienz und die Tatsache dass Substrate in der Regel hochspezifisch umgesetzt werden, prädestiniert deren Einsatz in vielen Bereichen der Medizin und der Biotechnologie. Allerdings erfüllt nur ein kleiner Teil der wildtypischen Enzyme die speziellen Anforderungen industrieller Verfahren. Es besteht daher ein großer Bedarf nach Verfahren mit denen Enzymeigenschaften maßgeschneidert verändert werden können.

Parallel zu biochemischen Ansätzen wurden hierfür in den letzten Jahren neue Methoden des computergestützten Enzymdesigns entwickelt. Dazu gehört TransCent, das entwickelt wird um Enzymfunktionen vom nativen auf alternative Proteingerüste zu übertragen. Der Entwurf der Strukturmodelle durch TransCent zielt darauf ab, vier Bedingungen, die von fundamentaler Bedeutung für die Enzymkatalyse sind, bestmöglich zu erfüllen. Dies sind (1) die Wahrung der Proteinstabilität, (2) die Optimierung der Ligandenbindung, (3) das Einstellen des korrekten Protonierungszustandes der katalytischen Residuen und (4) das Ausbilden funktionsspezifischer Wechselwirkungen zwischen Enzym und Ligand.

TransCent findet einen Kompromiss zwischen den zum Teil widersprüchlichen Forderungen, die vier spezifische Softwaremodule zur Erfüllung jeweils einer Bedingung stellen. Drei der vier Module basieren auf den *State-of-the-Art*-Methoden Rosetta, PROPKA und DSX. Für die Modellierung funktionsspezifischer Wechselwirkungen wurde ein neues Konzept entwickelt, das auf wissensbasierten Potentialen beruht. Aufgrund dieses Ansatzes kann TransCent auch dann verwendet werden, wenn der Übergangszustand des Substrates und der genau enzymatische Mechanismus der betrachteten Funktion unbekannt sind. Diese Eigenschaft unterscheidet TransCent von den anderen Enzymdesignverfahren.

Im Rahmen der vorliegenden Arbeit wurde TransCent um ein Modul erweitert, das die flexible Positionierung von Liganden unterstützt. Somit werden Limitationen der ersten Programmversion aufgehoben und neue Anwendungsmöglichkeiten erschlossen. Das Modul verwendet den neu entwickelten TransLig-Algorithmus zur Suche nach geeigneten Ligandpositionen. Hierbei wird die Lage des Liganden zusammen mit der Enzymsequenz unter Verwendung eines *Simulated Annealing* Verfahrens optimiert.

Die Performanz des Moduls zur Ligandpositionierung und des überarbeiteten Designalgorithmus wurde anhand von Rekapitulationsexperimenten basierend auf einem erweiterten und repräsentativen Enzymdatensatz bestimmt. Es zeigte sich, dass mithilfe des neuen Moduls zuverlässig nativ-ähnliche Ligandpositionen gefunden werden, die zu wildtypähnlichen Designmodellen umgesetzt werden.

Um die Qualität der TransCent-Modelle einer experimentellen Überprüfung zugänglich zu machen, wurden fünf Modelle für die Übertragung der Phosphoribosylanthranilat-Isomerase-Aktivität auf das Strukturgerüst zweier Enzyme aus der Histidin-Biosynthese von *Thermotoga maritima* berechnet. Anhand eines dieser Beispiele wurde der Designprozess ausführlich beschrieben und das resultierende Strukturmodell detailliert analysiert. Im Labor von R. Sterner wurden die aus den Modellen resultierenden Proteine hergestellt,

gereinigt und experimentell auf Stabilität, Aktivität und Bindung getestet. Alle Proteine waren stabil und drei der fünf Proteine waren in der Lage den Liganden mit annähernd wildtypischer Affinität zu binden. Allerdings besaß kein Protein eine nachweisbare Enzymaktivität. Diese experimentellen Ergebnisse belegen, dass die hier vorgestellte Version von TransCent zumindest in der Lage ist, Ligandenbindestellen korrekt zu modellieren. Damit zeigt dieses Programm einen Weg auf, ein wichtiges Teilproblem des Enzymdesigns anzugehen, welches bisher nicht zufriedenstellend gelöst wurde.

2 Einleitung und Motivation

Das Studium der komplexen Vorgänge im Inneren einer lebenden Zelle ist eines der spannendsten Forschungsgebiete unserer Zeit. Proteine gehören dabei zu den wichtigsten Komponenten, welche aufgrund ihrer vielfältigen Formen und Eigenschaften ein erstaunlich breites Spektrum von Aufgaben übernehmen: In Form von Antikörpern sind Proteine beispielsweise ein wichtiger Teil des Immunsystems höherer Lebewesen. Proteine bilden ebenfalls den Hauptbestandteil von Spinnenseide, mit der Spinnen ihre Netze konstruieren. Wiederum andere Proteine sind an der Replikation von DNA-Molekülen beteiligt.

Im letzten der genannten Beispiele treten Vertreter einer besonderen Gruppe von Proteinen in Aktion, der Gruppe der Enzyme. Dabei handelt es sich um hocheffiziente Biokatalysatoren, welche eine herausragende Rolle im Stoffwechsel aller lebenden Organismen spielen, da die meisten biochemischen Reaktionen auf zellulärer Ebene ohne Enzyme viel zu langsam oder gar nicht ablaufen würden. Enzyme haben ebenfalls eine große Bedeutung bei der Regulierung zahlreicher Stoffwechselprozesse und sind an der Übertragung von Signalen im Inneren einer Zelle beteiligt.

Auch bei industriellen Anwendungen nehmen Enzyme aufgrund ihrer Eigenschaften einen immer höheren Stellenwert ein. Durch ihre hohe Spezifität verhindern sie, dass bei der Katalyse unerwünschte Nebenprodukte entstehen. Da Enzyme an die Umweltbedingungen ihres Ursprungsorganismus angepasst sind, arbeiten sie im Vergleich mit anderen Katalysatoren bei relativ niedrigen Temperaturen. Außerdem sind sie natürlich biologisch abbaubar. Dieser Vorteil birgt in sich aber gleichzeitig den Nachteil, dass die Lebensdauer von Enzymen oft sehr begrenzt ist. Es werden daher große Anstrengungen unternommen, um sowohl die Stabilität als auch die Aktivität der verwendeten Enzyme zu verbessern.

Mit den etablierten Methoden der „gerichteten Evolution“ und des „rationalen Designs“ werden auf der Basis natürlich evolvierter Enzyme eindrucksvolle Ergebnisse bei der Optimierung der Katalyse und der Anpassung der Enzyme für neue Substrate erzielt. Diese Verfahren stoßen aber an ihre Grenzen, wenn es um die Herstellung „maßgeschneiderter“ Enzyme geht [1], mit Funktionen, welche in der Natur nicht vorkommen. Weiterführende Bestrebungen gehen sogar bis hin zur Erzeugung künstlicher metabolischer Pfade [2, 3], welche in der pharmazeutischen Industrie, der Medizin oder bei der Herstellung alternativer Kraftstoffe Anwendung finden können.

Da gerichtete Evolution und andere Methoden hier nicht weiterhelfen, hat sich in den letzten Jahren als Alternative die Disziplin des computergestützten Enzymdesigns etabliert. Sie ist zwar quasi eben erst ihren Kinderschuhen entwachsen, die rasanten Fortschritte der letzten Jahre und die Entwicklung neuer Methoden eröffnen aber ein spannendes Feld neuartiger Möglichkeiten [4] und geben Grund zur Hoffnung, dass computergestützte Verfahren eine wertvolle Ergänzung zu den etablierten Techniken darstellen können [5]. Gleichzeitig wird mit jedem Enzymdesign unser Verständnis für die funktionalen Zusammenhänge und die grundlegenden Mechanismen der Enzymkatalyse einer kritischen Überprüfung unterzogen und es können daraus möglicherweise wertvolle Erkenntnisse über die Evolution natürlicher Enzyme gewonnen werden [6].

2.1 Enzyme

Der Begriff „Enzym“ ist ein Kunstwort, welches 1878 von Wilhelm Friedrich Kühne geprägt wurde und das bis dahin gebräuchliche „Ferment“ ablöste [7]. Die damit bezeichneten Polypeptidketten sind wie alle anderen Proteine aus den 20 proteinogenen Aminosäuren aufgebaut, deren Abfolge in der DNA-Sequenz des zugehörigen Gens kodiert ist. Ihre Wirkungsweise beruht auf der Fähigkeit den energetisch ungünstigen Übergangszustand eines Substrats zu stabilisieren, wodurch die Aktivierungsenergie für dessen Umsetzung abgesenkt wird. Das chemische Gleichgewicht der Reaktion wird durch das Enzym dabei nicht verändert, es beschleunigt lediglich dessen Einstellung. Manche Enzyme erreichen dabei erstaunliche Ratenbeschleunigungen gegenüber der unkatalysierten Reaktion [8]. Die OMP-Decarboxylase aus *Saccharomyces cerevisiae*, das leistungsfähigste bekannte Enzym, reduziert die Halbwertszeit der Decarboxylierung von OMP von 78 Mio. Jahre auf 18 ms, was einer Ratenbeschleunigung um den Faktor $k_{\text{cat}}/k_{\text{uncat}} = 1,4 \cdot 10^{17}$ entspricht [9].

2.1.1 Grundlagen der Enzymkatalyse

In der Regel sind nur wenige Aminosäurereste an der Enzymkatalyse beteiligt. Diese befinden sich im sogenannten „aktive Zentrum“. Dies ist der Bereich des Enzyms, wo ein oder mehrere Substratmoleküle gebunden werden und die chemische Reaktion stattfindet. Die Bildung des Enzym-Substrat-Komplexes erfolgt dabei nach dem „Schlüssel-Schloss-Prinzip“, welches bereits 1894 von Emil Fischer postuliert wurde [10]. Die Voraussetzung für eine spezifische Bindung ist somit, dass die Formen und Eigenschaften des Liganden und des aktiven Zentrums komplementär zueinander passen. Eine dynamische Erweiterung des Prinzips stellt das „Induced-Fit-Modell“ dar [11]. Dabei treten bei beiden Bindungspartnern konformationelle Änderungen auf, welche die Bindung des Enzyms an das Substrat erst ermöglichen.

Enzyme katalysieren ein breites Spektrum chemischer Reaktionen und werden je nach Art der Reaktion in sechs EC-Klassen eingeteilt [12]. Bei der Umsetzung des Substrats wird in der Regel eine der folgenden Strategien zur Stabilisierung des Übergangszustandes verfolgt:

- Durch die Bindung zweier Substrate in der passenden Ausrichtung und Konformation werden deren reaktive Gruppen in räumliche Nähe gebracht, so dass die Reaktion beschleunigt ablaufen kann.
- Der Übergangszustand des Substrats wird stärker gebunden als das Produkt bzw. das Substrat selbst, was effektiv in einer Stabilisierung des Übergangszustandes resultiert.
- Bei der allgemeinen Säure-Base-Katalyse agieren Residuen wie z.B. Aspartat oder Histidin als Säure oder Base, d.h. sie nehmen Protonen vom Substrat auf oder geben Protonen an das Substrat ab.
- Bei der kovalenten Katalyse gehen die Seitenketten nukleophiler Residuen eine temporäre kovalente Bindung mit dem Substrat ein und bilden dabei ein kurzlebiges Zwischenprodukt (Intermediat).

Viele Enzyme benötigen für ihre Tätigkeit Cofaktoren in Form von Coenzymen oder Metallionen, welche nicht kovalent im aktiven Zentrum gebunden sind und durch Wechselwirkungen mit dem Substrat die Katalyse unterstützen.

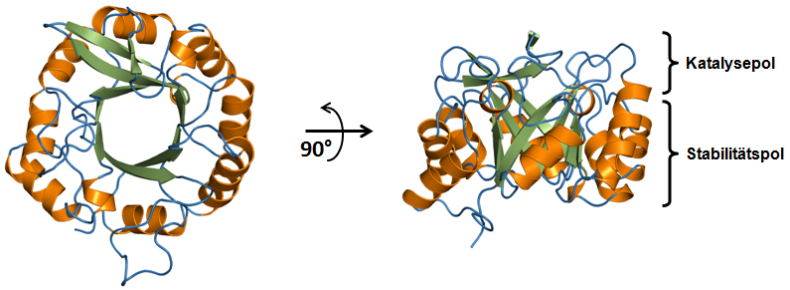


Abbildung 2.1: Der Faltungstyp eines $(\beta\alpha)_8$-Fasses
Als Beispiel für ein Enzym mit $(\beta\alpha)_8$-Fass-Faltungstyp ist die Struktur von HisF aus *Thermotoga maritima* (PDB-Code: 1thf) in Cartoon-Darstellung gezeigt. Die Färbung gibt die Einteilung in die Sekundärstrukturelemente α-Helix (orange), β-Faltblatt (grün) und ungeordnete Schleifenregionen (blau) wieder. In der Draufsicht links ist deutlich die zentrale fassartige Struktur zu erkennen, die von den acht β-Strängen gebildet wird. Die Seitenansicht rechts illustriert die räumliche Trennung der Bereiche, die für Stabilität (Stabilitätspol) und Funktionalität (Katalysepol) verantwortlich sind.

2.1.2 $(\beta\alpha)_8$-Fass-Enzyme

Proteine können aufgrund des Verlaufs der Hauptkette in sogenannte Faltungstypen aufgeteilt werden. Obwohl aufgrund der großen Genomsequenzierprojekte mittlerweile mehr als 26 Millionen Proteinsequenzen bekannt sind (UniProtKB/TrEMBL, *Release* 2012/09, [13]) gibt es nur eine erstaunlich kleine Anzahl von Faltungstypen, die diese Sequenzen annehmen. Die SCOP-Datenbank ([14], Version 1.75A), in der die bekannten Proteine anhand ihrer Struktur klassifiziert sind, umfasst aktuell 1194 unterschiedliche Faltungstypen. Schätzungen gehen davon aus, dass insgesamt etwa 4000 mögliche Topologien existieren, von denen ca. 2000 in Form natürlich vorkommender Proteine tatsächlich realisiert sind [15].

Der häufigste Faltungstyp bei Enzymen ist der des $(\beta\alpha)_8$-Fasses. Ungefähr 10% aller Enzyme gehören zu dieser Gruppe [16]. Es handelt sich dabei um globuläre Proteine mit einer Größe von etwa 250 Aminosäuren, welche in einem sich achtmal wiederholenden Motiv aus einem β-Strang und einer α-Helix angeordnet sind. Die β-Stränge lagern sich zu einem ringförmigen, parallelen β-Faltblatt zusammen, welches die zentrale, fassartige Struktur bildet (vgl. Abb. 2.1), welche von den α-Helices umgeben ist und dem Faltungstyp seinen Namen verleiht. Die Sekundärstrukturelemente sind jeweils durch Schleifenregionen miteinander verbunden.

Die hohe strukturelle Ähnlichkeit der TIM-Barrel-Enzyme, wie die Vertreter dieses Faltungstyps auch genannt werden, legt einen gemeinsamen evolutionären Ursprung der Proteine nahe [17, 18]. Der Name leitet sich von der Triosephosphat Isomerase (TIM) ab, bei der dieser Faltungstyp zum ersten Mal beobachtet wurde [19]. Außerdem lässt die Vielfalt der chemischen Reaktionen, welche von $(\beta\alpha)_8$-Fass-Enzymen katalysiert werden [20, 21], darauf schließen, dass es sich dabei um einen sehr alten Faltungstyp handelt. Gestützt wird diese Vermutung durch die Tatsache, dass TIM-Barrel-Enzyme nicht nur in fünf der sechs EC-Klassen vertreten sind, sondern auch in vielen der grundlegenden metabolischen Pfade wie z.B. der Aminosäuresynthese vorkommen [22]. Entstanden ist der Faltungstyp

vermutlich durch Genduplikation und Fusion zweier $(\beta\alpha)_4$-Halb-Barrel-Proteine [23, 24], welche ihrerseits aus Duplikations- und Fusionsereignissen von Viertel-Barrel-Enzymen hervorgegangen sein könnten [25, 26].

Eine besondere Eigenschaft der $(\beta\alpha)_8$-Fass-Enzyme ist, dass die Bereiche, welche die strukturelle Stabilität bzw. die Funktion des Enzyms ausmachen, räumlich voneinander getrennt sind [21]. Das aktive Zentrum befindet sich bei fast allen TIM-Barrel-Enzymen am C-terminalen Ende der β-Stränge (Katalysepol), wobei die katalytisch wichtigen Aminosäuren oft auch zu den βα-Schleifenregionen gehören. Der Rest des Proteins inklusive der αβ-Schleifen ist hingegen für die Stabilität der Struktur verantwortlich (Stabilitätspol, vgl. Abb. 2.1). Die weitgehende Entkopplung von Stabilität und Katalyse ist nicht nur eine der Ursachen, die den Faltungstyp aus evolutionärer Sicht so erfolgreich macht, sondern zudem auch der Grund, welcher TIM-Barrel-Proteine für Enzymdesignvorhaben so interessant macht. Die Trennung der beiden Aspekte ermöglicht, dass Aminosäureaustausche im aktiven Zentrum vorgenommen werden können, ohne die Stabilität zu beeinträchtigen.

2.2 Enzymdesign

Mit den heute zur Verfügung stehenden molekularbiologischen Verfahren ist es relativ einfach, die Basenreihenfolge eines Gens und somit auch die Aminosäurezusammensetzung des codierten Proteins zu manipulieren. Die Methoden der modernen Biotechnologie erlauben die künstliche Herstellung einzelner Gene bis hin zur Synthetisierung vollständiger Genome [27] und die Produktion der kodierten Enzyme in großen Mengen. Somit sind die technischen Voraussetzungen für das Design von Enzymen vorhanden. Diese Techniken werden dazu eingesetzt, um beispielsweise die Thermostabilität existierender Proteine zu erhöhen oder die Enantioselektivität von Enzymen zu verbessern.

2.2.1 Funktionsübertragungen

Ein weiteres Anwendungsgebiet ist die Übertragung von Enzymfunktionen zwischen verschiedenen Proteinen, im Folgenden „Funktionsübertragung“ genannt. Dabei wird versucht, durch den Austausch von Aminosäuren das Zielprotein so zu verändern, dass es die Reaktion des Vorlageenzyms katalysiert (siehe Abb. 2.2). Dies kann prinzipiell dadurch erreicht werden, dass zunächst über ein Sequenz- oder Strukturalignment eine Zuordnung zwischen den Proteinpositionen der beiden Enzyme hergestellt wird. Anschließend werden im Zielprotein sukzessive die wildtypischen gegen die entsprechenden Aminosäuren aus dem Vorlageenzym ausgetauscht und ggf. zusätzliche Residuen eingefügt bzw. gelöscht. Allerdings wäre der Informationsgewinn bei dieser Vorgehensweise eher zweifelhaft, denn im ungünstigsten Fall müssten alle Positionen ausgetauscht werden, um ein aktives Enzym zu erhalten. Daher muss die eigentliche Zielsetzung bei Funktionsübertragungen lauten, die minimale Anzahl an Änderungen zu identifizieren, die notwendig sind, um die gewünschte Funktion auf dem Zielgerüst zu etablieren.

Handelt es sich bei den beiden Proteinen um homologe Enzyme, welche durch divergente Evolution aus einem gemeinsamen Vorläufer hervorgegangen sind, so können anhand derartiger Experimente Erkenntnisse über die Entstehung der Vielfalt rezenter Enzyme gewonnen werden [28, 29]. Eine Analyse der Austausche, welche im Zielenzym eingeführt

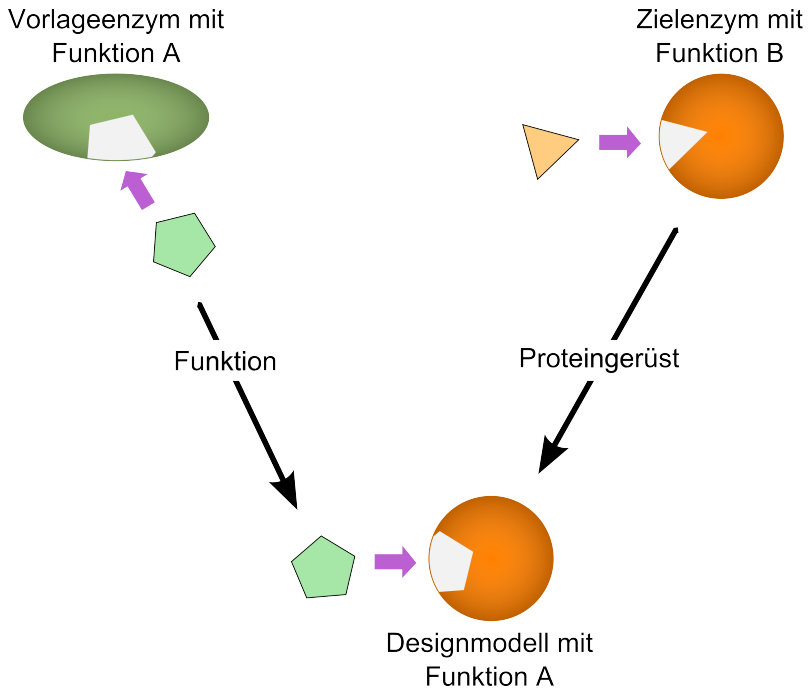


Abbildung 2.2: Schematische Darstellung einer Funktionsübertragung
Bei einer Funktionsübertragung soll die katalytische Aktivität des Vorlageenzyms A auf dem Proteingerüst des Zielenzyms B etabliert werden. Das Ziel ist hierbei, durch das Einführen geeigneter Mutationen Enzym B so zu verändern, dass es spezifisch das Substrat von Funktion A bindet und umsetzt. Im Allgemeinen geht bei einer Funktionsübertragung die ursprüngliche Funktion des Zielenzyms verloren.

werden, kann möglicherweise auch dabei helfen, den Reaktionsmechanismus des Vorlageenzyms aufzuklären. Da die Menge der Mutationen diejenigen Residuen umfassen muss, welche für die Aktivität und Spezifität des Enzyms verantwortlich sind, kann ein Vergleich mit den entsprechenden Positionen des Vorlageenzyms zumindest wertvolle Hinweise liefern.

2.2.2 Designmethoden

Die Strategien, welche zum Design von Enzymen verfolgt werden können, lassen sich grob in zwei Klassen unterteilen: gerichtete Evolution und rationales Design.

2.2.2.1 Gerichtete Evolution

Bei der gerichteten Evolution werden die Effekte der natürlichen Evolution künstlich im Labor nachgestellt, um Enzyme mit verbesserten bzw. veränderten Eigenschaften zu erzeugen. Durch ungenaues Kopieren mittels fehlerbehafteter Polymerasekettenreaktion (engl. *error-prone PCR* [30]) werden Varianten des ursprünglichen Enzyms erstellt, welche sich durch einen oder mehrere Austausche von diesem unterscheiden [31]. Die so erzeugten Mutationen sind zufällig über das gesamte Gen verteilt. Im anschließenden Selektions- oder Screening-Prozess werden dann in der erzeugten Genbank die Varianten identifiziert,

welche gegenüber dem wildtypischen Enzym eine Verbesserung zeigen. Durch wiederholte Anwendung des Verfahrens kann iterativ die Qualität der Resultate von Generation zu Generation verbessert werden. Der Vorteil dieser Methode besteht darin, dass weder die Struktur des Enzyms noch der genaue Mechanismus der Enzymfunktion bekannt sein müssen.

Dieses Verfahren kann dann erfolgreich angewandt werden, wenn nur eine kleine Anzahl von Mutationen notwendig ist, um eine messbare Verbesserung der betrachteten Eigenschaft zu erzielen. Durch Methoden wie DNA-*Shuffling* [32] kann die Zahl der Mutationen je Variante erhöht werden. Alternativ kann auch durch homologe Rekombination mehrerer natürlich vorkommender Enzyme versucht werden, verbesserte Varianten zu erzielen [33]. Die Größe der benötigten Genbanken macht die Verfahren der gerichteten Evolution allerdings sehr aufwändig und meist erweisen sie sich als ungeeignet für die Etablierung neuer Enzymfunktionen [34].

2.2.2.2 Rationales Design

Anders als bei der gerichteten Evolution wird beim rationalen Design nicht zufällig vorgegangen, sondern es werden z.B. anhand struktureller Informationen gezielt Proteinpositionen für Mutationsstudien ausgewählt. Dadurch wird die Anzahl möglicher Varianten dramatisch reduziert. Dennoch ergeben sich für die Totalrandomisierung von lediglich fünf Positionen $20^5 = 3,2 \cdot 10^6$ unterschiedliche Sequenzen. Genbanken dieser Größe können mit geeigneten Selektionssystemen bewältigt werden, Hochdurchsatz-Screening-Methoden stoßen hier allerdings bereits an ihre Grenzen.

Daher ist es notwendig, die Zahl der Varianten weiter zu reduzieren. Dies kann z.B. dadurch geschehen, dass nicht mehr alle 20 Aminosäuren an einer Position erlaubt sind, oder dass nur Einfach- und Doppelmutanten erzeugt werden. Beim CAST-Verfahren (*Combinatorial Active-site Saturation Test*, [35]) werden beispielsweise mehrere kleine Genbanken erstellt, bei denen jeweils nur zwei bis drei Positionen variiert werden. Die besten Varianten können dann miteinander kombiniert werden, um eine zusätzliche Steigerung der Aktivität oder Stabilität zu erhalten, sofern die erzielten Effekte voneinander unabhängig sind. Für die Verwendung der CAST-Methode muss allerdings die Position des aktiven Zentrums im Enzym bekannt sein.

Neben der Struktur kann auch ein Multiples Sequenzalignment (MSA) homologer Sequenzen als Informationsquelle für das rationale Design dienen. Beim Konsensus-Design [36] wird für jede Proteinposition diejenige Aminosäure gewählt, die in der entsprechenden Spalte des MSAs am häufigsten auftritt, was oft zu einer Stabilisierung des Enzyms führt [37]. Außerdem können durch den Vergleich mehrerer Sequenzen wichtige Positionen im Enzym identifiziert werden, welche sich für Mutationsstudien eignen. Auch Computeralgorithmen wie CLIPS-1D [38] können zur Auswertung von MSAs verwendet werden.

Da die biochemischen Grundlagen der Enzymkatalyse noch nicht vollständig verstanden sind, besteht beim rationalen Design allerdings immer die Gefahr, dass die Optimierung der Aktivität zu Lasten der Stabilität geht und umgekehrt [39]. Außerdem sind die Verfahren, welche beim rationalen Design eingesetzt werden, meist sehr zeitaufwändig und kostenintensiv.

Im Laufe der letzten Jahren konnte neben der gerichteten Evolution und dem rationalen Design eine neue Teildisziplin des Enzymdesigns etabliert werden, die eine vielversprechende Ergänzung zu den etablierten Verfahren darstellt: das computergestützte Enzymdesign.

2.2.3 Computergestütztes Enzymdesign

Enzyme und deren Substrate sind (Makro-)Moleküle, welche aus Atomen aufgebaut sind, die auf unterschiedliche Art und Weise miteinander interagieren. Die Molekülstrukturen und das komplexe Gefüge von Wechselwirkungen können im Rechner modelliert und bis zu einem gewissen Grad auch manipuliert werden. Es ist daher möglich, den Austausch von Aminosäuren mit dem Computer nachzuvollziehen und deren Auswirkungen zu berechnen. Die Methoden des computergestützten Enzymdesigns bilden somit eine kostengünstige, wenn auch nicht unbedingt weniger aufwändige Alternative zu molekularbiologischen Verfahren, welche neue Möglichkeiten wie z.B. die Verwendung *de novo* erzeugter Faltungstypen [40] bietet. Durch rasches „virtuelles Screening“ können mit einigen Methoden bis zu 10^{80} Sequenzen für ein Designexperiment untersucht werden [41]. Dies übersteigt den für experimentelle Ansätze zugänglichen Bereich um viele Größenordnungen.

2.2.3.1 QM/MM-Methoden

Für eine exakte Beschreibung der zugrunde liegenden physikalischen Gesetze ist prinzipiell eine *ab initio* quantenmechanische (QM) Behandlung des betrachteten Systems notwendig. QM-Algorithmen sind allerdings sehr rechenintensiv, weswegen bislang nur Systeme mit einer Größe von wenigen hundert Atomen behandelt werden können. Enzym-Ligand-Komplexe umfassen aber (inklusive Lösungsmittel) nicht selten weit mehr als 10000 Atome. Für das Design von Enzymen werden deswegen QM/MM-Hybrid-Verfahren verwendet [42], bei denen das betrachtete System in zwei Bereiche unterteilt wird. Die Teile des aktiven Zentrums, die direkt an der chemischen Reaktion beteiligt sind, werden wegen der höheren Genauigkeit quantenmechanisch behandelt. Für den Rest des aktiven Zentrums und das übrige Enzym werden aus Geschwindigkeitsgründen molekülmechanische (MM) Berechnungsmethoden angewendet.

QM/MM-Verfahren eignen sich hervorragend für die Untersuchung des katalytischen Mechanismus eines Enzyms [43], für die Bewertung der Auswirkung von Punktmutationen [44] und somit für die Optimierung der Aktivität eines Enzyms [45]. Aufgrund der benötigten Rechenleistung sind sie jedoch bisher zu aufwändig, um damit ganze aktive Zentren neu zu gestalten bzw. Enzymfunktionen zu übertragen, sondern dienen eher der Unterstützung des rationalen Designs [46].

2.2.3.2 Enzymdesign mittels MD-Simulationen

Etwas weniger rechenintensiv sind Verfahren, die auf Moleküldynamik-Simulationen (MD-Simulation) basieren. Dabei werden empirische Kraftfelder wie CHARMM (*Chemistry at Harvard Macromolecular Mechanics*, [47]), AMBER (*Assisted Model Building with Energy Refinement*, [48]) oder GROMOS (*Groningen Molecular Simulation*, [49]) eingesetzt, in denen sich die Atome des Enzyms, des Liganden und des Lösungsmittels gemäß den Gesetzen der Newton'schen Mechanik bewegen. Beim Enzymdesign werden typischerweise die Bewegungen der Atome für wenige Nanosekunden simuliert, um daraus Konfigurationsänderungen des Liganden, Umlagerungen von Seitenketten im aktiven Zentrum oder die Bildung bzw. das Aufbrechen von Wasserstoffbrücken ableiten zu können. Allerdings kann mit Hilfe von MD-Simulationen die tatsächliche chemische Umsetzung des Substrats nicht

modelliert werden. Dennoch sind sie dafür geeignet, um damit aktive von inaktiven Modellen zu unterscheiden [50]. Somit stellen MD-Simulationen eine nützliche Alternative zur zeitlich aufwändigeren labortechnischen Überprüfung der Designmodelle dar und können in Kombination mit schnelleren Verfahren beim Design eingesetzt werden [51].

2.2.3.3 Heuristische Verfahren für das Design von Enzymen

Da QM/MM-Verfahren und MD-Simulationen für viele Problemstellungen zu rechenintensiv sind, werden in den meisten Fällen stattdessen heuristische Verfahren zum Design von Enzymen verwendet. Die nachfolgend genannten heuristischen Methoden verfolgen zwar unterschiedliche Strategien, können aber prinzipiell in drei Komponenten eingeteilt werden: Modellierungseinheit, Energiefunktion und Optimierungsverfahren.

Modellierungseinheit: Die Aufgabe der Modellierungseinheit ist, Strukturmodelle mit atomistischer Auflösung für die betrachteten Sequenzen zu generieren. Bei den meisten Designverfahren wird das vorgegebene Proteinrückgrat starr gehalten. Somit beschränkt sich die Aufgabe der Modellierungseinheit darauf, die Seitenketten der einzelnen Residuen zu modellieren. Um die Anzahl möglicher Varianten zu reduzieren, werden dabei häufig Rotamerbibliotheken verwendet. Dies bedeutet, dass Aminosäureseitenketten keine beliebigen Konformationen annehmen dürfen, sondern auf einige wenige energetisch günstige „Rotamere“ beschränkt sind. Deren relative beobachtete Häufigkeiten sind zusammen mit den entsprechenden Diederwinkeln der drehbaren Bindungen in der Rotamerbibliothek abgelegt [52].

Energiefunktion: Die Energiefunktion dient zur Bewertung der von der Modellierungseinheit erzeugten Strukturmodelle. Die meisten Designmethoden verfügen über eine eigene Energiefunktion, wobei praktisch bei allen in irgendeiner Form Van-der-Waals Wechselwirkungen, elektrostatische Interaktionen, Wasserstoffbrücken und Solvatationseffekte berücksichtigen werden. Grundsätzlich können die existierenden Energiefunktionen in drei Gruppen eingeteilt werden:

Physikalisch motivierte Energiefunktionen basieren direkt auf den physikalischen Gesetzmäßigkeiten, welche für die Wechselwirkungen innerhalb eines Proteins und die Interaktionen mit dem umgebenden Lösungsmittel verantwortlich sind [53]. Obwohl damit relativ exakte Vorhersagen gemacht werden können, ist ihr Nachteil, dass sie aufwändig zu berechnen sind.

Empirische Energiefunktionen, wie sie bei MD-Simulationen zum Einsatz kommen, können auch für das Design von Enzymen adaptiert werden. Das Programm PROTDES [54] verwendet eine auf dem CHARMM-Kraftfeld basierende Energiefunktion, welche durch drei unterschiedliche Terme zur Behandlung des Lösungsmittels ergänzt wird.

Am besten für das Enzymdesign geeignet erscheinen jedoch Energiefunktionen, welche *wissensbasierte Potentiale* zur Bewertung der Strukturmodelle verwenden. Der Vorteil bei diesen Ansätzen ist, dass sie auch dann angewendet werden können, wenn die zugrunde liegenden Effekte noch nicht vollständig verstanden sind. Die Potentiale basieren stattdessen auf Häufigkeitsverteilungen von Eigenschaften (z.B. Bindungswinkel oder Abstände zwischen Atomen), die aus bekannten Strukturen abgeleitet werden. Motiviert wird dieses Vorgehen sowohl durch Ergebnisse der Testtheorie (Neyman-Pearson-Lemma, vgl. [55]) als

auch durch Konzepte aus der statistischen Physik (Invertierung des Boltzmann-Gesetzes, siehe [56]).

Die meisten Enzymdesignverfahren verwenden in ihren Energiefunktionen nur Ein- und Zwei-Körper-Energieterme, da diese einen Geschwindigkeitsvorteil bei der Suche nach der Struktur mit der besten Energie bieten.

Optimierungsverfahren: Mit der Energiefunktion können die von der Modellierungseinheit erzeugten Strukturmodelle bewertet werden. Die Aufgabe des Optimierungsverfahrens ist nun, im hochdimensionalen Raum möglicher Lösungen dasjenige Modell zu identifizieren, welches die Anforderungen der Energiefunktion am besten erfüllt. Betrachtet man den Designprozess als Optimierungsproblem, so entspricht dies der Suche nach dem globalen Minimum der Energiefunktion. Wie von Pierce *et al.* gezeigt werden konnte, ist dieses Optimierungsproblem NP-hart [57]. Um dennoch in angemessener Zeit eine gute Lösung zu finden, werden bei rotamerbasierten Designalgorithmen hauptsächlich zwei Klassen von Optimierungsverfahren eingesetzt.

Dies sind zum einen Methoden, die auf dem *Dead-End-Elimination* (DEE) Theorem beruhen [58, 59]. Dieses garantiert, dass mit einem DEE-Suchalgorithmus das globale Optimum einer Energiefunktion gefunden wird [60], wobei sich die Laufzeit gegenüber einer entsprechenden Aufzählmethode von $O(n^p)$ auf $O(n^3)$ reduziert (n =Anzahl der Proteinpositionen, p =Anzahl der Rotamere je Position). Der Algorithmus skaliert allerdings immer noch verhältnismäßig schlecht mit der Größe des Optimierungsproblems. Daher werden bei modernen Designprogrammen meist durch heuristische Elemente erweiterte Varianten des DEE-Theorems wie z.B. der FASTER-Algorithmus (*Fast and Accurate Side-Chain Topology and Energy Refinement*, [61, 62]) eingesetzt.

Beispiele für Designprogramme, welche DEE-Optimierungsverfahren verwenden, sind iMinDEE [60] und ORBIT (*Optimization of Rotamers By Iterative Techniques*, [53]). Mit Letzterem konnte erfolgreich die Bindungsspezifität des Proteins Calmodulin für seine natürlichen Bindungspartner verändert werden, wobei im Designprozess 10^{22} unterschiedliche Sequenzen berücksichtigt wurden.

Die zweite Klasse von Optimierungsverfahren, welche häufig beim Design von Enzymen eingesetzt wird, sind Monte-Carlo-Algorithmen wie z.B. *Simulated Annealing* [63]. Dabei handelt es sich um stochastische Näherungsverfahren, welche hinreichend gute Lösungen liefern, allerdings nicht garantieren können, dass es sich bei einer gefundenen Lösung um das globale Minimum der Energiefunktion handelt. Dafür erlaubt die hohe Geschwindigkeit des Optimierungsverfahrens eine mehrfache Wiederholung der Suche, so dass die Wahrscheinlichkeit für das Erreichen des globalen Minimums erhöht werden kann. Die genaue Vorgehensweise bei der Sequenzoptimierung mittels *Simulated Annealing* wird in Abschnitt 3.1.1 detailliert beschrieben.

Implementierungen von *Simulated Annealing*-Algorithmen finden sich z.B. in den Softwarebibliotheken SHARPEN (*Systematic Hierarchical Algorithms for Rotamers and Proteins on an Extended Network*, [64]) und EGAD (*EGAD: A Genetic Algorithm for protein Design*, [65]), welche für das computergestützte Design von Proteinen entwickelt wurden.

Rosetta Auch der Designalgorithmus von Rosetta [66], dem bislang erfolgreichsten Verfahren beim Design neuer Enzyme, verwendet ein *Simulated Annealing* Protokoll für die Optimierung von Designmodellen. Rosetta [67] stammt ursprünglich aus der Gruppe von

David Baker und wurde zunächst für die *de novo* Vorhersage von Proteinstrukturen entwickelt. Inzwischen umfasst die Rosetta Software Suite ein breites Spektrum an Methoden zur Manipulation und Modellierung von Proteinen und wird von den *RosettaCommons*, einem weltweiten Verbund von Arbeitsgruppen an mehreren Universitäten, kontinuierlich weiterentwickelt.

Zur Bewertung der Designmodelle verwendet der Enzymdesignalgorithmus von Rosetta eine semiempirische Energiefunktion bestehend aus mehreren Energietermen, welche sowohl die Wechselwirkungen innerhalb des Proteins (vgl. Tabelle 3.1) als auch die Interaktionen mit dem Ligandmolekül beschreiben (vgl. [68]).

Das Designverfahren konnte von Röthlisberger *et al.* dazu verwendet werden, Enzyme zu erzeugen, welche die Kemp-Eliminierungsreaktion katalysieren [69]. Dabei handelt es sich um das erste erfolgreiche *de novo* Design einer Enzymfunktion, welche bei natürlich evolvierten Enzymen nicht vorkommt. Weitere erfolgreiche Designexperimente unter Verwendung des Rosetta-Algorithmus umfassen die Etablierung der bimolekularen Diels-Alder-Reaktion [70] sowie das Design mehrerer aktiver Retro-Aldolasen [71]. Im letzteren Fall konnte die Enzymfunktion auf fünf verschiedenen Proteingerüsten mit zwei unterschiedlichen Faltungstypen eingeführt werden.

Allerdings ist die katalytische Aktivität aller erzeugten Varianten weit unter dem Niveau natürlich vorkommender Enzyme und es konnte im Fall der Kemp Eliminierung gezeigt werden, dass die anfangs schwache Aktivität durch gerichtete Evolution deutlich gesteigert werden kann [72, 73, 74]. Die experimentellen Ergebnisse weisen also darauf hin, dass es noch Spielraum für die Verbesserung der Designalgorithmen gibt und dass unter Umständen zusätzliche Eigenschaften oder Effekte mit einbezogen werden müssen, die bislang noch nicht ausreichend berücksichtigt sind.

Die wichtigste Eingabe für den Enzymdesignalgorithmus von Rosetta ist jeweils ein Modell für den Übergangszustand des Substrats, welches umgesetzt werden soll, zusammen mit den Seitenketten der katalytisch aktiven Residuen. Dieses sogenannte Theozym [75] bildet die Grundlage eines Designexperimentes. Die Aufgabe des Designalgorithmus ist es dann, das Theozym möglichst gut in ein passend gewähltes Proteingerüst einzubetten und die Wechselwirkungen zwischen dem Enzym und dem Modell des Übergangszustandes zu optimieren.

Eine Voraussetzung für die Anwendbarkeit des Verfahrens ist somit, dass der genaue Mechanismus der zu modellierenden Funktion inklusive der Konformation des Übergangszustandes bekannt sein muss. Für viele Reaktionen sind jedoch die strukturellen Grundlagen der Enzymkatalyse nicht ausreichend gut verstanden. Oft ist nicht einmal die genaue Konformation des Übergangszustandes bekannt. Daher ist in diesen Fällen die Festlegung der Optimierungskriterien schwierig.

2.2.4 TransCent

Um derartige Enzymfunktionen dennoch behandeln zu können, wurde von André Fischer in der Arbeitsgruppe von Rainer Merkl das Enzymdesignprogramm TransCent entwickelt [76], welches die Übertragung einer Funktionen von einem Vorlageenzym auf ein Zielprotein auch ohne die genaue Kenntnis des Übergangszustandes ermöglichen soll (vgl. Abb. 2.2). Für die Umsetzung diese Vorhabens berücksichtigt TransCent beim Design vier Bedingungen, welche von fundamentaler Bedeutung für die Enzymkatalyse sind: (1) die

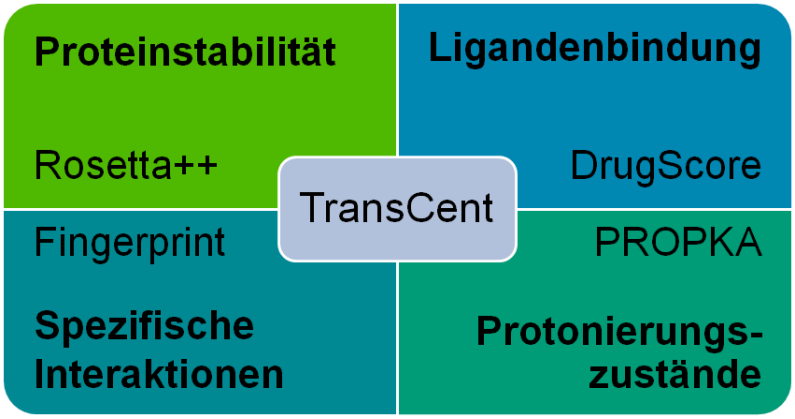


Abbildung 2.3: Schematische Darstellung des modularen Aufbaus von TransCent
Der TransCent-Designalgorithmus ist aus vier Modulen zur Optimierung von vier Nebenbedingungen aufgebaut. Drei der vier Module basieren auf den *State-of-the-Art*-Methoden Rosetta++ bzw. EGAD, DrugScore und PROPKA. Für die Berücksichtigung funktionsspezifischer Wechselwirkungen wurde das Fingerprint-Modul entwickelt.

Wahrung der Proteinstabilität, (2) die Optimierung der Ligandenbindung, (3) das korrekte Einstellen des Protonierungszustandes der katalytischen Residuen und (4) die Ausprägung funktionsspezifischer Wechselwirkungen zwischen Enzym und Ligand.

Der Aufbau von TransCent ist modular (siehe Abb. 2.3), d.h. jedes der vier Optimierungskriterien wird durch ein eigenes Modul erfasst und verarbeitet. Das Programm verwendet für die Behandlung der ersten drei Bedingungen jeweils Methoden, die den letzten Stand der Technik für die betrachtete Problemstellung repräsentieren. Das vierte Modul wurde speziell für TransCent entwickelt und basiert auf der Ableitung und Verwendung wissensbasierter Potentiale.

Die Forderungen, welche durch die einzelnen Module an die Designmodelle gestellt werden, können nicht unabhängig voneinander betrachtet werden und widersprechen sich teilweise sogar. Daher werden die durch die Module beschriebenen Effekte für die Optimierung zu einer einzigen Energiefunktion zusammengefasst und mittels Gewichtungsfaktoren ausbalanciert. Damit kommt in der Energiefunktion der Widerspruch zwischen Funktionalität und Stabilität zum Ausdruck, der auch für natürlich vorkommende Enzyme gilt [77]. In der Natur sorgen die Mechanismen der Evolution dafür, dass die Aminosäuresequenzen der Enzyme einen optimalen Kompromiss zwischen den beiden Eigenschaften darstellen. Bei TransCent ist es Aufgabe des *Simulated Annealing* Optimierungsverfahrens, Sequenzen zu identifizieren, welche möglichst gut die Vorgaben der Energiefunktion erfüllen. Da TransCent ein rotamerbasiertes Verfahren ist und das Rückgrat des Zielproteins beim Design starr gehalten wird, reduziert sich das Optimierungsproblem darauf, die Rotamerkombination mit der niedrigsten Energie zu finden. Die *Simulated Annealing* Routine wurde genau wie die Modellierungseinheit von Rosetta übernommen. Somit verwendet TransCent bei der Modellierung der Seitenketten die rückgratabhängige Version der Dunbrack-Rotamerbibliothek [52].

Die Konformation des Ligandmoleküls wird unverändert aus der Kristallstruktur des Vorlageenzyms übernommen und durch Überlagerung der Vorlagestruktur mit dem Zielenzym

in das neue aktive Zentrum übertragen. Sowohl die Konformation als auch die Position des Liganden ist während der Optimierung fixiert. Diese Vorgehensweise dient der Reduzierung der Rechenzeit und basiert auf der Annahme, dass es sich bei der Ligandpose in der Kristallstruktur um eine aktive Ligandkonformation handelt.

2.2.4.1 Modul für Proteinstabilität

Alle Proteine und insbesondere auch Enzyme müssen korrekt und stabil in eine wohldefinierte dreidimensionale Struktur falten, um ihre Aufgaben erfüllen zu können. Die Energiefunktionen von Rosetta++ wurde für die Vorhersage stabiler Proteinstrukturen entwickelt und optimiert [40] und wird daher in TransCent für die Bewertung der Proteinstabilität verwendet.

2.2.4.2 Modul für Ligandenbindung

Die Fähigkeit zur Bindung eines Ligandmoleküls ist eine der Grundvoraussetzungen der Enzymkatalyse. Nicht umsonst wird die „Katalysekraft“ $\frac{k_{cat}}{K_M}$ eines Enzyms über das Verhältnis von Umsetzungsrate k_{cat} und Michaeliskonstante K_M definiert, wobei letztere eine Maß für die Bindungsaffinität ist. Das Modul für die Optimierung der Ligandenbindung basiert daher auf dem Programm DrugScore [78], welches wissensbasierte Potentiale verwendet, um Wechselwirkungen zwischen Protein und Ligandmolekül zu bewerten und Enzym-Ligand-Komplexe mit hoher Bindungsaffinität zu identifizieren. Die verwendete Version von DrugScore wurde speziell für TransCent angepasst, um den Beitrag einzelner Rotamere zur Ligandenbindung energetisch bewerten zu können.

2.2.4.3 Modul für pK_a-Wert-Optimierung

In einer Analyse des Catalytic Site Atlas [79] haben Bartlett *et al.* gezeigt, dass die sieben Aminosäuren mit titrierbaren Seitenketten ca. 75% aller katalytisch wichtigen Residuen ausmachen [80]. Andererseits werden speziell für katalytische Residuen häufig relativ stark verschobene pK_a-Werte beobachtet [81, 82]. Ein entscheidender Faktor bei der Enzymkatalyse ist der korrekte Protonierungszustand der Seitenketten im aktiven Zentrum. Oft sind elektrostatische Interaktionen wichtig für die Bindungsspezifität oder es müssen Ladungen des Übergangszustandes stabilisiert werden. Im Fall einer Säure-Base-Katalyse müssen die daran beteiligten Seitenketten im richtigen Protonierungszustand vorliegen, so dass sie vom Liganden Protonen aufnehmen bzw. an diesen abgeben können. Da sich der physiologische pH-Wert bei den meisten Organismen im Bereich um 7,5 bewegt, ist der pK_a-Wert eines Residuums die maßgebliche Größe für dessen Protonierungszustand.

Das Modul zur pK_a-Wert-Optimierung von TransCent basiert auf einer geschwindigkeitsoptimierten Version von PROPKA. Dieses Programm verwendet ein empirisches Modell zur Vorhersage von pK_a-Wert-Perturbationen und ist im Mittel bis auf eine pH-Einheit genau [83]. In TransCent werden mit PROPKA sowohl die Referenz-pK_a-Werte im Vorlageenzym als auch die pK_a-Werte der entsprechenden Residuen im Designmodell berechnet. Die Energiefunktion des Moduls ist so angelegt, dass Abweichungen von den vorgegebenen Werten energetisch bestraft werden. Da mehrere Residuen gleichzeitig Einfluss auf die Verschiebung eines pK_a-Wertes haben können, handelt es sich dabei um einen Mehr-Körper-Energieterm.

2.2.4.4 Fingerprint-Modul

Die strukturellen Grundlagen der Enzymkatalyse sind in vielen Fällen unklar oder nicht ausreichend gut verstanden, um genaue Vorgaben für das Design der jeweiligen Funktion machen zu können. Dennoch ist klar, dass der katalytische Mechanismus eines Enzyms auf sehr spezifischen Wechselwirkungen beruhen muss, welche charakteristisch für die betrachtete Funktion sind und in ihrer Gesamtheit wie ein „Fingerabdruck“ (engl. *fingerprint*) ein Enzym identifizieren. Um diesen strukturellen Fingerprint erfassen, beschreiben und beim Designprozess berücksichtigen zu können, verfügt TransCent über ein entsprechendes Modul.

Dieses analysiert die Struktur des Protein-Ligand-Komplexes des Vorlageenzyms und erkennt polare Wechselwirkungen (v.a. Wasserstoffbrücken) zwischen den Seitenketten des aktiven Zentrums und dem Ligandmolekül. Um sicherzustellen, dass es sich bei den ermittelten Wechselwirkungen um solche handelt, die für die Enzymfunktion charakteristisch und wichtig sind, wird bei der Ableitung des Fingerprints nicht nur die Struktur des Vorlageenzyms sondern eine Strukturbibliothek homologer Enzyme verwendet. Da für eine statistisch belastbare Aussage i. A. bislang zu wenige Kristallstrukturen für Enzyme mit einer bestimmten Funktion bekannt sind, werden stattdessen Homologiemodelle verwendet, welche mit Hilfe des Programms MODELLER [84] erstellt werden.

Aus der räumlichen Verteilung der Wechselwirkungen und den beobachteten Aminosäurehäufigkeiten werden wissensbasierte Potentiale abgeleitet [56], mit denen beim Design die Rotamere im aktiven Zentrum des Zielenzyms energetisch bewertet werden. Durch eine geschickte Kombination der Energiewerte bei der Designoptimierung wird sichergestellt, dass jeweils nur ein Residuum ein Potential erfüllen kann.

Bei einer Funktionsübertragung sorgt das Fingerprint-Modul dafür, dass im neuen aktiven Zentrum Seitenketten in ähnlicher Form und Orientierung wie im Vorlageenzym zur Ausbildung der wichtigen Wechselwirkungen modelliert werden. In [85] wurde gezeigt, dass das Fingerprint-Modul entscheidend zur Qualität der mit TransCent erstellten Modelle beiträgt.

2.2.4.5 Überprüfung der Vorhersagequalität der ersten TransCent-Version

Um den Designalgorithmus einem kritischen Test zu unterziehen, wurden für Funktionsübertragungen zwischen fünf Ribulosephosphat-bindenden $(\beta\alpha)_8$-Fass-Enzymen mit der ersten Version von TransCent Designmodelle erstellt und eine detaillierte Analyse durchgeführt. Eine der Varianten wurde darüber hinaus für eine labortechnische Überprüfung ausgewählt und von Bernd Reisinger am Lehrstuhl von Prof. Reinhard Sterner im Rahmen seiner Diplomarbeit [86] umgesetzt und getestet. Dabei handelte es sich um ein Designmodell für die Übertragung der von TrpF aus *Thermotoga maritima* katalysierten Phosphoribosylanthranilat-Isomerisierungsreaktion auf das Proteingerüst von HisF ebenfalls aus *T. maritima*. Für die Variante konnte jedoch keine Aktivität im messbaren Bereich festgestellt werden.

Eine Analyse der Ergebnisse wies zusammen mit den *in silico* Ergebnissen darauf hin, dass die in der ersten Version von TransCent verwendete Strategie zur Positionierung des Liganden im aktiven Zentrum zu ungenau und fehleranfällig ist, obwohl es sich bei Vorlage- und Zielenzym um Strukturen mit sehr großer struktureller Ähnlichkeit handelte. Aus

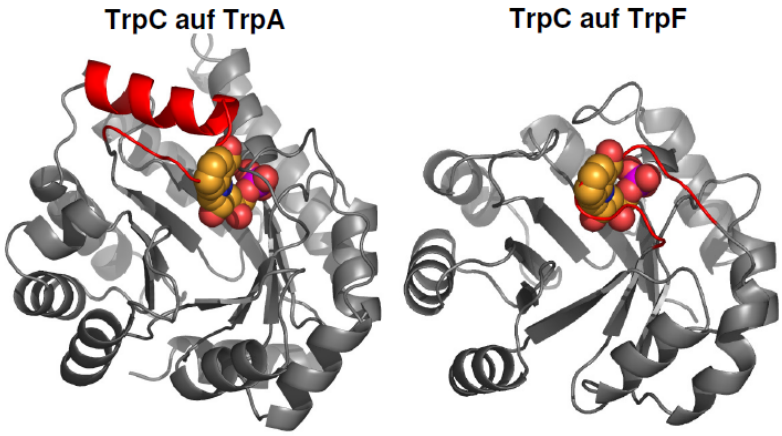


Abbildung 2.4: Superpositionsmethode: Sterische Kollisionen zwischen Ligand und Proteinrückgrat
Die Positionierung des Liganden im aktiven Zentrum des Zielproteins mit Hilfe der Superpositionsmethode führt häufig zu sterischen Kollisionen zwischen Ligand und Proteinrückgrat. Auch bei sehr großer struktureller Ähnlichkeit zwischen Vorlage- und Zielenzym können Unterschiede in den Schleifenregionen dazu führen, dass keine Funktionsübertragung mit TransCent möglich ist. Die Problematik wird illustriert anhand von Beispielen aus [85] für die Übertragung der Funktion von TrpC (PDB-Code: 1lbf) auf das Gerüst von TrpA (PDB-Code: 1qoq) bzw. TrpF (PDB-Code: 1lbm). Die Abbildung wurde aus [85] übernommen.

demselben Grund konnten für einige der Funktionsübertragungen überhaupt keine Modelle berechnet werden, da die mit der Superpositionsmethode ermittelten Ligandpositionen zu sterischen Kollisionen mit dem Proteinrückgrat geführt haben (vgl. Abb. 2.4).

Aus den gewonnenen Erkenntnissen wurde gefolgert, dass TransCent für die erfolgreiche Übertragung von Enzymfunktionen um ein Modul erweitert werden muss, das die flexible Positionierung von Liganden unterstützt.

2.3 Zielsetzung der vorliegenden Arbeit

Die flexible Behandlung der Ligandposition stellt eine konzeptionelle Erweiterung des TransCent-Algorithmus dar, für deren Realisierung drei Teilprobleme gelöst werden müssen.

Zunächst müssen im aktiven Zentrum des Zielenzyms passende Ligandpositionen identifiziert werden. Diese Aufgabenstellung ist schwieriger als die bereits relativ gut gelöste Problematik des Liganden-Docking, da das aktive Zentrum in dem der Ligand positioniert werden soll, erst im nachfolgenden Designprozess erstellt wird. Anstatt wie beim Docking die (wahrscheinliche) Bindungspose zu berechnen, müssen Ligandpositionen ermittelt werden, welche dem Designalgorithmus die Möglichkeit bieten, alle wichtigen Wechselwirkungen zwischen Enzym und Ligand zu modellieren. Der Ligand darf also weder zu weit weg noch zu nahe am Proteinrückgrat positioniert werden, so dass ausreichend Platz für die

Seitenketten vorhanden ist und gleichzeitig deren Abstand zum Liganden nicht zu groß wird.

Im zweiten Schritt muss aus den gefundenen möglichen Ligandpositionen diejenige ausgewählt werden, die sich am besten für die Realisierung aller geforderten notwendigen Wechselwirkungen eignet. Da dies ihm Rahmen des Optimierungsprozesses geschehen soll, müssen Anpassungen an der TransCent-Energiefunktion vorgenommen werden, damit zusätzlich Bewertungskriterien für die Positionierung des Liganden berücksichtigt werden können. Die Aufgabe der Energiefunktion darf sich also nicht mehr darauf beschränken, mögliche Rotamerkombinationen zu beurteilen. Sie muss gleichzeitig auch die Eignung der Ligandpositionen vergleichend bewerten zu können.

Das dritte Teilproblem besteht darin, die Optimierungsroutine von TransCent so anzupassen, dass auch die flexible Behandlung des Liganden berücksichtigt wird. Ähnlich wie die Bewertung eines Rotamerwechsels von der jeweils aktuellen Rotamerkombination abhängt, ist die Beurteilung einer Ligandposition abhängig von der jeweils vorliegenden Ausrichtung der Seitenketten im aktiven Zentrum. Um eine gleichzeitige Optimierung der Rotamerkombination und der Ligandposition zu ermöglichen, muss daher die Optimierungsroutine von TransCent geeignet überarbeitet werden.

3 Material und Methoden

In diesem Kapitel werden die Programme, Algorithmen und Methoden vorgestellt, welche im Rahmen dieser Arbeit verwendet wurden. Außerdem wird auch die Auswahl und Aufbereitung der Daten erläutert, auf denen die Optimierung und Evaluation des Designalgorithmus durchgeführt wurden.

3.1 TransCent

Zentrales Element dieser Arbeit ist das Enzymdesignprogramm TransCent, welches von André Fischer im Rahmen seiner Doktorarbeit entwickelt wurde, und dessen Funktionsumfang durch die flexible Positionierung des Liganden erweitert werden soll. Der modulare Aufbau, der sich auch in der Energiefunktion (Gl. (3.1)) widerspiegelt, bleibt dabei erhalten.

$$E_{\text{TransCent}} = \omega_{\text{Rosetta}} \cdot E_{\text{Rosetta}} + \omega_{\text{DSX}} \cdot E_{\text{DSX}} + \omega_{\text{Fingerprint}} \cdot E_{\text{Fingerprint}} + \omega_{\text{PROPKA}} \cdot E_{\text{PROPKA}} \tag{3.1}$$

Die TransCent-Energiefunktion ist eine Linearkombination aus den Beiträgen der einzelnen Module. Die ersten beiden Terme E_{Rosetta} (siehe 3.2.3) und E_{DSX} (siehe 3.3) bewerten die Proteinstabilität und die Ligandenbindung und sind ihrerseits wieder Kombinationen aus Ein- und Zwei-Körper-Energietermen. Bei $E_{\text{Fingerprint}}$ (siehe 3.4.3) und E_{PROPKA} (siehe 3.5.3) handelt es sich um Mehr-Körper-Energien, die für die Ausprägung essentieller Wechselwirkungen zwischen Ligand und Proteinseitenketten bzw. das korrekte Einstellen des Protonierungszustände verantwortlich sind.

Basierend auf dieser Energiefunktion ist dann der bestmögliche Designvorschlag dasjenige Modell, das den niedrigsten Energiewert aufweist. Ein Suchvorgang im Raum der möglichen Lösungen wird im Folgenden „Designoptimierung“ genannt. Dabei wird mit Hilfe der Optimierungsroutine das globale Minimum der TransCent-Energiefunktion gesucht und das gefundene Ergebnis als „Modell“ ausgegeben.

3.1.1 Optimierungsroutine

Neben Genetischen Algorithmen und *Dead-End-Elimination* zählt *Simulated Annealing* [63] zu den am häufigsten eingesetzten Optimierungsstrategien im Bereich des Proteindesigns. Es handelt sich dabei um ein heuristisches Optimierungsverfahren, welches den physikalischen Prozess des langsamen Abkühlens eines flüssigen Metalls nachbildet. Die Atome geben schrittweise ihre thermische Energie ab und finden dabei nach und nach ihren Platz im Kristallgitter, so dass schlussendlich eine geordnete Struktur mit minimaler Energie entsteht. Dieser Vorgang wird in abstrakter Form beim *Simulated Annealing* nachgestellt und kann für die Optimierung unterschiedlicher Systeme angewendet werden.

Die Voraussetzungen sind, dass eine Energiefunktion zur Bewertung des aktuellen Systemzustandes und eine Operation existieren, mit der der Systemzustand in einen anderen überführt werden kann, d.h. eine Methode die das schrittweise Abtasten des Lösungsraumes erlaubt. Ausgangspunkt der Optimierung ist eine zufällig gewählte Startkonfiguration. Diese wird anschließend durch Zufallsschritte verändert, wobei ein Schritt, welcher zu einer niedrigeren Gesamtenergie führt immer akzeptiert wird. Dies entspricht einem Gradientenabstieg in der gegebene Energielandschaft. Zusätzlich werden aber auch Schritte akzeptiert, welche die Gesamtenergie erhöhen, und zwar mit einer Wahrscheinlichkeit, die gemäß dem Metropolis-Kriterium [87] berechnet wird:

$$p = \exp\left(-\frac{\Delta \mathrm{E}}{T_i}\right) \tag{3.2}$$

Dabei ist $\Delta \mathrm{E} = \mathrm{E}_{\mathrm{neu}} - \mathrm{E}_{\mathrm{alt}}$ die Differenz zwischen den Energiewerten vor und nach dem Optimierungsschritt und T_i die Pseudotemperatur zum Zeitpunkt i. Diese wird ausgehend von einem relativ hohen Wert schrittweise erniedrigt, was dazu führt, dass die Wahrscheinlichkeit, mit der eine Verschlechterung zugelassen wird, ebenfalls abnimmt. Schlussendlich werden nur noch Verbesserungen akzeptiert und das System konvergiert i.A. gegen das nächste Minimum. Insgesamt führt das gelegentliche Akzeptieren von Verschlechterungen dazu, dass das System während der Optimierung lokale Minima in der Energielandschaft wieder verlassen kann und somit oft bessere Lösungen gefunden werden.

In TransCent wird *Simulated Annealing* Optimierung dazu verwendet, die optimale Rotamerkombination für die zu etablierende Enzymfunktion zu finden. Für jede Position im Protein existiert eine fest vorgegebene Menge von Rotameren (vgl. 3.2.2). Diese repräsentieren entweder eine Aminosäure an den rotierbaren Positionen oder mehrere verschiedene Aminosäuren an den mutierbaren Positionen. Gesucht ist diejenige Kombination an Rotameren, welche die niedrigste Gesamtenergie besitzt und somit die durch die Energiefunktion (siehe Gl. (3.26)) festgelegten Nebenbedingungen am besten erfüllt.

Die Implementierung wurde von [85] übernommen, welche ihrerseits auf [88] basiert. Eine Optimierung startet mit einer beliebig gewählten Rotamerkombination. Bei einem *Simulated Annealing*-Schritt wird dann zufällig eine Proteinposition ausgewählt, für die ebenfalls zufällig ein neues Rotamer bestimmt wird. Anhand des Metropolis-Kriteriums wird dann bestimmt, ob der Rotamertausch akzeptiert wird oder nicht. Der Startwert $T_0 = 100$ für die Pseudotemperatur wird innerhalb von 20 Temperaturschritten gemäß Gl. (3.3) abgesenkt, wobei die Anzahl der *Simulated Annealing*-Schritte, die bei jeder Temperaturstufe ausgeführt werden, proportional zur Gesamtzahl der wählbaren Rotamere ist.

$$T_i = (T_0 - 0,3) \cdot e^{-i} + 0,3 \tag{3.3}$$

Kann innerhalb von vier Temperaturschritten keine weitere Verbesserung gefunden werden, so wird das System wieder „aufgeheizt“, d.h. die Temperatur zurück auf den Ausgangswert gesetzt. Dadurch kann das gefundene Energieminimum wieder verlassen und ggf. eine energetisch noch günstigere Lösung gefunden werden. Am Ende jeder Simulation erfolgt noch eine „Quenchphase“. Hierbei wird die Pseudotemperatur auf 0 gesetzt und ausgehend von der letzten Rotamerkonfiguration in zufälliger Reihenfolge an allen Positionen jedes mögliche Rotamer gesetzt, um zu prüfen, ob dadurch eine weitere Verbesserung möglich ist (Gradientenabstieg). Schlussendlich wird die Rotamerkombination

mit der niedrigsten Energie, die während der Simulation beobachtet wurde, als Ergebnis ausgegeben.

3.2 Proteinstabilität

Proteinstabilität ist die wichtigste Voraussetzung für Enzymaktivität. Selbst wenn prinzipiell alle für die Funktion erforderlichen Aminosäuren vorhanden sind, so kann ein Enzym nur dann katalytisch aktiv werden, wenn es korrekt und stabil gefaltet ist. Insbesondere im aktiven Zentrum muss eine stabile Ausrichtung der Seitenketten gegeben sein, damit diese mit dem Ligandmolekül wechselwirken können. Die Mechanismen, welche für die Proteinstabilität verantwortlich sind, sind gut verstanden [89] und können von Computeralgorithmen relativ zuverlässig erfasst werden. Die Algorithmen sind allerdings bislang zu langsam, um *ab initio* Proteine von der Größe eines durchschnittlichen Enzyms zu falten.

In TransCent basiert das Modul für die Proteinstabilität auf der *score12*-Energiefunktion [90] von Rosetta3 [67]. Die Rosetta Software Suite ist eine umfangreiche Sammlung von Methoden und Energiefunktionen für Anwendungen in unterschiedlichen Bereichen wie z.B. der *ab initio* Proteinstrukturvorhersage [91], dem Proteindesign [92, 93], dem Docken kleiner Ligandmoleküle in Proteinbindetaschen [68, 94] oder dem Protein-Protein-Docking [95]. Die für TransCent relevanten Teile, nämlich die Modellierungseinheit und die Energiefunktion, stammen aus dem Designalgorithmus *fixbb* für starr gehaltenes Proteinrückgrat [40].

Aus der Sicht einer Funktionsübertragung kann die Aufgabe des Moduls für Proteinstabilität so umschrieben werden, dass dadurch nur solche Mutationen zugelassen werden sollen, welche die Proteinstabilität nicht entscheidend negativ beeinflussen oder sogar noch verbessern.

3.2.1 Die Rosetta-Energiefunktion

In der Rosetta Software Suite sind mehrere Energiefunktionen implementiert. Für TransCent wurde die sogenannte *score12*-Funktion übernommen, welche die Standard-Energiefunktion von Rosetta für Proteindesign mit atomarer Auflösung ist (vgl. Supplement von [96]). Die Energiefunktion ist eine Linearkombination mehrerer wissensbasierter Potentiale, welche unter anderem Solvatationseffekte, Wasserstoffbrücken und elektrostatische Wechselwirkungen berücksichtigen.

Die Gesamtenergiefunktion ist eine gewichtete Summe von 19 Einzelenergien, wobei teilweise mehrere Energieterme die gleiche Wechselwirkung beschreiben. So wird z.B. die Energiefunktion für Wasserstoffbrücken aufgeteilt in langreichweitige und kurzreichweitige Wechselwirkungen des Proteinrückgrats, Interaktionen zwischen Seitenketten und Rückgrat, und Wasserstoffbrücken zwischen zwei Seitenketten. Das Lennard-Jones-Potential zur Berücksichtigung der Van-der-Waals-Kräfte (vgl. Abb. 3.1) wird in einen anziehenden und einen abstoßenden Teil zerlegt. Tabelle 3.1 enthält eine Übersicht über alle Energiebeiträge und die dazugehörigen Gewichte. In TransCent wurde der Standardgewichtssatz von Rosetta für *score12* übernommen. Eine detaillierte Beschreibung der Energieterme findet sich in [90] und den darin erwähnten Referenzen.

Energieterm	**Beschreibung**	**Gewicht**
fa_atr	Anziehender Teil des Lennard-Jones-Potential	0,8
fa_rep	Abstoßender Teil des Lennard-Jones-Potential	0,44
fa_sol	Implizites Solvatationsmodell nach Lazaridis & Karplus [97]	0,65
fa_intra_rep	Residuen-interne Van-der-Waals-Wechselwirkungen	0,004
pro_close	Korrekturterm für Prolin-Seitenketten	1
fa_pair	Statistische Paarenergie zur Berücksichtigung elektrostatischer Wechselwirkungen	0,49
hbond_sr_bb	Langreichweitiges Potential für Proteinrückgrat-Wasserstoffbrücken	0,585
hbond_lr_bb	Kurzreichweitiges Potential für Proteinrückgrat-Wasserstoffbrücken	1,17
hbond_bb_sc	Energie für Wasserstoffbrücken zwischen Rückgrat und Seitenkette	1,17
hbond_sc	Energie für Wasserstoffbrücken zwischen zwei Seitenketten	1,1
dslf_ss_dst	Statistische Potentiale für Abstände und Winkel bei Disulfidbrücken	0,5
dslf_cs_ang		2
dslf_ss_dih		5
dslf_ca_dih		5
rama	Ramachandran-Score für Phi-Psi-Winkel-Kombinationen	0,2
omega	Potential für ω-Winkel der Peptidbindung	0,5
fa_dun	Rotamerspezifische Energie abgeleitet aus beobachteten Rotamerhäufigkeiten	0,56
p_aa_pp	Aminosäurespezifische Energie für Phi-Psi-Winkel-Kombinationen	0,32
ref	Referenzenergie für freie Aminosäuren (chemisches Potential)	1

Tabelle 3.1: Beschreibung der Energieterme von *score12*
Die Namen der Energieterme entsprechen der Rosetta-internen Bezeichnungsweise. Die Spalte „Gewicht“ enthält die Gewichtungsfaktoren, die bei der Berechnung von *score12* verwendet werden.

3.2.2 Modellierungseinheit und Rotamerbibliothek

Das wichtigste Ziel des Enzymdesign ist es, eine Proteinsequenz vorherzusagen, welche zu einem Enzym mit der gewünschten Funktion faltet. Um die optimale Primärstruktur zu ermitteln muss jedoch der „Umweg“ über die Tertiärstruktur in atomarer Auflösung gegangen werden, denn die Wechselwirkungen, die beim Enzymdesign von Bedeutung sind, spielen sich auf atomarer Ebene ab. Deswegen verfügt jeder Enzymdesignalgorithmus über eine Modellierungseinheit, mit der die Strukturmodelle erstellt werden können, welche dann durch die Energiefunktion bewertet werden.

Da das Proteinrückgrat beim Design als starr angenommen wird, beschränkt sich in TransCent die Aufgabe der Modellierungseinheit auf das Modellieren der Aminosäureseitenketten. Diese haben je nach Art der Aminosäure eine unterschiedliche Anzahl drehbarer Bindungen, die den Konformationsraum einer Seitenkette definieren. Alanin und Glycin besitzen keine drehbare Bindung, für alle anderen Aminosäuren ergibt sich jedoch

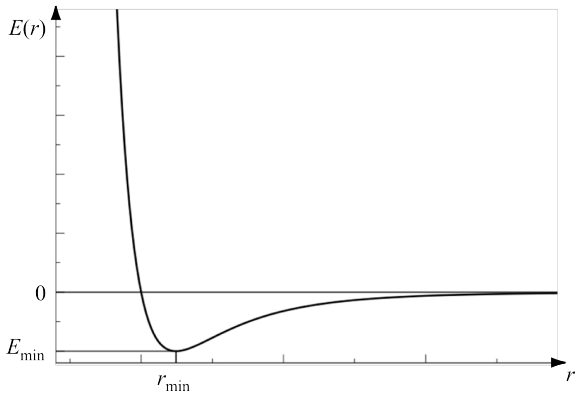


Abbildung 3.1: Das Lennard-Jones Potential
Es beschreibt näherungsweise die Wechselwirkungen zwischen ungeladenen, nicht kovalent verbundenen Atomen. Für große Abstände r wirkt das Potential anziehend ($\sim\frac{1}{r^6}$), bei kleinen Abständen überwiegt die Abstoßung ($\sim\frac{1}{r^{12}}$)

eine mehr oder minder große Anzahl möglicher Konformationen. Um die Komplexität des Designproblems auf ein handhabbares Maß zu reduzieren, wird der Konformationsraum für gewöhnlich auf eine begrenzte Anzahl diskreter Rotamere abgebildet. Für jeden Bindungswinkel χ (vgl. Abb. 3.2) gibt es dann drei mögliche Werte. Für ein Arginin-Residuum mit vier drehbaren Bindungen existieren somit $3^4 = 81$ verschiedene Rotamere, die in natürlich vorkommenden Proteinen unterschiedlich oft beobachtet werden. Die Rotamere werden zusammen mit ihren Häufigkeiten zu einer Rotamerbibliothek zusammengefasst, welche die Grundlage für die Modellierungseinheit bildet. Aufgrund der Menge bekannter Proteinstrukturen ist es möglich, rückgratabhängige Rotamerbibliotheken (engl. *backbone dependent*) zu erstellen, d.h. die Häufigkeiten werden in Abhängigkeit von der Phi-Psi-Winkel-Kombination angegeben. In Rosetta und damit auch in TransCent wird die Dunbrack-Rotamerbibliothek verwendet [52].

Rotamere sind idealisierte Seitenkettenkonformationen, die für sich betrachtet energetisch optimal sind. In Proteinen müssen allerdings auch Wechselwirkungen mit den benachbarten Residuen berücksichtigt werden, was dazu führen kann, dass leicht veränderte Konformationen eine bessere Gesamtenergie ergeben. Vor allem das Lennard-Jones-Potential für Van-der-Waals-Wechselwirkungen (siehe Abb. 3.1) hat hier einen großen Einfluss, da es bei kleinen Abständen sehr sensitiv auf minimale Änderungen reagiert. Die Rosetta-Modellierungseinheit bietet daher die Möglichkeit, zusätzliche Rotamere mit leicht abweichenden χ-Winkeln zu erzeugen. Sofern nicht anders angegeben wurden bei allen Rechnungen die Optionen `-ex1` und `-ex2aro` verwendet. Dies bedeutet, dass neben den „Standardrotameren“ auch Rotamervarianten mit veränderten χ_1-Winkeln und bei den aromatischen Aminosäuren auch mit leicht variierten χ_2-Winkeln erzeugt wurden.

Darüber hinaus gibt es noch die Option `-use_input_sc`, welche die Modellierungseinheit dazu veranlasst, Seitenketten, wie sie in der Eingabestruktur vorliegen, als weitere Rotamervarianten zu verwenden.

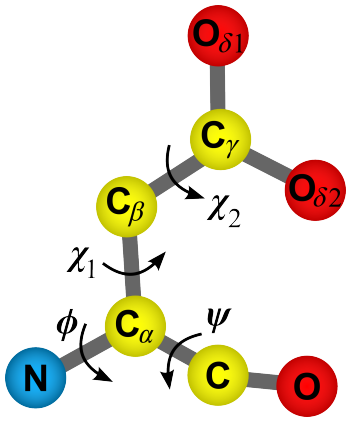


Abbildung 3.2: Schematische Darstellung der Torsionswinkel einer Seitenkette
Die Diederwinkel einer Aminosäureseitenkette, hier am Beispiel von Aspartat gezeigt, werden durchnummeriert je nach Abstand zum C_α-Atom.

3.2.3 Rosetta-TransCent-Schnittstelle

Damit TransCent auf die von Rosetta stammenden Ergebnisse zugreifen kann, musste für Rosetta3 eine Schnittstelle implementiert werden, über die auf die Ergebnisse der Modellierungseinheit und die Energiewerte der Rosetta-Energiefunktion zugegriffen werden kann. Realisiert wurde dies mit Unterstützung von Sam DeLuca aus der Arbeitsgruppe von Prof. Jens Meiler an der Vanderbilt Universität. Die Schnittstelle, genannt *RotamerDumpMover*, ist nun auch Teil der Entwicklerversion von Rosetta3.

Der Export der Rotamerkoordinaten erfolgt atomweise inklusive der Zuordnung zu den einzelnen Proteinpositionen. Bei der Ausgabe der Energiewerte wird die Rosetta-Energiefunktion unterteilt in einen rotamerspezifischen Ein-Körper-Energie-Anteil $E^{1K}_{Rosetta}$ und die entsprechenden Zwei-Körper-Energien $E^{2K}_{Rosetta}$, die sich zwischen zwei Rotameren r_i und r_j ergeben. Daraus kann dann für eine gegebene Rotamerkombination die Energie des Rosetta-Moduls wie folgt berechnet werden:

$$E_{Rosetta} = \sum_{i} E^{1K}_{Rosetta}(r_i) + \sum_{\substack{i,j \\ i<j}} E^{2K}_{Rosetta}(r_i, r_j) \tag{3.4}$$

Alle Berechnungen wurden mit der Entwicklerversion von Rosetta3 (Revisionsnummer: 44897) durchgeführt.

3.3 Ligandenbindung

Genau wie Proteinstabilität ist auch die Bindung eines oder mehrerer Ligandmoleküle eine der Grundvoraussetzungen für Enzymaktivität. Bei manchen Enzymen (z.B. der Antranilat-Phosphoribosyltransferase aus *Sulfolobus solfataricus* [98]) beschränkt sich deren enzymatische Tätigkeit darauf, zwei Substrate korrekt zueinander orientiert und lange genug in räumliche Nähe zu bringen, damit die entsprechende Reaktion ablaufen kann.

Aber auch bei allen anderen Enzymen ist Katalyse nur dann möglich, wenn ein oder mehrere Ligandmoleküle gebunden werden können. Daher ist es unerlässlich, dass beim Design der Energiefunktion auch Anforderungen berücksichtigt werden, welche für die Ligandenbindung wichtig sind.

In vielen Fällen widersprechen diese den Vorgaben für Proteinstabilität, eine möglichst dichte Packung des Proteins zu erreichen [77]. Der Energieterm für Ligandenbindung muss also dafür sorgen, dass das aktive Zentrum eines Enzyms nicht mit großen Seitenketten aufgefüllt wird, welche die Bindung eines Ligandmoleküls verhindern. Stattdessen soll das aktive Zentrum so mit Aminosäuren ausgekleidet werden, dass nach dem „Schlüssel-Schloss-Prinzip“ [99] die spezifische Bindung des Liganden (genauer dessen Übergangszustandes) mit hoher Affinität möglich ist.

In TransCent wird diese Aufgabe vom DSX-Modul übernommen, welches das DrugScore-Modul aus [85] ablöst. Darüber hinaus kommt ihm in der neuen TransCent-Version neben dem Design der Seitenketten noch eine zweite Aufgabe zu. Da mittels DSX die Ligandenbindung bewertet werden kann, hat das Modul auch entscheidenden Einfluss auf die Wahl der Ligandposition (siehe 3.6).

3.3.1 DSX

Das Programm DSX (DrugScore eXtended, [100]) ist eine Weiterentwicklung des DrugScore-Algorithmus [78] und stammt wie sein Vorgänger aus der Gruppe von Gerhard Klebe an der Philipps-Universität Marburg. Es wurde entwickelt, um qualitativ Bindungsaffinitäten von Protein-Ligand-Komplexen vorherzusagen. Die Scoringfunktion $\mathrm{Score_{DSX}}$ verwendet wissensbasierte Potentiale, ähnlich dem Fingerprint-Modul (siehe 3.4), und besteht aus drei Komponenten, die gewichtet aufsummiert werden ($w_{\mathrm{tors}}, w_{\mathrm{SR}}, w_{\mathrm{pair}} \in \{0, 1\}$):

$$\mathrm{Score_{DSX}} = \mathrm{w_{tors}Score_{tors}} + \mathrm{w_{SR}Score_{SR}} + \mathrm{w_{pair}Score_{pair}} \tag{3.5}$$

Der Torsionswinkel-Score $\mathrm{Score_{tors}}$ bewertet die Diederwinkel der Bindungen im Ligandmolekül und soll unnatürliche Winkel bestrafen. Die zugehörigen Potentiale wurden aus den Strukturen der Cambridge Structural Database (CSD, [101]) abgeleitet. Desolvatationseffekte bei der Ligandenbindung werden für Ligandatome und Proteinatome in dem Term $\mathrm{Score_{SR}}$ zusammengefasst. Die entsprechenden SR-Potentiale wurden anhand von Beispielstrukturen aus der PDB berechnet.

Den bedeutendsten Anteil an der DSX-Scoringfunktion haben jedoch die distanzbasierten Paarpotentiale $\mathrm{Score_{pair}}$. In ihnen sind die typischen Abstände zwischen wechselwirkenden Ligand- und Proteinatomen kodiert. Für die Paarpotentiale existieren zwei Versionen, die aus der PDB bzw. der CSD abgeleitet wurden. Die Berechnung des Paar-Scores erfolgt gemäß:

$$\mathrm{Score_{pair}} = \sum_{\mathrm{a_i} \in \mathrm{P}} \sum_{\mathrm{a_j} \in \mathrm{L}} \mathrm{score_{pair}} \left(\mathrm{c} \left(\mathrm{a_i}, \mathrm{a_j} \right), \mathrm{r} \left(\mathrm{a_i}, \mathrm{a_j} \right) \right) \tag{3.6}$$

Dabei ist $r\left(a_i, a_j\right)$ der Abstand zwischen zwei Atomen a_i und a_j aus der Menge der Proteinatome P bzw. der Ligandatome L und $c\left(a_i, a_j\right)$ die Zuordnungsfunktion für das entsprechende Paarpotential. Dieses richtet sich nach dem „Kontakttyp“. DSX kennt insgesamt 68 verschiedene Atomtypen für 16 Elemente (Wasserstoffe werden nicht betrachtet). Die möglichen Atomtypkombinationen werden auf 300 (PDB) bzw. 600 (CSD) Kontakttypen abgebildet, wobei jeder sein eigenes Paarpotential besitzt.

Bei den Proteinatomen ist die Bestimmung der Atomtypen relativ einfach, da im Allgemeinen nur die 20 kanonischen Aminosäuren vertreten sind. Die Vielfalt möglicher Ligandmoleküle ist allerdings beträchtlich, was die Festlegung der Atomtypen verkompliziert. In DSX wird daher der fconv-Algorithmus [102] verwendet, um automatisiert die Atomtypen zu ermitteln. Bei der Ableitung der Potentiale werden Atompaare mit einem Abstand von maximal 6 Å berücksichtigt. Dadurch kann sichergestellt werden, dass nur direkte Wechselwirkungen zwischen Ligand und Protein erfasst werden und die Potentiale nicht durch indirekte wasservermittelte Kontakte verzerrt sind.

3.3.2 DSX-Energiefunktion

Die Scoringfunktion von DSX ist prinzipiell so ausgelegt, dass sie eine Protein-Ligand-Kombination als Ganzes bewertet. Für das Enzymdesign mit TransCent wird jedoch eine rotamerbasierte Variante benötigt. Gerd Neudert hat daher DSX so angepasst, dass dem Programm eine Menge von Rotameren im PDB-Format übergeben werden kann, deren Scores einzeln berechnet und ausgegeben werden. Dabei verwendet DSX statt der gesamten Scoringfunktion (3.5) lediglich die distanzbasierten Paarpotentiale. Diese Beschränkung hat zwei Gründe: Zum einen liegt der Ligand in der Positionsbibliothek nur in einer einzigen Konformation vor (siehe 3.6). Der Torsionswinkel-Score ist somit in allen Fällen gleich und braucht daher nicht berücksichtigt zu werden. Zum anderen kann sich potentiell bei jedem Rotamer- bzw. Ligandpositionswechsel die lösungsmittelzugängliche Oberfläche des Protein-Ligand-Komplexes ändern und muss neu berechnet werden. Dies führt zu einem enormen Mehraufwand an Rechenzeit, weswegen auf die Hinzunahme von $\mathrm{Score_{SR}}$ verzichtet wird.

Da die Minimierung der DSX-Scoringfunktion äquivalent zur Optimierung der Bindungsaffinität ist, kann der Score $\mathrm{Score_{DSX}}$ ohne weitere Anpassung als Energiefunktion $\mathrm{E_{DSX}}$ in TransCent übernommen werden. Der Energieterm setzt sich somit zusammen aus einer Ein-Körper-Energie, welche die Wechselwirkungen zwischen Ligand und Proteinrückgrat umfasst, und einem Zwei-Körper-Energieterm. Dieser entspricht den Scores, welche sich für die Seitenketten der einzelnen Rotamere und die Ligandpositionen der Positionsbibliothek ergeben.

3.4 Fingerprint

Mit dem DSX-Modul wird allgemein die Ligandenbindung bewertet. Somit werden durch die DSX-Energiefunktion nur relativ abstrakte Bedingungen an die Positionen im aktiven Zentrum gestellt. Bei jeder Enzymfunktion gibt es jedoch sehr spezifische Anforderungen an einzelne Proteinpositionen, die nur von wenigen, oft nur einer einzigen Aminosäure erfüllt werden können. Vergleicht man Strukturen homologer Enzyme gleicher Funktion so sind gewisse charakteristische Merkmale in allen aktiven Zentren vorhanden. Nicht nur sind die katalytischen Reste strikt konserviert, sondern auch die relative Geometrie der Seitenketten-Ligand-Kombination ist erhalten, da essentielle Wechselwirkungen für den Ablauf der Katalysereaktion ausgebildet werden müssen. Analysiert man die Gemeinsamkeiten der aktiven Zentren, so erhält man den „Fingerabdruck“ einer Enzymfunktion. Die Anforderungen sind allerdings nicht bei allen Positionen gleich. Manche sind sehr genau

definiert, andere zeigen mehr Flexibilität was die Wahl der Aminosäuren und die Platzierung der wechselwirkenden Atome betrifft.

Mit dem Fingerprint-Modul werden die eben beschriebenen Bedingungen in wissensbasierte Potentiale übersetzt, welche in Form eines Energieterms $E_{Fingerprint}$ (siehe Gl. (3.15)) in die Optimierung eingehen. Das Fingerprint-Modul sorgt also dafür, dass nur solche Seitenketten im aktiven Zentrum platziert werden, die genau definierte, charakteristische Wechselwirkungen ausbilden können. Hierbei handelt es sich vor allem um die polaren Wechselwirkungen des Wasserstoffbrücken-Netzwerkes zwischen Protein und Ligandmolekül.

Bei der Übertragung einer Enzymfunktion auf ein anderes Proteingerüst ist aufgrund der Unterschiede im Rückgratverlauf *a priori* keine Zuordnung der Fingerprint-Potentiale zu bestimmten Positionen im Zielprotein möglich, zumal wenn mehrere Ligandpositionen gleichzeitig beim Design berücksichtigt werden sollen (vgl. 3.6). Stattdessen wird der „Fingerabdruck“ einer Enzymfunktion in TransCent durch Potentiale repräsentiert, welche relativ zum Liganden definiert sind. Erst während der Designoptimierung wird entschieden, welche Proteinposition am besten zu welchem Potential passt bzw. ob überhaupt eine solche Position existiert.

Die Implementierung des Fingerprint-Moduls wurde aus [85] (dort noch Funktionsdefinition genannt) übernommen und überarbeitet. Auf Stellen, an denen bedeutende Änderungen vorgenommen wurden, wird im Folgenden ausdrücklich hingewiesen.

3.4.1 Strukturbibliothek

Um die charakteristischen Merkmale eines aktiven Zentrums herauszuarbeiten, ist eine einzige Enzymstruktur als Vorlage i.A. nicht ausreichend. Zwar müssen in jedem Fall alle für die Katalyse erforderlichen Wechselwirkungen vorhanden sein, allerdings können diese dann nicht von Interaktionen unterschieden werden, welche nur typisch für das betrachtete Enzym sind. Andererseits hat die Natur auch gezeigt, dass es verschiedene Realisierungsmöglichkeiten für ein und die selbe Enzymfunktion gibt. So können z.B. äquivalente Wasserstoffbrücken sowohl von Glutamin- als auch von Asparagin-Seitenketten ausgebildet werden. Um diese Variabilität erfassen zu können und gleichzeitig eine statistisch aussagekräftige Grundlage zur Bestimmung des Fingerprints zu haben, muss eine ausreichend große Menge an Beispielstrukturen vorhanden sein. Für die allermeisten Enzyme existiert jedoch, wenn überhaupt, nur eine bekannte Struktur und nur in seltenen Fällen sind es mehr als zehn.

Um dennoch auf einer breiten Datenbasis arbeiten zu können, wird in TransCent eine Strukturbibliothek erstellt. Diese kann neben der Vorlagestruktur auch weitere Kristallstrukturen aus der PDB enthalten. Den Hauptbestandteil machen allerdings Homologiemodelle aus, welche mit Hilfe des Programms MODELLER [84] erstellt werden.

3.4.1.1 Homologe Sequenzen

Grundlage für die Berechnung der Strukturbibliothekmodelle ist eine Vorlagestruktur sowie ein MSA homologer Sequenzen. Dieses kann aus unterschiedlichen Quellen stammen. In [85] wurden z.B. MSAs aus der PFAM-Datenbank [103] als Basis für die Homologiemodellierung verwendet. In dieser Arbeit wurden die MSAs, sofern nicht ausdrücklich

anders angegeben, mit dem Programm S2MSAAA erstellt [104], welches Laura Schiller in ihrer Bachelorarbeit entwickelt hat. Die einzelnen Programmschritte werden im Folgenden kurz umrissen.

Über eine BLAST-Suche am NCBI-Server werden in der nicht-redundanten Sequenzdatenbank [105] Treffer mit hoher Ähnlichkeit zur Anfragesequenz identifiziert und heruntergeladen. In einem ersten Filterschritt werden anschließend Sequenzen ausgeschlossen, deren Länge zu stark von der der Anfragesequenz abweicht. Aus den verbliebenen Sequenzen wird mittels MAFFT [106, 107, 108] ein erstes grobes MSA erstellt. Daraus werden dann paarweise Sequenzähnlichkeitswerte abgeleitet und für Paare mit einem zu hohen bzw. zu niedrigen Wert jeweils ein Vertreter gelöscht. Nach diesem zweiten Filterschritt stehen mehrere Programme zur Auswahl (MAFFT, T-Coffee [109], MSAProbs [110]), mit denen das endgültige MSA erstellt werden kann. Aus Konsistenzgründen fiel die Wahl im Rahmen dieser Arbeit auf MAFFT. Die verwendete Version von S2MSAAA wurde für die Benutzung mit TransCent erweitert und akzeptiert auch PDB-Dateien als Eingabe, aus denen automatisch die Proteinsequenz ausgelesen wird.

TransCent erwartet die Sequenzdaten als Eingabe im MultipleFASTA-Format. Sollten diese noch nicht aligniert sein, so wird mit Hilfe von MAFFT (Version 6.857) zunächst ein MSA erstellt. Unabhängig von der Quelle des MSAs wird in TransCent eine erneute Filterung der Sequenzen vorgenommen. Für eine korrekte Berechnung des Enzym-Fingerprints muss die Strukturbibliothek Modelle möglichst hoher Qualität enthalten. Homologiemodellierungen sind aber umso genauer, je größer die Ähnlichkeit zur Vorlagestruktur ist. Die Modellierung aktiver Zentren stellt hierbei besonders hohe Anforderungen [85]. Dementsprechend erfolgt die Filterung der homologen Sequenzen basierend auf der Sequenzidentität der Residuen im aktiven Zentrum. Als solche werden in diesem Zusammenhang alle Proteinpositionen definiert, die in der Vorlagestruktur einen Abstand von maximal 6 Å zum Liganden aufweisen. Über das MSA erhält man die Zuordnung zu den äquivalenten Positionen in den anderen Sequenzen.

Somit hängt der Wert für die Sequenzidentität aber auch stark von der Alignmentqualität ab, denn Fehlalignments können dazu führen, dass der Identitätswert zu hoch oder (wahrscheinlicher) zu niedrig geschätzt wird. Deswegen wird in TransCent nicht nur auf Sequenzidentität gefiltert, sondern es werden zusätzlich Sequenzen aussortiert, bei denen keine zuverlässige Zuordnung möglich ist. Zur Bewertung der Qualität des Alignments wird der CORE-Wert verwendet. Dieser kann mit T-Coffee (Version 8.99) berechnet werden (Option `-score`) und ist detailliert in [111] beschrieben. Jeder Position im MSA wird damit ein Wert zwischen 0 und 9 zugeordnet, wobei ein hoher Wert eine hohe lokale Alignmentqualität anzeigt. Für die Bewertung einer Sequenz A wird der CORE-Wert über alle Spalten C_{as} gemittelt, die zum aktiven Zentrum gehören:

$$\mathrm{sCORE\,(A)} = \frac{1}{9\mathrm{N}} \sum_{\mathrm{i} \in \mathrm{C_{as}}} \mathrm{CORE\,(A_i)} \tag{3.7}$$

Die Normierung sorgt dafür, dass der Wertebereich unabhängig von der Anzahl der Positionen zwischen 0 und 1 liegt. Neben der zu modellierenden Sequenz A muss zusätzlich die Alignmentqualität der Vorlagesequenz R mit berücksichtigt werden, da beide das paarweise Alignment beeinflussen, welches aus dem MSA abgeleitet wird und die Grundlage der Homologiemodellierung bildet.

$$\mathrm{alCORE\,(A,R)} = \mathrm{sCORE\,(A)} \cdot \mathrm{sCORE\,(R)} \cdot 100 \tag{3.8}$$

In TransCent wird daher der alCORE-Wert neben der Sequenzidentität als Filterkriterium verwendet. Die Schwellwerte wurden empirisch in [85] ermittelt und liegen bei 40 für den alCORE-Wert und 60% für die Sequenzidentität im aktiven Zentrum. Die verbliebenen Sequenzen werden abschließend mit MAFFT realigniert.

3.4.1.2 Homologiemodellierung

Für die Berechnung der Homologiemodelle für die Strukturbibliothek wird in TransCent das Programm MODELLER [84, 112] (Version 9.9) verwendet. Der Algorithmus ist ein bewährtes Verfahren zur Homologiemodellierung und basiert auf dem Konzept der Erfüllung räumlicher Abstands- und Winkelbedingungen. Diese stammen zum einen aus dem Alignment mit der Vorlagestruktur. Zum anderen sind sie statistischer Natur und werden aus einer Strukturdatenbank abgeleitet. Darüber hinaus kommen ergänzend stereochemische Nebenbedingungen hinzu, welche auf dem CHARMM-Kraftfeld [47] basieren. Außerdem können auch Ligandatome in eingeschränktem Umfang bei der Modellierung berücksichtigt werden. Durch das Platzieren des Liganden im aktiven Zentrum des Modells als sogenannter BLK-Rest erzeugt die sterische Hinderung der Ligandatome zusätzliche Nebenbedingungen.

Diese Überlagerung aus Abstands- und Winkelbedingungen kann durch Wahrscheinlichkeitsdichtefunktionen ausgedrückt werden, die zu einer Zielfunktion für den Modellierungsprozess zusammengefasst werden. Mit einem Kombinationsansatz aus CG-Verfahren (von engl. *conjugate gradients*), Moleküldynamik und *Simulated Annealing* versucht MODELLER, das Optimum der Zielfunktion zu ermitteln.

Bei TransCent ist der MODELLER-Algorithmus in die Berechnungsroutine für die Strukturbibliothek eingebettet. Aus dem MSA der homologen Sequenzen wird für alle Sequenzen automatisch das paarweise Alignment mit der Vorlagesequenz ermittelt und weiterprozessiert. Das Ergebnis der Modellierung wird mit Hilfe des Programms TM-Align ([113], Version vom 30.01.2011) mit der Vorlagestruktur aligniert. So werden Abweichungen, die während der Homologiemodellierung entstanden sein können, ausgeglichen.

Die MODELLER-Routinen wurden von [85] übernommen und angepasst, um Kompatibilitätsprobleme mit der MODELLER-Version 9.9 zu beheben.

3.4.2 Fingerprint Potentiale

Das Fingerprint-Modul übernimmt in TransCent zwei Aufgaben. Zum einen sorgt seine Energiefunktion dafür, dass während des Designprozesses das Setzen von Rotameren, welche eines der Fingerprint-Potentiale erfüllen können, energetisch belohnt wird. Dadurch wird erreicht, dass alle wichtigen Wechselwirkungen, die den „Fingerabdruck“ der Enzymfunktion ausmachen, im Design etabliert werden, sofern dies möglich ist. Zuvor jedoch müssen die charakteristischen Wechselwirkungen des aktiven Zentrums erst identifiziert und in Potentiale übersetzt werden.

3.4.2.1 Ableiten der Fingerprint-Potentiale

Die Basis für die Berechnung der Fingerprint-Potentiale bilden die Strukturen der Strukturbibliothek und das dazugehörige MSA. Zunächst werden im aktiven Zentrum einer

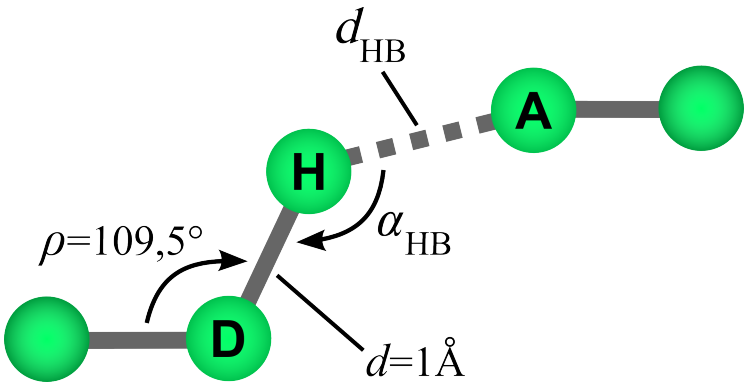


Abbildung 3.3: Schematische Darstellung einer Wasserstoffbrücke
Die angegebenen Abstände zwischen Donoratom D, Wasserstoffatom H und Akzeptoratom A beziehen sich auf die Atommittelpunkte.

Struktur alle Seitenketten identifiziert, welche Teil einer polaren Interaktion mit dem gebunden Ligandmolekül sind. Diese können sowohl als Wasserstoffbrücken-Donor wie auch als Akzeptor agieren. Da Wasserstoffatome in den meisten Proteinstrukturen nicht enthalten sind, werden die Wasserstoffpositionen für die entsprechenden Gruppen modelliert. Es wird dabei eine optimale Geometrie angenommen mit einem Abstand zwischen Donoratom und Wasserstoff von 1 Å und einem Bindungswinkel ρ von 109,5°.

Aus energetische Sicht sind lineare Wasserstoffbrücken mit einer Bindungslänge von 1,8 Å am günstigsten. Allerdings treten in Proteinen auch Wasserstoffbrücken auf, deren Eigenschaften mehr oder weniger von dieser optimalen Anordnung abweichen [114]. TransCent verwendet daher Schwellwerte von $d_{HB} < 4,5$ Å und $\alpha_{HB} > 90°$ bei der Erkennung von Wasserstoffbrücken (siehe Abb. 3.3). Das Winkelkriterium bezieht sich dabei auf den Winkel α_{HB} zwischen Donor-, Wasserstoff- und Akzeptoratom. Da die Position des Wasserstoffatoms nur modelliert ist, wird der Abstand d_{HB} nicht wie üblich zwischen Akzeptor- und Wasserstoffatom sondern zwischen Akzeptor- und Donoratom gemessen.

Die relativ großzügige Wahl der Schwellwerte ist, neben der Bandbreite real existierender Wasserstoffbrücken, zusätzlich begründet durch die begrenzte Auflösung von Proteinkristallstrukturen, die für gewöhnlich nicht besser als 1 Å ist. Außerdem handelt es sich bei Kristallstrukturen um „Momentaufnahmen“, welche keinen Aufschluss über die Flexibilität von Aminosäureseitenketten geben.

Ob eine Wasserstoffbrücke tatsächlich ausgebildet werden kann, hängt neben der räumlichen Anordnung auch vom Protonierungszustand der Seitenketten ab. Diese Einschränkung wird bei der Erkennung von Wasserstoffbrücken in TransCent jedoch ignoriert, sodass implizit auch Wechselwirkungen zwischen geladenen Gruppen berücksichtigt werden.

Insgesamt werden so alle wichtigen Interaktionen zwischen dem Liganden und dem aktiven Zentrum eines Enzyms erfasst. Zur Ableitung der Fingerprint-Potentiale wird eben beschriebenes Verfahren für alle Strukturen der Strukturbibliothek angewendet. Jede gefundene Wechselwirkung wird durch das Donor- bzw. Akzeptoratom der beteiligten Seitenkette repräsentiert, wobei die Atompositionen äquivalenter Proteinpositionen zu Gruppen zusammengefasst werden (siehe Abb. 3.4).

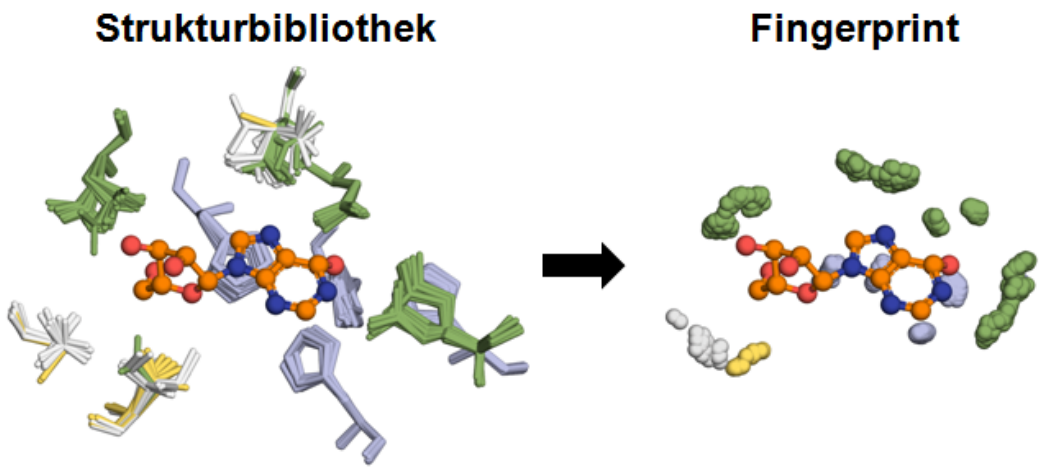


Abbildung 3.4: Übersetzung der Strukturbibliothek in Fingerprint-Punktwolken
Zur Veranschaulichung der Vorgehensweise bei der Ableitung der Fingerprint-Punktwolken ist exemplarisch eine Überlagerung von 46 Strukturen aus der Strukturbibliothek der Adenosin-Desaminase aus *Mus musculus* (PDB-Code: 1a4m) dargestellt. Neben dem Ligandmolekül (orange, Ball & Stick-Modell) werden auch die Seitenketten der mit ihm wechselwirkenden Residuen in der Stick-Darstellung gezeigt. Die Farbwahl gibt dabei die Art der Aminosäuren wieder. Basische Seitenketten sind blau (hier nur Histidin), negativ geladene grün (Glutamat, Aspartat). Cystein-Residuen sind gelb markiert. Hydrophobe Seitenketten sind in weiß dargestellt. Die Punktwolken der wechselwirkenden Schweratome sind entsprechend ihrer Zugehörigkeit zu einer der Seitenketten eingefärbt.

Daraus resultieren Punktwolken unterschiedlicher Form und Ausdehnung, welche stellvertretend für die räumliche Variabilität einer Wechselwirkung des Enzym-Fingerprints in der Strukturbibliothek stehen. Kleine und dichte Wolken zeigen an, dass die Orientierung der entsprechende Seitenkette zum Liganden in allen Fällen sehr ähnlich ist und folglich exakt eingehalten werden muss, wohingegen weit verstreute Wolken bedeuten, dass unterschiedliche Seitenkettenkonformationen als gleichwertig betrachtet werden können.

Im Fingerprint-Modul wird die räumliche Verteilung einer Punktwolke durch eine Wahrscheinlichkeitsdichtefunktion $\rho_{\text{Fingerprint}}$ angenähert. Diese hat die Form einer dreidimensionalen Gaußfunktion:

$$\rho_{\text{Fingerprint}}\left(\vec{x}\right) = \frac{1}{(2\pi)^{\frac{3}{2}}\left|\Sigma\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^{T}\Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \tag{3.9}$$

Aus den Koordinaten der Punktwolke kann deren Kovarianzmatrix Σ und der Mittelpunktsvektor $\vec{\mu}$ berechnet werden. $|\Sigma|$ bezeichnet hier die Determinante der Kovarianzmatrix. Folglich kann jedem Punkt $\vec{x}$ ein Wahrscheinlichkeitswert $\rho_{\text{Fingerprint}}\left(\vec{x}\right)$ zugeordnet werden, wobei der Wert im Mittelpunkt μ maximal ist und mit zunehmendem Abstand zum Mittelpunkt gegen 0 konvergiert. Die Isoflächen einer Gaußfunktion haben die Form von Ellipsoiden, deren Halbachsenlängen gleich einem Vielfachen der Standardabweichungen der zugrunde liegenden Punkteverteilung entlang dieser Halbachsen sind [115].

Dadurch kann jedes Fingerprint-Potential mit einem Ellipsoid E identifiziert werden (vgl. Abb. 4.2).

Es wird allerdings nicht jede Punktwolke in ein Fingerprint-Potential übersetzt. Eine Voraussetzung ist, dass mindestens 5% der Strukturbibliothekmodelle zu einer Wolke beitragen. Außerdem werden Punktwolken verworfen, die einer Proteinposition entsprechen, für welche die Vorlagesequenz im MSA eine Lücke hat, da die dadurch beschriebene Wechselwirkung keine Entsprechung in der Vorlagestruktur hat. Somit wird sichergestellt, dass der Fingerprint nur wichtige Protein-Ligand-Interaktionen umfasst.

Auch die Punktwolken selbst werden gefiltert, bevor sie in Wahrscheinlichkeitsdichtefunktionen umgewandelt werden. Das Filterkriterium hierzu wurde für die in dieser Arbeit vorgestellte Version von TransCent überarbeitet und ist nun konsistent mit der Verwendung der Ellipsoide an anderen Stellen des Algorithmus. Dazu wird zunächst ein „3σ-Roh-Ellipsoid“ berechnet, d.h. die Oberfläche des Ellipsoiden wird definiert als die Isofläche der Gaußfunktion im Abstand der dreifachen Standardabweichung von dessen Mittelpunkt. Alle Punkte außerhalb dieses Ellipsoiden werden gelöscht und die verbliebene gefilterte Punktwolke in die Wahrscheinlichkeitsdichtefunktion $\rho_{\text{Fingerprint}}$ übersetzt. Dadurch können „Ausreißer“ eliminiert werden, welche sonst die Orientierung und die Kompaktheit der Ellipsoide verzerren würden.

Über die Wahrscheinlichkeitsdichtefunktion kann die räumliche Orientierung von Seitenketten relativ zum Liganden erfasst werden. Für eine vollständige Beschreibung eines Fingerprint-Potentials muss aber auch festgelegt werden, welche Aminosäuren für die betrachtete Wechselwirkung geeignet sind. Dazu werden die relativen Häufigkeiten $f_{\text{Fingerprint}}$ der Aminosäuren in den Punktwolken berechnet, normiert auf die Gesamtzahl aller Strukturen in der Strukturbibliothek. Dadurch wird nicht nur bestimmt, welche Aminosäuren für ein Potential in Frage kommen, sondern auch anhand der Konserviertheit einer Wechselwirkung, wie wichtig sie für die Enzymfunktion ist.

Empirisch wurde in [85] ermittelt, dass für eine zuverlässige Berechnung des Fingerprints mindestens 80 Strukturen bzw. Strukturmodelle notwendig sind.

Die neue Version von TransCent wurde mit einer Schnittstelle versehen, über die Fingerprint-Potentiale in Form von Textdateien ausgegeben bzw. eingelesen werden können. Dadurch ist es möglich, die abgeleiteten Potentiale zu überarbeiten, zu entfernen bzw. weitere Potentiale hinzuzufügen. Darüber hinaus ist mit dieser Schnittstelle auch die Möglichkeit geschaffen, *de novo* generierte Fingerprints zu verwenden und damit in der Natur nicht vorkommende Enzymfunktionen zu modellieren.

3.4.2.2 Referenzmodell

Für die Berechnung wissensbasierter Potentiale wird neben den beobachteten Häufigkeiten und Verteilungen auch immer ein Referenzmodell benötigt [56]. Bei vielen bioinformatischen Fragestellungen wird hierfür die Annahme getroffen, dass die zu erwartenden „Beobachtungen“ zufällig verteilt sind. Häufig werden die Verteilungen aus entsprechenden Referenzdatensätzen geschätzt. Nachdem es keine Beispielstrukturen mit „zufälligen“ Protein-Ligand-Interaktionen gibt, kann die Wahrscheinlichkeitsdichtefunktion ρ_{Referenz} nicht aus einem Referenzdatensatz geschätzt werden. Stattdessen wird eine Gleichverteilung innerhalb eines Ellipsoiden unterstellt wird, d.h. die Seitenketten werden zufällig

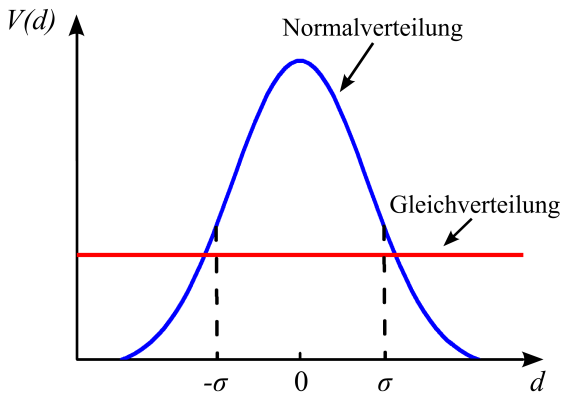


Abbildung 3.5: Vergleich von Normalverteilung und Gleichverteilung
Der energetisch begünstigte Bereich liegt symmetrisch um das Maximum der Normalverteilung und wird durch die Schnittpunkte mit der Gleichverteilung definiert. Die Höhe der Gleichverteilung hängt wiederum vom „Volumen“ der Normalverteilung ab. Im eindimensionalen entspricht dies einem Vielfachen der Standardabweichung σ.

platziert:

$$\rho_{\text{Referenz}}(\vec{x}) = \frac{1}{V_{3\sigma}} \tag{3.10}$$

$V_{3\sigma}$ bezeichnet hier das Volumen des entsprechenden „3σ-Ellipsoiden“. Somit wird ein Bereich definiert, innerhalb dessen die Platzierung eines Donor- bzw. Akzeptoratoms energetisch begünstigt ist (vgl. eindimensionale Darstellung in Abb. 3.5).

Die Wahl eines Referenzmodells für die Aminosäurehäufigkeiten gestaltet sich ähnlich schwierig, da auch hierfür keine Beispiele existieren. Um dennoch Referenzhäufigkeiten schätzen zu können, werden für alle 20 Aminosäuren as Rotamere r aus einer Rotamerbibliothek [52] auf die entsprechenden Proteinpositionen modelliert und die Teilmenge R^{ww} der Rotamere ermittelt, die tatsächlich mit dem Liganden wechselwirken können.

$$f_{\text{Referenz}}(as) = f_{\text{Amino}}(as) \cdot \sum_{r \in R^{\text{ww}}} f_{\text{Dunbrack}}(r) \tag{3.11}$$

Die Referenzhäufigkeiten $f_{\text{Referenz}}(as)$ ergeben sich also aus dem Produkt von Aminosäure- und Rotamerhäufigkeiten. In der neuen Version von TransCent werden die Aminosäuren nicht mehr als gleich wahrscheinlich angenommen. Stattdessen werden für $f_{\text{Amino}}(as)$ die Häufigkeiten aus der Swiss-Prot Datenbank verwendet (siehe Tabelle 3.3).

3.4.3 Fingerprint-Energie und Zuordnungsproblem

Zur Bewertung einzelner Rotamere müssen die Fingerprint-Potentiale in eine Energiefunktion überführt werden. Bei wissensbasierten Potentialen erfolgt dies für gewöhnlich in Form von logarithmierten Chancenquotienten [56]. Für ein Rotamer r in einem Potential P wird daher die Energie wie folgt berechnet:

$$E_{\text{Fingerprint}}(r, P) = -ln\left(\frac{\rho^{P}_{\text{Fingerprint}}(\vec{x}_r) \cdot f^{P}_{\text{Fingerprint}}(as_r)}{\rho^{P}_{\text{Referenz}}(\vec{x}_r) \cdot f^{P}_{\text{Referenz}}(as_r)}\right) \tag{3.12}$$

Hierbei steht $\vec{x}_r$ für den Ortsvektor des wechselwirkenden Atoms und as_r für die Aminosäureart des Rotamers. Aus numerischen Gründen werden für die Fingerprint-Häufigkeiten $f_{\mathrm{Fingerprint}}(as)$ Pseudocounts eingeführt. Der minimale Wert wird auf 10^{-4} gesetzt, wodurch das Divergieren der Energiefunktion verhindert wird. Falls mehrere Atome einer Seitenkette als Wechselwirkungspartner in Frage kommen (z.B. beide Sauerstoffatome der Carboxylat-Gruppe einer Aspartat-Seitenkette), so wird das Rotamer mit dem besten der sich ergebenden Energiewerte bewertet.

Aufgrund ihrer physikochemischen Ähnlichkeit werden Aspartat- und Glutamat- bzw. Asparagin- und Glutamin-Residuen bei der Energieberechnung äquivalent behandelt, d.h. für ein Fingerprint-Potential werden immer beide Aminosäuren als passend definiert, auch wenn nur eine der beiden in der Strukturbibliothek vertreten ist. Durch die unterschiedliche Länge der Seitenketten können Abweichungen im Verlauf des Proteinrückgrats zwischen Vorlage- und Zielprotein ausgeglichen werden.

Mit $\mathrm{E}_{\mathrm{Fingerprint}}(\mathrm{r},\mathrm{P})$ steht also ein Ausdruck zur Verfügung, mit dem eine Rotamer-Potential-Kombination energetisch bewertet werden kann. Zur Berechnung des energetischen Gesamtbeitrags des Fingerprint-Moduls können jedoch nicht wie bei der Rosetta- und der DSX-Energiefunktion die Einzelbeiträge einfach aufsummiert werden. Zum einen gibt es nur sehr wenige Rotamere, die ein Potential „erfüllen“ können, d.h. ein Donor- bzw. Akzeptoratom so platzieren, dass sie die entsprechende Wechselwirkung wirklich eingehen können. Zum anderen gibt es aber auch den Fall, dass Rotamere von zwei, meist benachbarten Positionen dasselbe Potential erfüllen können, und so eine Wechselwirkung doppelt ausgebildet würde. Ziel des Designs ist es jedoch, jedes Potential mit genau einer passend orientierten Seitenkette abzudecken, wie es der Situation im Vorlageenzym entspricht. Es besteht also ein Zuordnungsproblem, welches in TransCent mit Hilfe der Ungarischen Methode [116] gelöst wird. Diese liefert die aus energetischer Sicht optimale Kombination von Potentialen und Rotameren. Für eine detaillierte Beschreibung des Algorithmus wird auf [85] verwiesen.

Bei der Berechnung der Gesamtenergie $\mathrm{E}_{\mathrm{Fingerprint}}$ müssen drei Situationen unterschieden werden. Im Optimalfall sind alle N Fingerprint-Potentiale erfüllt und die Gesamtenergie ist gleich der Summe der Einzelenergien:

$$\mathrm{E}_{\mathrm{Fingerprint}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{E}_{\mathrm{Fingerprint}}\left(\mathrm{r}_{\mathrm{z(i)}}, \mathrm{P}_{\mathrm{i}}\right) \tag{3.13}$$

wobei $z(i)$ die Zuordnungsfunktion ist, welche die Rotamere auf die Potentiale abbildet. Wird eine Seitenkette modelliert, die zwar mit dem Liganden wechselwirken kann, jedoch keinem Fingerprint-Potential entspricht, so wird diese mit einer Strafenergie E_{p} belegt. Zum einen wird dadurch verhindert, dass Potentiale mehrfach erfüllt werden. Noch wichtiger ist allerdings der Effekt, dass dadurch implizit auch hydrophobe Bereiche mit apolaren „Wechselwirkungen“ zwischen Ligand und Enzym durch den Fingerprint beschrieben werden.

Im dritten Fall bleiben ein oder mehrere Potentiale unerfüllt, d.h. die Rotamerkonfiguration enthält kein Rotamer, welches die entsprechende Wechselwirkung ausbilden kann. Je nach Potential tritt dieser Fall aber auch mehr oder weniger oft bei den Modellen der Strukturbibliothek auf. Es gibt also Beispiele anhand derer Häufigkeiten $m_{Fingerprint}$ für das Fehlen der Wechselwirkungen geschätzt werden können. Die Referenzhäufigkeiten $m_{Referenz}$ werden analog zu den Referenzhäufigkeiten f_{Referenz} für die Fingerprint-Poten-

tiale ermittelt (siehe Gl. (3.11)) und zusammen mit $m_{Fingerprint}$ in einen Energieterm übersetzt:

$$E_m(P) = -ln\left(\frac{m^P_{Fingerprint}}{m^P_{Referenz}}\right) \tag{3.14}$$

Die Energien $E_m(P)$ werden für alle nicht erfüllten Potentiale P zur Gesamtenergie $\mathrm{E_{Fingerprint}}$ hinzuaddiert. Je nach Verhältnis von $m_{Fingerprint}$ und $m_{Referenz}$ wird demnach das Nichterfüllen eines Potentials energetisch mehr oder weniger stark bestraft, in seltenen Fällen sogar belohnt. Für den allgemeinen Fall muss also Gl. (3.13) wie folgt abgewandelt werden:

$$\mathrm{E_{Fingerprint}} = \sum_{\mathrm{i}\in\mathrm{N_e}} \mathrm{E_{Fingerprint}}\left(\mathrm{r_{z(i)}}, \mathrm{P_i}\right) + \sum_{\mathrm{i}\in\mathrm{N_m}} \mathrm{E_m}\left(\mathrm{P_i}\right) + \mathrm{N_p}\cdot\mathrm{E_p} \tag{3.15}$$

Dabei ist N_e die Indexmenge der erfüllten, N_m die Menge der unerfüllten Potentiale und N_p die Anzahl der wechselwirkenden Rotamere, die keinem Potential zugeordnet werden können.

Da es sich bei den Fingerprint-Energien um wissensbasierte Potentiale handelt, ist die Summation der Einzelenergien gleichbedeutend mit der Unterstellung von statistischer Unabhängigkeit der einzelnen Proteinpositionen. Nachdem Abhängigkeiten durch die Lösung des Zuordnungsproblems implizit berücksichtigt werden ist diese Annahme zulässig.

3.4.4 Mitrotation/-translation

Die räumliche Komponente der Fingerprint-Potentiale ist relativ zur Position des Ligandmoleküls definiert. Daher entsprechen die Lage und die Orientierung der Ellipsoide nach deren Berechnung der Situation in der Vorlagestruktur. Nachdem die Positionierung des Liganden nicht wie in der alten TransCent-Version durch die Superpositionierung von Ziel- und Vorlagestruktur erfolgt, können die Fingerprint-Potentiale nicht einfach unverändert übertragen werden. Stattdessen werden die Ligandpositionen aus der Positionsbibliothek (siehe 3.6) verwendet, für welche die „Vorlagestruktur-Potentiale“ passend verschoben und rotiert werden müssen.

Die dazu notwendigen Operationen sind dieselben, mit denen die Originalposition des Liganden in die designrelevante Pose l überführt werden kann. Diese können mit Hilfe des Kabsch-Algorithmus [117, 118] ermittelt werden. Zunächst werden dabei die geometrischen Schwerpunkte $\vec{m}_0$ und $\vec{m}_l$ der beiden Ligandpositionen sowie die Kovarianzmatrix C_{0l} der Ligandatomkoordinaten berechnet. Über die Singulärwertzerlegung der Kovarianzmatrix:

$$C_{0l} = VSW^T \tag{3.16}$$

erhält man die Rotationsmatrix R_{0l} mit der die Potential-Ellipsoide rotiert werden müssen (vgl. Abb. 3.6):

$$R_{0l} = WV^T \tag{3.17}$$

Der Translationsvektor $\vec{t}_{0l}$ für die Verschiebung ergibt sich aus der Differenz der Schwerpunktsvektoren:

$$\vec{t}_{0l} = \vec{m}_l - R\vec{m}_0 \tag{3.18}$$

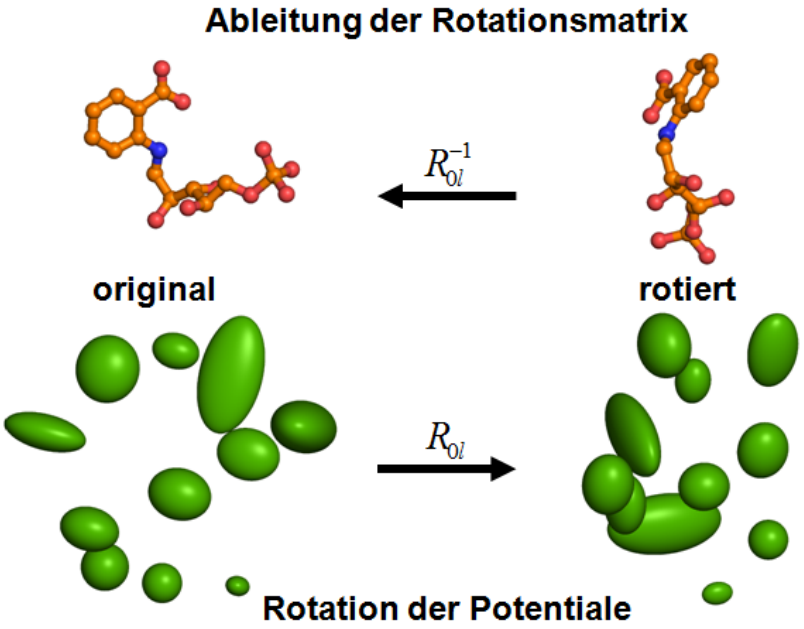


Abbildung 3.6: Schematische Darstellung von Berechnung und Anwendung der Rotationsmatrix
Die Berechnung der Rotationsmatrix R_{0l} erfolgt unter Verwendung der Ligandposition l und der Originalposition im Vorlageenzym. Die Anwendung der Rotationsmatrix auf die Ellipsoid-Matrizen führt, zusammen mit der entsprechenden Translation, zur korrekten Positionierung der Fingerprint-Potentiale.

Mit den beiden Operationen R_{0l} und $\vec{t}_{0l}$ können der Ellipsoid-Mittelpunkt $\vec{\mu}_0$ und die Ellipsoid-Matrix Σ_0 entsprechend transformiert werden:

$$\vec{\mu}_l = R_{0l}\vec{\mu}_0 + \vec{t}_{0l} \tag{3.19}$$

$$\Sigma_l = R_{0l}^T \Sigma_0 R_{0l} \tag{3.20}$$

Σ_l und $\vec{\mu}_l$ definieren dann die Orientierung und Lage der Ellipsoide relativ zur Ligandposition l.

3.5 Optimierung der pK_a-Werte essentieller Reste

Neben Proteinstabilität und Ligandenbindung ist die richtige räumliche Orientierung der Residuen im aktiven Zentrum relativ zum Liganden ein weiterer entscheidender Faktor für die Aktivität eines Enzyms. Dies ist jedoch nicht ausreichend, denn zusätzlich zur passenden räumlichen Ausrichtung müssen die Seitenketten auch im korrekten Protonierungszustand vorliegen, damit die Enzymreaktion ablaufen kann. Ob eine Seitenkette bei einem bestimmten pH-Wert protoniert ist oder nicht wird durch deren pK_a-Wert bestimmt. Dieser hängt wiederum ab von der Umgebung des Residuums im Protein, d.h. der pK_a-Wert einer Seitenkette ergibt sich aus dem pK_a-Wert für die freie Aminosäure und den Perturbationen, welche durch die benachbarten Reste verursacht werden. Somit ergeben

Titrierbare Gruppe	**Referenz-pK_a-Wert**
C-Terminus	3,2
Aspartat	3,8
Glutamat	4,5
Histidin	6,5
N-Terminus	8,0
Cystein	9,0
Tyrosin	10,0
Lysin	10,5
Arginin	12,5

Tabelle 3.2: Referenz-pK_a-Werte für titrierbare Aminosäuregruppen
Die angegebenen Werte beziehen sich auf freie Aminosäuren in wässriger Lösung.

sich beim Enzymdesign zusätzliche Nebenbedingungen bei der Wahl der Aminosäuren in und um das aktive Zentrum. Bei TransCent werden diese Nebenbedingungen durch den Energiebeitrag, den das PROPKA-Modul liefert, berücksichtigt.

Die Implementierung des Moduls, welche auf Version 1.0 des Programms PROPKA [83] basiert, wurde von [85] übernommen und auf die neue TransCent-Architektur abgestimmt. Die wichtigen Einzelheiten der Erläuterungen aus [85] werden im Folgenden zusammengefasst dargestellt.

3.5.1 Berechnung der pK_a-Werte mit PROPKA

Ausgehend von der Proteinstruktur bestimmt das Programm PROPKA alle titrierbaren Gruppen und berechnet in einem iterativen Verfahren deren pK_a-Werte. Die zugrunde liegende Gleichung:

$$pK_a = pK_{Modell} + \Delta pK_a \tag{3.21}$$

setzt sich aus einem Modell-Wert für die freie Aminosäure pK_{Modell} (vgl. Tabelle 3.2) und der Summe aller von der Proteinumgebung verursachten Perturbationen ΔpK_a zusammen. Beide Werte werden bei PROPKA empirisch bestimmt, d.h. die Parameter des Berechnungsmodells werden so gewählt, dass die Vorhersagen des Programms mit einer Menge experimentell bestimmter Daten optimal übereinstimmen.

Bei der Ermittlung der Perturbationswerte ΔpK_a werden insgesamt drei mögliche Ursachen für die pK_a-Wert-Verschiebung berücksichtigt. Am häufigsten führen Wasserstoffbrücken, welche zwischen dem titrierbaren Rest und benachbarten Seitenketten bzw. dem Proteinrückgrat ausgebildet werden, zu einer Verschiebung ΔpK_{HB}, insbesondere bei Aspartat- und Glutamat-Residuen. Darüber hinaus erfahren vergrabene Seitenketten eine pK_a-Wert-Verschiebung ΔpK_{Des}, die auf globalen und lokalen Desolvatationseffekten beruht. Die größten Änderungen ΔpK_{chgchg} werden hingegen durch Coulomb-Wechselwirkungen zwischen geladenen Gruppen hervorgerufen, die im Protein vergraben sind. Der Gesamtwert für die pK_a-Verschiebung folgt dann aus der Summe der Einzelbeiträge:

$$\Delta pK_a = \Delta pK_{HB} + \Delta pK_{Des} + \Delta pK_{chgchg} \tag{3.22}$$

Für eine detaillierte Beschreibung der einzelnen Terme sei auf [83] verwiesen.

Durch die iterative Berechnung der pK_a-Werte bei PROPKA wird berücksichtigt, dass sich die Protonierungszustände und somit auch die pK_a-Werte der ionisierbaren Reste gegenseitig beeinflussen. Das Programm verfolgt dabei die Strategie, die Protonierungszustände der Gruppen mit den extremsten pK_a-Werten als erste festzulegen, da sie mit der höchsten Zuverlässigkeit vorhergesagt werden können. Anschließend werden die Berechnungen für die noch unbestimmten Reste wiederholt, und zwar so lange, bis der Algorithmus konvergiert, d.h. alle Protonierungszustände eindeutig festgelegt sind. Unter Verwendung aller Berechnungsmodelle kann PROPKA pK_a-Werte im Mittel bis auf eine pH-Einheit genau vorhersagen. Die Parameter und Berechnungsvorschriften wurden unverändert in die TransCent-Version von PROPKA übernommen.

3.5.2 Geschwindigkeitsoptimierung der pK_a-Wert-Berechnung

Im Gegensatz zu anderen Ansätzen, welche auf der nummerischen Lösung der linearisierten Poisson-Boltzmann-Gleichung basieren, erlaubt PROPKA als empirische Methode die Berechnung der pK_a-Werte in relativ kurzer Zeit. Allerdings sind im Zusammenhang mit Proteindesign Programmlaufzeiten im Bereich weniger Sekunden um mehrere Größenordnungen zu langsam. Der *Simulated Annealing* Prozess umfasst typischerweise mehrere Millionen Einzelschritte, von denen jeder potentiell die Neuberechnung der pK_a-Werte erfordert. Somit würde eine einzige Designoptimierung mehrere Wochen in Anspruch nehmen. Daher war es unumgänglich, den PROPKA-Algorithmus an die Anforderungen des Enzymdesigns anzupassen und dabei signifikant zu beschleunigen.

Ermöglicht wird dies durch zwei Maßnahmen. Zum einen wird das rotamerbasierte Konzept von TransCent auf die PROPKA-Implementierung übertragen. Dies erlaubt, analog zur Ermittlung der Rotamerenergien mit Rosetta (vgl. 3.2.3), die Vorausberechnung und Tabellierung aller pK_a-Wert-Verschiebungen, welche zwischen zwei Rotameren auftreten können. Da die Menge an Rotameren fest vorgegeben ist, ändern sich auch die pK_a-Wert-Verschiebungen während einer Designoptimierung nicht und aufwändige Mehrfachberechnungen können so vermieden werden. Diese Strategie des *Preprocessing* findet sich in vielen Anwendungen in der Bioinformatik wieder. Die Berechnung der pK_a-Werte erfolgt dann gemäß Gleichung (3.21), wobei sich die Verschiebung ΔpK_a aus der Summe aller Beiträge der aktuellen Rotamerkonfiguration ergibt. Zusätzlich kann hierbei zwischen statischen Beiträgen, wie Wasserstoffbrücken zum Proteinrückgrat, und dynamischen Beiträgen, die von der Rotamerkonfiguration abhängen, unterschieden werden. Die statischen Effekte können direkt aufaddiert werden, wohingegen für die Bestimmung der dynamischen Beiträge zunächst eine Optimierung der Protonierungszustände durchgeführt wird.

Die zweite PROPKA-Anpassung zur Laufzeitverkürzung ergibt sich ebenfalls direkt aus der Einbettung von PROPKA in die TransCent-Optimierungsroutine. Da bei einem *Simulated Annealing* Schritt nur ein Rotamer getauscht wird, bleibt der größte Teil des Designmodells unverändert. Anstatt eine vollständige Neuberechnung aller pK_a-Werte durchzuführen, ist es zulässig und sinnvoll, nur die Änderungen einzuarbeiten, welche durch den Rotamerwechsel zustande kommen. Die Beiträge des abgewählten Rotamers werden dabei gelöscht und stattdessen die Effekte des neuen Rotamers addiert. Die prinzipiell nachfolgende Neuberechnung der Protonierungszustände erfolgt nur, falls der Rotamertausch auch tatsächlich zu einer Änderung der pK_a-Wert-Verschiebungen geführt hat. Ansons-

ten wird einfach der bestehende Zustand übernommen. Dadurch wird der Rechenaufwand weiter optimiert.

Insgesamt wird durch die Anpassungen die pK_a-Wert-Berechnung in TransCent etwa um den Faktor 5000 gegenüber der Standard-PROPKA-Version beschleunigt [85], ohne dass dabei die Vorhersagequalität des Moduls beeinträchtigt wird. Dieser Geschwindigkeitsgewinn kommt allerdings nur während der Designoptimierung zum Tragen. Die einmalige Berechnung der pK_a-Werte eines Proteins bleibt davon unbeeinflusst.

3.5.3 Beitrag des PROPKA-Moduls zur TransCent-Energiefunktion

Um bei der Designoptimierung berücksichtigt werden zu können, müssen die berechneten pK_a-Werte in einen Energieterm übersetzt werden, welcher dann in die TransCent-Energiefunktion (Gl. (3.1)) eingeht. Gemäß der Zielsetzung des Moduls, die pK_a-Werte der katalytisch essentiellen Reste korrekt einzustellen, hat die PROPKA-Energie die Form eines Strafterms, welcher Abweichungen von den vorgegebenen Referenz-pK_a-Werten mit einem energetischen Malus belegt:

$$\mathrm{E}_{\mathrm{PROPKA}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{w_i} \cdot \left| \mathrm{pK_{a\,i}^{Modell}} - \mathrm{pK_{a\,i}^{Referenz}} \right|^{\mathrm{p_{PROPKA}}} \tag{3.23}$$

Für N titrierbare Reste wird die betragsmäßige Differenz zwischen dem pK_a-Wert bei der aktuellen Rotamerkonfiguration $pK_{a\,i}^{\mathrm{Modell}}$ und dem angestrebten Referenzwert $pK_{a\,i}^{\mathrm{Referenz}}$ gewichtet aufsummiert. Der funktionale Zusammenhang kann über den Exponenten p_{PROPKA} eingestellt werden. Für diesen wird der Wert 1,5 aus [85] übernommen. Die Gewichte w_i entsprechen den relativen Häufigkeiten der katalytisch essentiellen Aminosäuren, für die die Referenz-pK_a-Werte berechnet wurden. Diese werden aus den entsprechenden Spalten des MSAs homologer Sequenzen für die Strukturbibliothek (siehe 3.4.1) abgeleitet.

Die Differenzbildung in der Energiefunktion führt zu einem zusätzlichen positiven Effekt. Da die pK_a-Wert-Berechnung mit PROPKA im Mittel nur auf eine pH-Einheit genau ist, führen systematische Fehler ggf. dazu, dass ein Referenz-pK_a-Wert zu hoch oder zu niedrig angesetzt wird. Diese Fehler treten jedoch sowohl bei der Referenzrechnug als auch während der Designoptimierung auf und werden aufgrund der Differenzbildung eliminiert, so dass sie sich nicht negativ auf das Designergebnis auswirken.

In Gleichung (3.23) ist nicht berücksichtigt, dass vor der Energieberechnung zunächst ein Zuordnungsproblem zu lösen ist. *A priori* ist nämlich nicht klar, welche pK_a-Werte miteinander verglichen werden müssen. Bei Rekapitulationsrechnungen (siehe 3.10.1) ist die Situation vergleichsweise einfach, da die Referenz-pK_a-Werte eindeutig mit bestimmten Positionen im Enzym identifiziert werden können. Allerdings ist auch hier eine pK_a-Berechnung nur sinnvoll, wenn das aktuelle Rotamer einem titrierbaren Rest entspricht. Andernfalls ist eine pK_a-Wert-Bestimmung gar nicht möglich. Bei Funktionsübertragungen ist schon die Zuordnung eines Referenz-pK_a-Wertes zu einer Position im Enzym *a priori* nicht durchführbar.

Die Lösung dieses Problems gelingt durch die Verknüpfung des PROPKA-Moduls mit dem Fingerprint-Modul (vgl. 3.4). Statt mit einer Proteinposition wird ein Referenz-pK_a-Wert mit einem Fingerprint-Potential assoziiert, wobei sich das Zuordnungsschema ganz natürlich aus der Vorlagestruktur ergibt. Während einer Designoptimierung wird also immer zunächst das Fingerprint-Modul aufgerufen, welches das Assoziationsmuster zwischen

Positionen und Potentialen ermittelt. Darauf aufbauend erfolgt dann die Berechnung der pK_a-Werte und der PROPKA-Energie. Allerdings ist je nach Wahl der Rotamere nicht für jedes Fingerprint-Potential eine korrespondierende Seitenkette vorhanden. Dementsprechend kann es auch Referenz-pK_a-Werte geben, für die kein passender titrierbarer Rest existiert. In solchen Fällen wird eine maximale Abweichung von sieben pH-Einheiten angenommen, während für die anderen der berechnete Differenzwert in die Energiefunktion (Gl. (3.23)) eingeht. Dadurch wird auch das Nichterfüllen einer pK_a-Wert-Nebenbedingung energetisch bestraft.

Die Verknüpfung der beiden Module hat auch zur Folge, dass das Fingerprint-Modul bei einer Designoptimierung immer „im Hintergrund“ mitlaufen muss, wenn eine pK_a-Wert-Optimierung durchgeführt werden soll, auch wenn es selbst eigentlich deaktiviert ist. Die Zuordnungsroutine (siehe 3.4.3) wird also ausgeführt, jedoch ohne Einfluss auf die Gesamtenergiefunktion. Außerdem ist das PROPKA-Modul über den Fingerprint mit der Ligandpositionierung verknüpft, da der Fingerprint relativ zum Liganden definiert ist.

Aufgrund der iterativen Optimierung der Protonierungszustände und der Verknüpfung mit der Fingerprintzuordnung handelt es sich bei E_{PROPKA} um einen Mehr-Körper-Energieterm.

3.6 Verwendung einer Bibliothek von Ligandpositionen

Das zentrale Thema dieser Arbeit ist die Erweiterung von TransCent für die flexible Handhabung der Ligandposition. Die korrekte Positionierung des Liganden entscheidet zusammen mit der Wahl eines geeigneten Proteingerüsts über den möglichen Erfolg eines Enzymdesigns. Üblicherweise verwendet man Dockingalgorithmen wie z.B. Glide [119], GOLD [120], FlexX [121] oder AutoDock [122] zur Suche nach katalytisch aktiven Ligandposen. Diese Möglichkeit entfällt jedoch beim Enzymdesign, da das aktive Zentrum, in welches der Ligand gedockt werden soll, erst noch modelliert werden muss. Prinzipiell bestünde die Möglichkeit, ähnlich wie bei einer Moleküldynamik-Simulation den Liganden während der Designoptimierung „mitzubewegen“. Nach einigen *Simulated Annealing*-Schritten wird für den Liganden die für die aktuelle Rotamerkonfiguration optimale Pose ermittelt und dieser dann entsprechend verschoben. Davon ausgehend folgt wieder eine Runde Seitenkettenoptimierung und schließlich wird so iterativ die bestmögliche Kombination aus Ligandposition und Rotamerkonfiguration gefunden. Der Nachteil dieser Vorgehensweise ist, dass alle Interaktionsenergien zwischen Ligand und Protein nach jeder Verschiebung neu berechnet werden müssen, was zu einer dramatischen Zunahme der Rechenzeit führt.

TransCent verwendet stattdessen die neu entwickelte Methode TransLig (siehe 3.8), um geeignete Ligandpositionen im Voraus zu identifizieren. Diese werden dann, in konsequenter Erweiterung des rotamerbasierten Rahmenkonzeptes von TransCent, als „Pseudorotamere“ des Ligandmoleküls behandelt (vgl. Abb. 3.7). Analog zu den „echten“ Rotameren der Aminosäureseitenketten werden die Ligandpositionen als Wahlmöglichkeiten beim *Simulated Annealing* präsentiert, wodurch der Suchraum um eine Dimension erweitert wird. Während der Optimierung wird durch zufällige Pseudorotamer-Wechsel zwischen zwei Ligandpositionen versucht, die bestmögliche Position des Liganden zu ermitteln.

Ob ein *Simulated Annealing*-Schritt bei der Ligandpositionierung akzeptiert wird, wird ebenfalls über das Metropolis-Kriterium (siehe Gl. (3.2)) entschieden. Bei der Berechnung

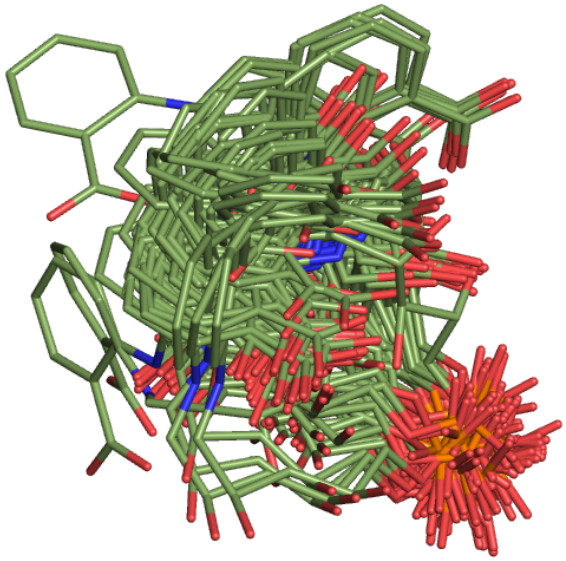


Abbildung 3.7: Überlagerung mehrerer Positionen einer Ligandpositionsbibliothek
Die Abbildung zeigt die Überlagerung von 100 identischen Ligandmolekülen, von denen jedes eine Position der Ligandpositionsbibliothek repräsentiert.

der Energiedifferenz ΔE müssen allerdings nur drei der vier Modulenergien berücksichtigt werden. Die Rosetta-Energie umfasst nur Wechselwirkungen innerhalb des Proteins und ist somit unabhängig von der Ligandposition.

Da die Fingerprint-Potentiale relativ zum Liganden definiert sind, hat hier jede Ligandposition ihren eigenen Satz an Fingerprint-Energien, welche während des Preprocessing für jedes Rotamer berechnet werden (vgl. 3.4.3). Die Differenz in der Fingerprint-Energie resultiert also zum einen aus der unterschiedlichen Bewertung der einzelnen Rotamere. Zum anderen kann sich aber beim Ligandpositionswechsel auch eine neue Zuordnung zwischen Fingerprint-Potentialen und Proteinpositionen ergeben (vgl. 3.4.3) und so zu einer Änderung der Fingerprint-Energie führen.

Zusätzlich kann sich durch den Zuordnungswechsel aufgrund der Verknüpfung mit dem Fingerprint-Modul die PROPKA-Energie ändern (siehe 3.5.3). Nachdem die Ligandatome bei der pK_a-Wertbestimmung nicht berücksichtigt werden (siehe 3.5.1), ist allerdings keine Neuberechnung der pK_a-Werte beim Wechsel der Ligandposition notwendig.

Der DSX-Energieterm wird völlig analog zur Rosetta-Energie bei den Proteinrotameren behandelt. Die Wechselwirkungen mit dem fixen Proteinrückgrat entsprechen einer Ein-Körper-Energie, welche nur von der Position des Liganden abhängt, während die Interaktionen mit den Seitenketten als Zwei-Körper-Energieterm berücksichtigt werden (vgl. 3.3.2). Beide Beiträge ergeben dann die DSX-Energiedifferenz beim Ligandpositionswechsel. Insgesamt resultiert daraus eine Wahrscheinlichkeit

$$p = \exp\left(-\frac{(\Delta E_{\mathrm{DSX}} + \Delta E_{\mathrm{Fingerprint}} + \Delta E_{\mathrm{PROPKA}})}{T_i}\right) \tag{3.24}$$

für das Akzeptieren der neuen Ligandposition bei einer Pseudotemperatur T_i.

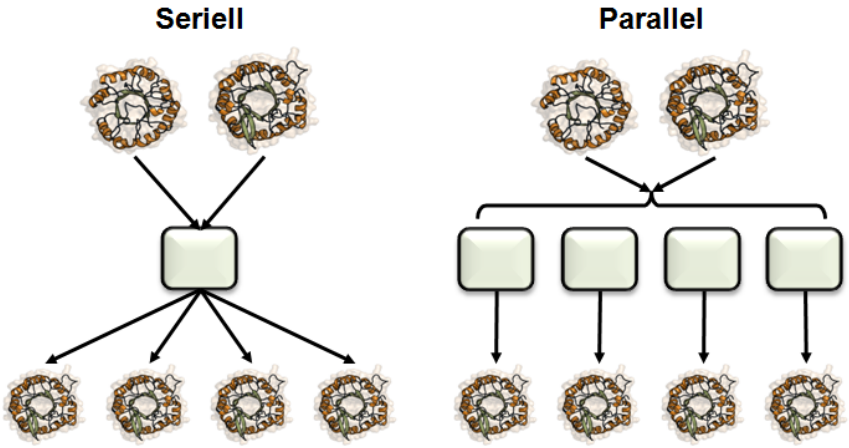


Abbildung 3.8: Schematische Darstellung der Parallelisierung von TransCent Im Gegensatz zum seriellen Fall werden beim parallelisierten Programm von mehreren Threads *Simulated Annealing* Optimierungen durchgeführt und Designmodelle erstellt. Die Berechnungen erfolgen auf einer gemeinsamen Datenbasis, welche nur einmal im Speicher vorgehalten wird. Die Laufzeitbeschleunigung ist dann optimal, wenn die Anzahl der Threads gleich der Anzahl der Prozessorkerne ist.

3.7 Parallelisierung des Designalgorithmus

Bei der Entwicklung von Prozessoren geht der Trend hin zu einer steigenden Anzahl von Kernen pro CPU. Diese erlauben die gleichzeitige Verarbeitung mehrerer Aufgaben und somit, zumindest prinzipiell, eine Verkürzung der Rechenzeit. Voraussetzung dafür ist, dass die Software, die auf dem Prozessor läuft, das parallele Arbeiten auch unterstützt. Aufgrund der gestiegenen Laufzeitkomplexität von TransCent und der verbesserten Hardware-Ausstattung wurde das Programm für eine parallelisierte Ausführung angepasst. Dabei handelt es sich um eine triviale Parallelisierung der *Simulated Annealing* Optimierung.

Typischerweise wird beim Enzymdesign nicht nur ein Modell berechnet, sondern einige hundert bis mehrere tausend Designoptimierungen durchgeführt. Dabei ist die Datenbasis, also im Falle von TransCent die Energietabellen und die zur Wahl stehenden Rotamere (vgl. 3.2.2), bei jeder *Simulated Annealing* Optimierung dieselbe. Es genügt also, die Datenbasis einmal zu berechnen bzw. einzulesen, um anschließend darauf gleichzeitig mehrere Optimierungen durchzuführen. In der parallelisierten Version von TransCent ist dies so realisiert, dass mehrere Threads gestartet werden können, von denen jeder eine Designoptimierung übernimmt (siehe Abb. 3.8).

Der modulare Aufbau von TransCent wurde auch bei der Parallelisierung berücksichtigt. So können auch Designoptimierungen mit unterschiedlichen Gewichten und unterschiedlichen Modulkombinationen gleichzeitig ausgeführt werden.

3.8 TransLig

Die Positionierung des Liganden ist für den Erfolg eines Designs von zentraler Bedeutung, da dadurch bestimmt wird, ob an den designbaren Positionen die Seitenketten so

gewählt werden können, dass sie die katalytisch essentiellen Wechselwirkungen zum Liganden eingehen. Andererseits ist zu Beginn des Designprozesses unklar, welche Position im Protein für welche Seitenkette bzw. Wechselwirkung geeignet ist. Daher ist auch die Wahl der Ligandposition nicht eindeutig. Als Lösung dieses Dilemmas bietet es sich an, dem Designprozess solche Ligandpositionen zu präsentieren, welche im Prinzip das Setzen der entsprechenden Seitenketten erlauben.

Friedemann Paulini hat diese Strategie im Rahmen seiner Diplomarbeit umgesetzt und dafür das Modul TransLig entwickelt [123]. Es liefert Vorschläge für geeignete Ligandpositionen im aktiven Zentrum des Zielproteins unter Berücksichtigung vorgegebener Abstandskriterien.

3.8.1 Motivation

Das Ziel von TransLig ist, den Liganden im Zielprotein so zu positionieren, dass im Designschritt die Freiheit bei der Wahl der Seitenketten möglichst wenig eingeschränkt wird. Ist der Ligand zu nah am Rückgrat des Proteins platziert, so kommen aus sterischen Gründen nur sehr kleine Aminosäuren wie Alanin oder Glycin in Frage. Befindet sich der Ligand hingegen zu weit weg vom aktiven Zentrum, so kann dieser, wenn überhaupt, nur mit sehr langen Seitenketten wie Lysin oder Arginin erreicht werden. Dadurch ist die Bandbreite designbarer Wechselwirkungen stark limitiert. Daraus ergibt sich ein optimaler Abstandsbereich, in dem der Ligand zu liegen kommen sollte.

Andererseits unterscheiden sich die „Anforderungen“ der einzelnen Ligandatome gegebenenfalls beträchtlich. Zwischen den Sauerstoffatomen von Phosphatgruppen und Stickstoffatomen des Proteinrückgrats bilden sich häufig Wasserstoffbrücken aus [124, 125]. Deswegen muss der Abstand in diesem Fall relativ klein gewählt werden. Für andere Ligandatome, deren positive Ladung durch eine entsprechende negative Ladung stabilisiert werden muss, ergibt sich der optimale Abstandsbereich aus der Länge der entsprechenden Aspartat- oder Glutamat-Seitenkette. Wieder andere gehen überhaupt keine Wechselwirkungen mit dem Enzym ein und erfordern daher keine speziellen Abstandskriterien.

Da vor Abschluss des Designschritts weder Art noch Lage der Seitenketten im aktiven Zentrum des Zielproteins bekannt sind, kommen als Abstandskriterien die optimalen Distanzen zwischen den tatsächlich wechselwirkenden Atomen nicht in Frage. Stattdessen können die Abstände der Ligandatome zu den Atomen des Proteinrückgrates und den C_β-Atomen betrachtet werden, da deren Position fest vorgegeben ist.

All dies berücksichtigt das Modul TransLig bei der Suche nach Ligandpositionen, dessen Umsetzung detailliert in [123] beschrieben ist. Im Folgenden werden die wichtigsten Aspekte zusammengefasst erläutert.

3.8.2 Algorithmus zur Bestimmung der Ligandpositionen

Der Optimierungsalgorithmus von TransLig ermittelt geeignete Ligandpositionen in einem iterativen Prozess durch abwechselnde Rotation und Translation des Ligandmoleküls im „leergeräumten“ aktiven Zentrum des Zielproteins. Dies bedeutet, dass die Seitenketten der mutierbaren Proteinpositionen (vgl. 4.2.1) auf das C_β-Atom gestutzt sind.

Gesteuert wird die Bewegung durch ein Kräftemodell (siehe 3.8.2.2), aus dem sich die Bewegungsrichtung und die Drehachse ableiten. Der Ligand wird dabei als starr angenommen, da unterstellt wird, dass er in der Kristallstruktur eine katalytisch aktive Konformation einnimmt, d.h. es werden keine inneren Freiheitsgrade des Liganden berücksichtigt. Eine geeignete Position ist dann gefunden, wenn die Abstände der Ligandatome zu den Proteinatomen mit den vorgegebenen Distanzen möglichst gut übereinstimmen. Durch die Verwendung des Kräftemodells und der dazugehörigen Energiefunktion entspricht die Optimierung einer Suche nach dem globalen Minimum in der Energielandschaft. Eine Nebenbedingung ist, dass Kollisionen mit den Proteinatomen vermieden werden müssen, d.h. ein minimaler Abstand zwischen Ligand- und Proteinatom darf nicht unterschritten werden.

Der Algorithmus unterteilt dabei die Problemstellung in zwei Teilprobleme. Zunächst muss eine Startgeometrie gefunden werden, die dem Liganden die größtmögliche Bewegungsfreiheit im aktiven Zentrum zugesteht. Im zweiten Schritt folgt dann die gerichtete Suche nach lokalen Minima der Energiefunktion.

3.8.2.1 Bestimmung der Startposition

Im ersten Schritt platziert TransLig den Liganden im zu designenden aktiven Zentrum so, dass er ungehindert bewegt werden kann. Dazu wird zunächst das 3D-Voronoi-Diagramm des aktiven Zentrums berechnet [126] und mit dessen Hilfe die größte leere Kugel innerhalb des aktiven Zentrums bestimmt. Der Ligand wird dann in dieser Kugel positioniert und zwar so, dass der Schwerpunkt des Liganden auf dem Kugelmittelpunkt zu liegen kommt. Falls Kollisionen mit Proteinatomen vorliegen sollten, wird der Ligand noch in eine kollisionsfreie Position rotiert. Eine Kollision liegt dann vor, wenn der Abstand d zweier Atome, gemessen von den Atommittelpunkten aus, die Summe der beiden Van-der-Waals-Radien abzüglich eines Korrekturterms Δ unterschreitet.

$$d\left(\mathrm{Atom_A}, \mathrm{Atom_B}\right) < R^{\mathrm{VdW}}_{\mathrm{Atom_A}} + R^{\mathrm{VdW}}_{\mathrm{Atom_B}} - \Delta \tag{3.25}$$

Der Korrekturterm ist in TransLig auf $\Delta = 0,3$ Å gesetzt und bewirkt zweierlei. Zum einen kann dadurch berücksichtigt werden, dass bindungsvermittelnde Atompaare häufig ihren „normalen“ Abstand unterschreiten [114]. Zum anderen erlangt TransLig dadurch ein gewisses Maß an Toleranz gegenüber Fehlern, die aufgrund der beschränkten Auflösung von Kristallstrukturen beim Ableiten der optimalen Abstände gemacht werden (siehe 3.8.3).

Um eine umfassendere Abtastung des Suchraumes zu erreichen, können auch weitere Knoten des Voronoi-Pfades als Startpunkte gewählt werden.

3.8.2.2 Das Kräftemodell

Ausgehend von der Startgeometrie wird der Ligand abwechselnd rotiert und verschoben, wobei das Kräftemodell die Bewegungsrichtung bestimmt. Im einfachsten Fall würde der Ligand aus einem einzelnen Atom bestehen. Die optimale Position für diesen Ein-Atom-Liganden erhält man dann, wenn die Distanz d zum nächsten (passenden) Proteinatom exakt dem vorgegebenen Abstand d^{opt} entspricht. Um den Liganden dort hin zu bewegen, kann man sich vorstellen, dass er über eine Feder mit dem Proteinatom verbunden ist, die

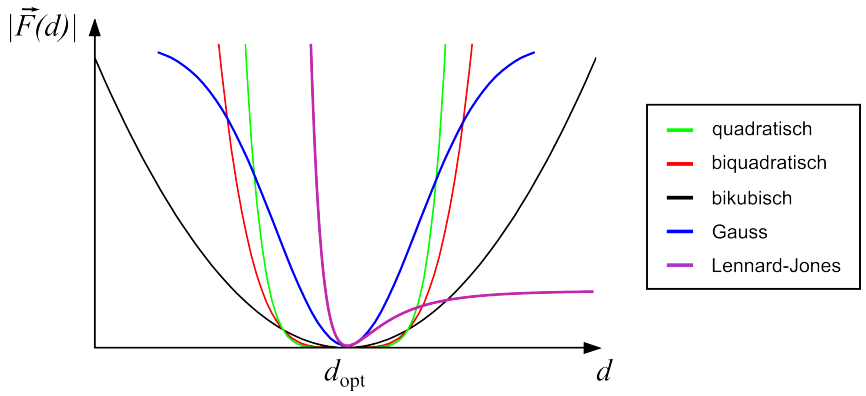


Abbildung 3.9: Qualitativer Vergleich der TransLig-Kraftfunktionen
Die Kurvenverläufe sind für die Darstellung angepasst und entsprechen nicht der tatsächlichen Skalierung.

so lange eine Kraft $\vec{\mathcal{F}}$ auf das Ligandatom ausübt, bis sich der Ligand im Energieminimum befindet. Die Kraft wird in Anlehnung an das Hookesche Gesetz wie folgt berechnet:

$$\vec{\mathcal{F}} = -K \cdot sgn\left(d - d^{\text{opt}}\right) \cdot \left|d - d^{\text{opt}}\right|^n \frac{\vec{r}_{\text{Ligandatom}} - \vec{r}_{\text{Proteinatom}}}{\|\vec{r}_{\text{Ligandatom}} - \vec{r}_{\text{Proteinatom}}\|} \qquad n \in \mathbb{N}^+ \tag{3.26}$$

Die Richtung der Kraft leitet sich aus den Positionen des Ligandatoms $\vec{r}_{\text{Ligandatom}}$ und des Proteinatoms $\vec{r}_{\text{Proteinatom}}$ ab und zeigt entweder zum Proteinatom hin ($d > d^{\text{opt}}$) oder von diesem weg ($d < d^{\text{opt}}$). Der Betrag der Kraft wird bestimmt durch die Pseudo-Federkonstante K, welche bei der Implementierung den Wert 2 hat, und den Betrag der Differenz von aktuellem und angestrebtem Abstand $\left|d - d^{\text{opt}}\right|^n$, wobei der Exponent n die Werte 2, 4 und 6 annehmen kann. Alle Abstände sind dabei in Ångström angegeben, was dazu führt, dass bei der Wahl eines größeren Exponenten Abweichungen größer 1 Å stärker bestraft werden, wohingegen kleinere Abweichungen eher toleriert werden. Darüber hinaus sind auch zwei Varianten implementiert, deren Kraftfunktionen vom Lennard-Jones Potential bzw. von einer Normalverteilung abgeleitet sind (siehe 3.9).

Das Kräftemodell lässt sich ohne weiteres auf einen mehratomigen Liganden erweitern, bei dem dann jedes Ligandatom über eine Feder mit dem ihm zugeordneten Proteinatom verbunden ist. Die Bewegung des Ligandmoleküls wird dann durch die Summe aller wirkenden Kräfte $\vec{\mathcal{F}}_{\text{ges}}$ bestimmt (siehe Abb. 3.10). Die resultierende Gesamtkraft greift am Schwerpunkt des Liganden an und gibt die Translationsrichtung vor.

$$\vec{\mathcal{F}}_{\text{ges}} = \sum_{\text{Ligandatome}} a_i \cdot \vec{\mathcal{F}}_{\text{i}} \tag{3.27}$$

Über die Wahl der Koeffizienten a_i können die Beiträge einzelner Ligandatome zur Gesamtkraft gewichtet werden. Der Algorithmus versucht also, bestimmte Abstände bevorzugt einzuhalten und bestraft Abweichungen von den vorgegebenen Distanzen stärker. Durch diese Anpassung des Kräftemodells kann also berücksichtigt werden, dass manche Ligandatome „wichtiger“ sind als andere, da sie z.B. Wasserstoffbrücken ausbilden sol-

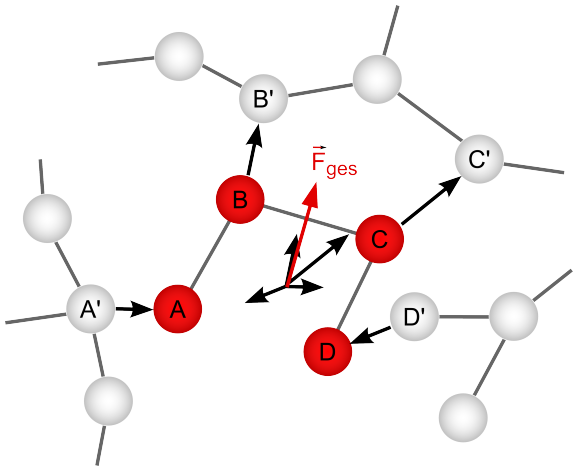


Abbildung 3.10: Vektoraddition der Einzelkräfte für die Translationsrichtung
Die Summe $\vec{\mathcal{F}}_{\text{ges}}$ der einzelnen Kräfte zwischen Ligandatom X und zugeordnetem Proteinatom X' gibt die Bewegungsrichtung für den Translationsschritt vor. Angriffspunkt der Gesamtkraft ist der Schwerpunkt des Liganden. Nach Abbildung 4.2 in [123].

len oder am chemischen Schritt der Enzymfunktion direkt beteiligt sind und daher ihre korrekte Positionierung essentiell für den Erfolg des Designs ist.

Bei einem ausgedehnten Liganden können zusätzlich zu den Verschiebungskräften auch Drehmomente wirken. Der Ligandschwerpunkt $\vec{r}_{\text{Schwerpunkt}}$ wird als Bezugspunkt für die an den Ligandatomen angreifenden Kräfte $\vec{\mathcal{F}}$ definiert und das Vektorprodukt aus Kraft und Hebelarm ergibt dann das Drehmoment:

$$\vec{\mathcal{M}} = (\vec{r}_{\text{Ligandatom}} - \vec{r}_{\text{Schwerpunkt}}) \times \vec{\mathcal{F}} \tag{3.28}$$

Über die gewichtete Vektorsumme

$$\vec{\mathcal{M}}_{\text{ges}} = \sum_{\text{Ligandatome}} a_i \cdot \vec{\mathcal{M}}_{\text{i}} \tag{3.29}$$

erhält man das Gesamtdrehmoment $\vec{\mathcal{M}}_{\text{ges}}$, welches die Richtung der Drehachse bestimmt.

3.8.2.3 Optimierung der Abstände

Die Optimierung der Abstände erfolgt bei TransLig in einem iterativen Prozess mit abwechselnden Translations- und Rotationsschritten, wobei die Richtung der Verschiebung und der Drehung bei jeder Iteration von neuem gemäß dem Kräftemodell bestimmt wird. Vor der Berechnung der Kräfte muss allerdings zunächst eine Zuordnung zwischen Ligandatomen und Proteinatomen hergestellt werden.

Ob ein Proteinatom grundsätzlich als Partneratom in Frage kommen kann, ist im Temperaturfaktor der Eingabe-PDB-Datei kodiert. Nur solche Atome, deren B-Faktor auf einen Wert von 1.00 gesetzt ist, dürfen bei der Zuordnung überhaupt in Betracht gezogen werden, wohingegen solche mit einem Wert von 0.00 unberücksichtigt bleiben. Bei der

Untersuchung von Ligandpositionen auf Kollisionen werden jedoch konsequenterweise alle Atome einbezogen.

Darüber hinaus wird für jedes Ligandatom eine bevorzugte Atomart definiert, wodurch die Wahl möglicher Partneratome weiter eingeschränkt ist. TransLig verwendet zur Spezifikation der Atomart die PDB-Nomenklatur (CA = C_α-Atom, CB = C_β-Atom, N = Rückgratstickstoff). Von den verbliebenen Proteinatomen wird dann dasjenige mit dem kleinsten Abstand zum Ligandatom als dessen Partner definiert und die Kräfte können damit entsprechend berechnet werden. Für C_β-Partner gilt zusätzlich die Regelung, dass der Winkel zwischen Ligandatom, C_β-Atom und dem dazu gehörenden C_α-Atom größer als 90° sein muss. Andernfalls wird das C_β-Atom verworfen und stattdessen das C_α-Atom gewählt. Hintergrund dieser Einschränkung ist, dass C_β-Atome immer in Richtung des Liganden „zeigen“ sollen, um zu gewährleisten, dass die Seitenketten, welche an die entsprechenden Positionen modelliert werden, dies ebenfalls tun und Wechselwirkungen ausbilden können.

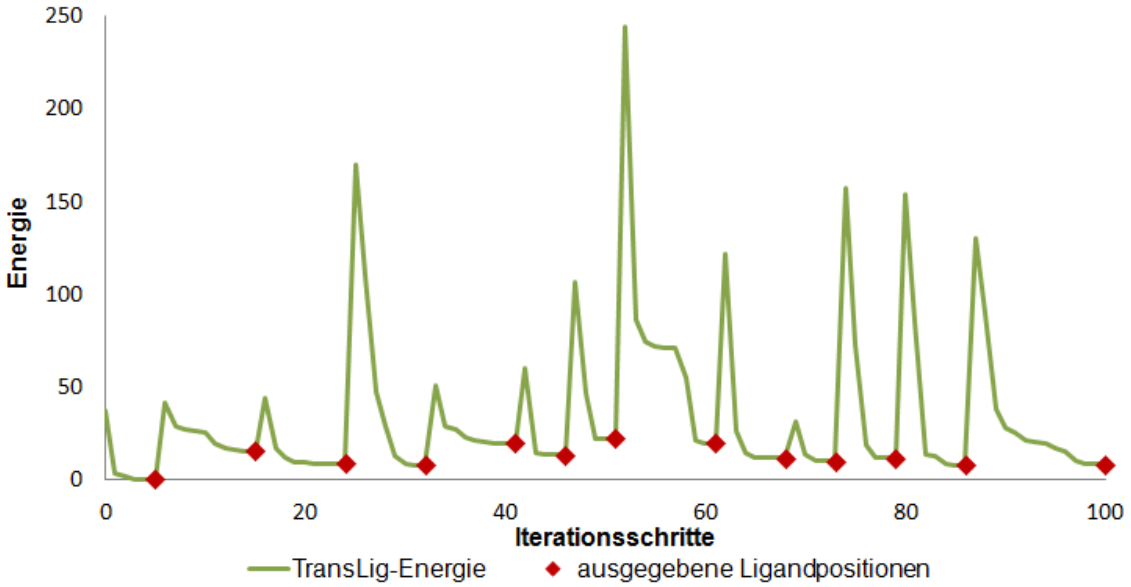


Abbildung 3.11: Beispielhafter Verlauf der TransLig-Energie bei der Suche nach geeigneten Ligandposition
Die Kurve zeigt den Energieverlauf für 100 Iterationsschritte ausgehend von einer Startposition. Die roten Rauten markieren die lokalen Energieminima, welche als mögliche Ligandpositionen von TransLig ausgegeben werden.

Sind die Gesamtkraftrichtung bzw. der Gesamtdrehmomentvektor berechnet, so tastet der Algorithmus einen Bereich um die aktuelle Position ab, bewegt den Liganden in das (lokale) Energieminimum dieses Bereichs und führt die Berechnungen ausgehend von der neuen Ligandposition fort. Bei Translationen wird ein Bereich von ±3 Å mit einer Auflösung von 0,1 Å abgetastet. Für Rotationen sucht der Algorithmus den kleinsten Wert in einem Winkelbereich von ±120° in Schritten von 0,5°.

Der Algorithmus terminiert, wenn der Ligand sich nicht mehr bewegt, d.h. dasselbe lokale Minimum zweimal in Folge gefunden wird. Die so bestimmte Position ist eine mögliche

Lösung des Positionierungsproblems und wird von TransLig ausgegeben. Allerdings ist dadurch nicht gesichert, dass bereits das globale Energieminimum erreicht ist. Um die Suche fortführen zu können, wird eine zufällige Störung auf die Ligandposition ausgeübt und die Berechnung auf der so erzeugten Anordnung fortgeführt (vgl. Abb. 3.11).

Um die Energielandschaft noch besser abtasten zu können, werden zufällige Störungen auch auf die Startpositionen angewendet. Als Parameter können dem Programm die Anzahl der Startpunkte, die Zahl der Variationen je Startpunkt, die Anzahl der Iteration je Startpunktvariation sowie die Art der Kraftfunktion übergeben werden.

3.8.3 TransLig in Verbindung mit TransCent

Bei der bisherigen Beschreibung von TransLig wurde immer betont, dass die Abstände zwischen Ligand- und Proteinatomen optimiert werden sollen. Allerdings wurde bei den bisherigen Ausführungen darauf verzichtet, diese optimalen Distanzen näher zu spezifizieren, da dies für den allgemeinen Fall nicht möglich ist. Nachdem TransLig aber Ligandpositionen für eine Funktionsübertragung mit TransCent liefern soll, können im konkreten Anwendungsfall die erwünschten Abstände bestimmt werden. Die Vorgehensweise ist dabei konsistent mit der „TransCent-Philosophie“, denn die optimalen Distanzen werden so übernommen, wie sie in der Ausgangsstruktur vorliegen, da sie sich dort als zielführend erwiesen haben (vgl. Abb. 3.12). Gleiches gilt für die Wahl der bevorzugten Partner-Atomart.

Zusätzlich zur Festlegung der Abstandsbedingungen muss auch eine Auswahl für die erlaubten Protein-Partneratome im Zielprotein getroffen werden. Prinzipiell kommen dafür alle C_β-Atome sowie die Atome des Proteinrückgrats in Frage. Allerdings ist es sinnvoll, den Suchraum geeignet zu beschränken und so die Suche nach passenden Positionen zu erleichtern. Ist das Zielprotein ebenfalls ein Enzym, bietet es sich daher häufig an, das bereits bestehende aktive Zentrum wiederzuverwenden und, nach Entfernung der „überschüssigen“ Seitenkettenatome, die entsprechenden Positionen über die Temperaturfaktoren als erlaubte Partner zu markieren. Falls auch Positionen, die im Zielprotein wildtypisch mit einem Glycin besetzt sind, als C_β-Partner berücksichtigt werden sollen, müssen diese um „virtuelle“ C_β-Atome erweitert werden.

Im Allgemeinen ist aufgrund des unterschiedlichen Verlaufs des Proteinrückgrats von Vorlage- und Zielprotein nicht zu erwarten, dass im aktiven Zentrum des Zielproteins eine Ligandposition existiert, bei der alle Abstandsbedingungen gleichzeitig exakt erfüllt werden können. Stattdessen kann TransLig dazu verwendet werden, Positionen zu finden, die den gestellten Anforderungen möglichst nahe kommen. Sofern die Unterschiede nicht zu groß sind, können die Abweichungen von der optimalen Ligandgeometrie durch geschickte Wahl der Rotamere bzw. Aminosäuren kompensiert und trotzdem alle wichtigen Wechselwirkungen im Design berücksichtigt werden.

3.8.3.1 Spezifikation individueller Interaktionspartner

TransLig bietet eine zusätzliche Möglichkeit, die Zuordnung zwischen Ligand- und Proteinatom zu präzisieren. Sowohl bei den Ligandatomen als auch bei den Proteinatomen können Teilmengen spezifiziert werden, welche ausschließlich einander zugeordnet werden dürfen. Dadurch kann zum Beispiel erreicht werden, dass ein bestimmtes Ligandatom den

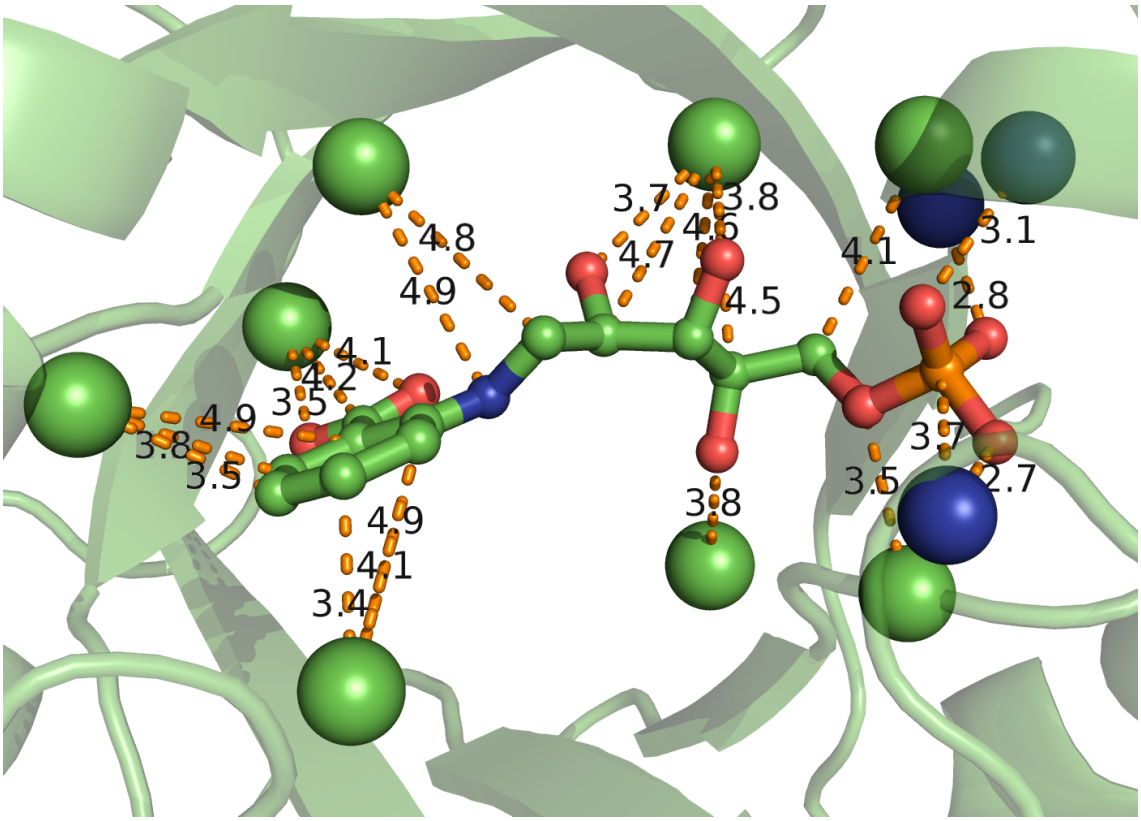


Abbildung 3.12: Ableitung der Distanzvorgaben aus der Ausgangsstruktur für TransLig
Die Abbildung veranschaulicht das Ableiten der Distanzvorgaben anhand der Protein-Ligand-Komplexstruktur von PriA aus *Mycobacterium tuberculosis* (PDB-Code: 2y85). Das Ligandmolekül ist im *Ball & Stick*-Modell dargestellt. Die gestrichelten orangen Linien repräsentieren die Abstände der Ligandatome zum jeweils nächsten C_β-Atom, C_α-Atom (grüne Kugeln) oder Rückgrat-Stickstoffatom (blaue Kugeln), welche als Distanzvorgaben verwendet werden. Die Abstände sind in Å angegeben. Das Proteinrückgrat ist in Cartoon-Darstellung wiedergegeben.

vorgegebenen Abstand zu einem eindeutig definierten C_β-Atom einhalten muss. Ziel dieser Erweiterung ist es, Teile des Liganden zielgerichtet im aktiven Zentrum zu platzieren. Falls ein Ligand z.B. eine Phosphatgruppe besitzt und das Zielprotein im wildtypischen Zustand bereits eine Phosphatbindestelle aufweist, so bietet es sich an, diese „wiederzuverwenden“ und die Phosphatgruppe durch geschickte Wahl der erlaubten Partneratome darin zu verankern.

3.8.4 Bewertung der TransLig-Vorschläge

Da TransLig bei der Suche nach geeigneten Ligandpositionen im Allgemeinen mehr als einen Vorschlag liefert, werden Maße benötigt, mit denen die erzeugten Positionen verglichen und bewertet werden können.

Der RMSD-Wert (engl. *root mean square deviation*) ist ein in der Bioinformatik übliches Maß für den durchschnittlichen Abstand zwischen zwei Mengen von Atomen und wird unter anderem beim Vergleich von Proteinstrukturen oder Ligandkonformationen verwendet. Zur Berechnung des RMSD müssen zunächst zwei Teilmengen gefunden werden, die bijektiv einander zugeordnet werden können. Damit erhält man den RMSD-Wert der Atommengen A und B wie folgt:

$$\mathrm{RMSD}\,(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\|\vec{r}_{A,i} - \vec{r}_{B,i}\right\|^2} \tag{3.30}$$

Die Koordinatenvektoren $\vec{r}_{A,i}$ und $\vec{r}_{B,i}$ geben die Positionen der assoziierten Atompaare an. Üblicherweise wird der RMSD-Wert in Ångström angegeben und ist gleich 0 für eine perfekte Überlagerung.

Je nach Anwendung gibt es verschiedene Verfahren zur Berechnung des RMSD. Werden zwei identische Strukturen in unterschiedlichen Konformationen verglichen, so können alle Atome berücksichtigt werden und deren Abweichungen in den RMSD-Wert eingehen. Dies gilt z.B. für verschiedene Ligandpositionen oder Proteine mit identischer Sequenz. Bei der RMSD-Wert-Berechnung für Strukturen mit unterschiedlicher Sequenz kann keine vollständige Zuordnung der Atome erfolgen. Stattdessen wird der RMSD-Wert in solchen Fällen üblicherweise nur für die C_α-Atome oder alle Rückgratatome äquivalenter Positionen bestimmt. In der Literatur existieren mehrere Lösungen für das Zuordnungsproblem (z.B. TM-Align [113], CE-Align [127], Dali [128]), welche zu unterschiedlichen RMSD-Werten führen können. Daher muss beim Vergleich von RMSD-Werten immer auch beachtet werden, welches Zuordnungsverfahren verwendet wurde.

Der cRMSD-Wert (engl. *constraint RMSD*) für eine Ligandposition gibt an, wie gut die tatsächlichen Abstände d_i^L zwischen Ligand- und Proteinatomen mit den vorgegebenen optimalen Distanzen d_i^{opt} im Mittel übereinstimmen. Der cRMSD für eine Ligandposition L folgt dann in Analogie zum RMSD (Gl. (3.30)) gemäß:

$$\mathrm{cRMSD}\,(L) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left|d_i^L - d_i^{\mathrm{opt}}\right|^2} \tag{3.31}$$

Für eine perfekte Übereinstimmung mit den Vorgaben ist der cRMSD gleich 0 und steigt dann mit zunehmender Abweichung vom Optimum an. Wie der RMSD wird auch der cRMSD in Ångström angegeben.

3.9 Hardware & zusätzliche Software

Alle Rechnungen für Funktionsübertragungen mit TransCent und zur Ligandpositionssuche mit TransLig wurden auf einem Teil des Athene-Linux-Clusters der Universität Regensburg durchgeführt. Die verwendeten neun Knoten verfügen über jeweils 32 GB Arbeitsspeicher und je zwei AMD K10 Quad-Core-Prozessoren (Opteron 2354, 2,2 GHz).

Die Implementation der TransCent-Version, welche in dieser Arbeit vorgestellt wird, baut auf der ersten Version von [85] auf und verwendet daher ebenfalls die MBT-Bibliothek (*Molecular Biology Toolkit*, [129]) zur Ver- und Bearbeitung von Strukturinformationen und

die BioJava-Bibliothek (Version 1.4, [130]) zum Prozessieren von Sequenzdaten. Darüber hinaus benötigt die neue Version von TransCent die Java-Bibliothek *ArbitraryAxisRotation* für die Transformation der Fingerprint-Potentiale [131].

Die Abbildungen von Proteinstukturen und einzelnen Aminosäuren in dieser Arbeit wurde mit Hilfe des Moleküldarstellungsprogrammes PyMOL [132], Version 1.3.1) erstellt. Die weiteren Abbildungen und Tabellen wurden mit Microsoft Office Excel 2007, Inkscape (Version 0.48.2) bzw. Origin (Version 6.1) erzeugt oder direkt in LaTeXgesetzt.

3.10 Optimierung und Evaluation

3.10.1 Rekapitulationsdesigns

Ob eine Funktionsübertragung zwischen zwei Enzymen erfolgreich ist, kann abschließend nur durch die Umsetzung des Designs im Labor und die Messung der entsprechenden Enzymaktivität entschieden werden. Dieser Prozess ist jedoch aufwändig, langwierig und kostspielig und somit nur für Einzelfälle anwendbar. Um die Gewichte einer Energiefunktion einzustellen (siehe 3.1) oder anschließend die Performanz eines Designprogramms zu bestimmen, sind jedoch ein oder wenige Beispiele nicht ausreichend.

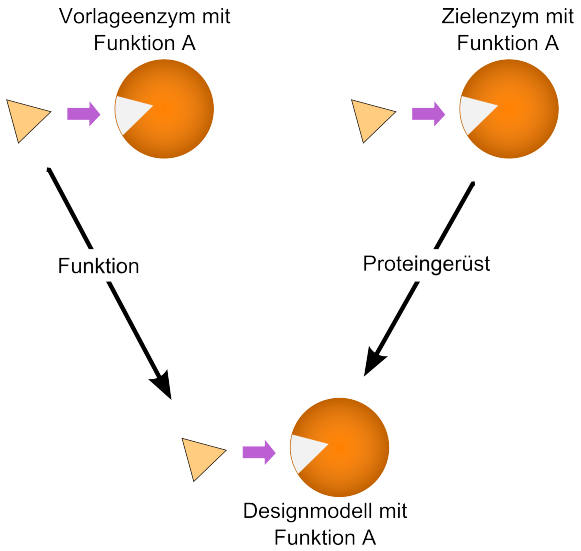


Abbildung 3.13: Schematische Darstellung einer Rekapitulationsrechnung
Das Verfahren beim Rekapitulationsdesign ist dasselbe wie beim Design einer Funktionsübertragung (vgl. Abb. 2.2). Die Funktion des Vorlageenzyms wird auf das Proteingerüst des Zielenzyms übertragen. Der Unterschied besteht darin, dass dasselbe Enzym sowohl als Vorlage- als auch als Zielenzym verwendet wird. Der Vorteil bei dieser Vorgehensweise ist, dass im Gegensatz zu Funktionsübertragungen das erwartete Ergebnis bekannt ist.

Eine *in silico* Alternative stellen sogenannte Rekapitulationsrechnungen dar, wobei die Funktion eines Enzyms „auf sich selbst“ übertragen wird. Es dient also dieselbe Struktur als Vorlage wie auch als Zielgerüst (siehe Abb. 3.13). Da natürlich vorkommende Enzyme über einen langen Zeitraum durch evolutionäre Prozesse optimiert worden sind, kann deren

Aminosäuresequenz sowohl für die Struktur als auch für die Funktion als optimal angesehen werden [88]. Also ist bei diesem Szenario das erwünschte Ergebnis der „Funktionsübertragung“ bekannt und ein computergeneriertes Design kann hinsichtlich seiner Qualität *in silico* bewertet werden. Dabei stehen verschiedene Maße zur Wahl (siehe 3.10.2).

3.10.2 Ähnlichkeitsmaße

Der Vergleich von Proteinsequenzen ist ein wichtiges Element vieler bioinformatischer Methoden, für gewöhnlich mit dem Ziel, die Ähnlichkeit der Sequenzen zu bestimmen, da damit z.B. auf einen gemeinsamen Faltungstyp geschlossen werden kann [133]. Bekannte Algorithmen sind der Needleman-Wunsch- [134] und der Smith-Waterman-Algorithmus [135] für globale und lokale Alignments von Sequenzen unterschiedlicher Länge.

Im vorliegenden Fall ist das Erstellen eines Alignments nicht nötig, da die Positionen eindeutig zugeordnet sind, und der Sequenzvergleich reduziert sich auf einen positionsweisen Vergleich der Aminosäuren aus der wildtypischen und der berechneten Sequenz. Im Folgenden werden drei mögliche Ähnlichkeitsmaße vorgestellt.

3.10.2.1 Sequenzidentität

Das einfachste Maß zur Berechnung der Ähnlichkeit zweier Sequenzen ist gleich dem Anteil von Positionen mit identischen Aminosäuren:

$$\mu_{\mathrm{SID}}\left(A,B\right)=\frac{1}{N_{\mathrm{des}}}\sum_{i=1}^{N_{\mathrm{des}}}\delta\left(a_i,b_i\right) \tag{3.32}$$

Die Sequenzen A und B werden an den N_{des} mutierbaren Positionen (siehe 4.2.1) verglichen, das Kronecker-Delta $\delta\left(a_i,b_i\right)$ gibt den Wert 1 falls die Aminosäuren gleich sind, sonst 0. Alle nicht mutierbaren Positionen werden ignoriert, da diese immer identisch sind und somit nichts zur Ähnlichkeitsbewertung beitragen. Damit liegt der Wert von μ_{SID} zwischen 0 und 1.

3.10.2.2 Blosum-Score

Aminosäuren unterscheiden sich aufgrund ihrer physikalischen und chemischen Eigenschaften. Dabei gibt es jedoch Aminosäuren, die einander ähnlicher sind bzw. sich stärker unterscheiden. Dementsprechend wirken sich Mutationen zwischen unähnlichen Aminosäuren (z.B. Glutamat (geladen) zu Isoleucin (ungeladen)) im Allgemeinen stärker auf ein Enzym aus als zwischen ähnlichen Aminosäuren (z.B. Leucin zu Isoleucin). Diesem Unterschied kann Rechnung getragen werden, indem man bei μ_{SID} das Kronecker-Delta durch den BLOSUM_{62}-Score ersetzt:

$$\mu_{\mathrm{BLOSUM}}\left(A,B\right)=\frac{1}{N_{\mathrm{des}}}\sum_{i=1}^{N_{\mathrm{des}}}\mathrm{BLOSUM}_{62}\left(a_i,b_i\right) \tag{3.33}$$

Der Score $\mathrm{BLOSUM}_{62}\left(\mathrm{a_i},\mathrm{b_i}\right)$ wird der BLOSUM_{62}-Substitutionsmatrix [136] entnommen und gibt die Ähnlichkeit der Aminosäuren a_i und b_i an, wobei negative Scores für sehr unähnliche und positive Scores dementsprechend für ähnliche Aminosäurepaare stehen.

Bei der Berechnung von μ_{BLOSUM} werden wiederum nur die mutierbaren Positionen berücksichtigt. Der allgemeine Wertebereich liegt zwischen -4 und 11, wobei positivere Werte für ähnlichere Sequenzen und somit (potentiell) erfolgreichere Designs sprechen.

3.10.2.3 Positionsspezifische Scoring-Matrizen

Enzyme sind zwar hochspezialisierte Proteine, deren Aminosäuresequenz im Laufe der Evolution optimiert wurde, allerdings zeigt der Vergleich homologer Sequenzen aus unterschiedlichen Spezies, dass es mehrere Möglichkeiten gibt, dieselbe Enzymfunktion zu realisieren. In den aktiven Zentren sind hauptsächlich die katalytischen Residuen und solche, die direkt an der Ligandenbindung beteiligt sind, strikt konserviert. An anderen Positionen ist die Wahlfreiheit zwar häufig eingeschränkt, es herrscht aber trotzdem eine gewisse Variabilität hinsichtlich der „erlaubten“ Aminosäuren.

Um dies bei der Bewertung von Rekapitulationsdesigns zu berücksichtigen kann diese Variabilität aus einem MSA von zum Zielprotein homologen Sequenzen abgeleitet werden. Für jede Position wird dazu aus der entsprechenden Spalte die Aminosäureverteilung ermittelt und mit den Hintergrundwahrscheinlichkeiten unter Verwendung von logarithmierten Chancenquotienten (engl. *log-odds ratios*) zu einer positionsspezifischen Scoringmatrix (PSSM) $\mathcal{M}$ verrechnet.

$$\mathcal{M}_{a_i,i} = log\left(\frac{f_{\text{MSA}}\left(a_i\right)}{f_{\text{Hintergrund}}\left(a_i\right)}\right) \tag{3.34}$$

Als Hintergrundwahrscheinlichkeiten $f_{\text{Hintergrund}}(a_i)$ werden die relativen Häufigkeiten der 20 Aminosäuren in der Swiss-Prot Datenbank [137] verwendet (Stand 14.11.2011, siehe Tabelle 3.3), während die spaltenweisen Häufigkeiten $f_{\text{MSA}}(a_i)$ der Aminosäuren aus dem MSA ermittelt werden.

Um auch Sequenzen unterschiedlicher Länge bewerten zu können, wird jede Spalte der Scoringmatrix um einen Wert für Lücken erweitert. Die Vorgehensweise ist analog zu oben, allerdings wird der Wert für die Lücken-Hintergrundwahrscheinlichkeit ebenfalls aus dem MSA ermittelt und zwar als relative Häufigkeit für das Vorkommen von Lücken über alle Spalten hinweg. Für die Berechnung des Gesamt-Scores μ_{PSSM} einer Sequenz A werden auch hier wiederum nur die designbaren Positionen herangezogen:

$$\mu_{\text{PSSM}}\left(A\right) = \frac{1}{N_{\text{des}}} \sum_{i=1}^{N_{\text{des}}} \mathcal{M}_{a_i,i} \tag{3.35}$$

Das Ähnlichkeitsmaß gibt also Auskunft darüber, wie sehr eine (Design-)Sequenz natürlich vorkommenden Enzymen ähnelt und beschränkt sich nicht nur auf den Vergleich mit einer Wildtypsequenz. Wie beim Blosum-Score werden bei zunehmender Ähnlichkeit die Werte positiver. Das Auftreten korrelierter Mutationen [138] oder anderer Abhängigkeiten in der Besetzung mehrerer Positionen wird bei diesem Verfahren allerdings nicht berücksichtigt, da die Verteilung der Aminosäuren an den einzelnen Positionen als unabhängig betrachtet wird.

3.10.3 Enzymdatensatz für Optimierung und Evaluation

Die Evaluation von Ergebnissen anhand bekannter Daten ist eine in der Bioinformatik übliche Vorgehensweise (vgl. z.B. [139]) und stellt oft die einzige Möglichkeit dar, eine

Aminosäure	**rel. Häufigkeit**
Alanin	0,0826
Arginin	0,0553
Asparagin	0,0406
Aspartat	0,0546
Cystein	0,0136
Glutamat	0,0675
Glutamin	0,0393
Glycin	0,0708
Histidin	0,0227
Isoleucin	0,0597
Leucin	0,0966
Lysin	0,0585
Methionin	0,0242
Phenylalanin	0,0386
Prolin	0,0470
Serin	0,0655
Threonin	0,0534
Tryptophan	0,0108
Tyrosin	0,0292
Valin	0,0687

Tabelle 3.3: Aminosäure-Hintergrundwahrscheinlichkeiten
Relative Häufigkeiten der 20 kanonischen Aminosäuren in der Swiss-Prot Datenbank ([137], Stand 14.11.2011)

statistisch signifikante Aussage über die Qualität von Vorhersageergebnissen zu erzielen. Um die Performanz von TransCent bewerten und entsprechend optimieren zu können, ist daher ein Satz von Enzymen erforderlich, für welchen Designrechnungen durchgeführt, begutachtet und bewertet werden können. Ein Enzymdatensatz besteht dabei aus einer Struktur des Enzym-Ligand-Komplexes im PDB-Format zusammen mit einem MSA homologer Sequenzen im FASTA-Format. Bei Funktionsübertragungen kommt die Struktur des Zielproteins im PDB-Format hinzu.

Für eine statistisch signifikante Aussage über die Performanz von TransCent muss eine hinreichende Menge von Enzymen verwendet werden, welche in Trainings-, Test- und Benchmarkdatensatz unterteilt werden muss. Unterscheiden sich diese Enzyme in ihrer Funktion und Struktur relativ stark, d.h. sie stammen aus verschiedenen Familien der SCOP-Datenbank [140] und katalysieren Reaktionen aus unterschiedlichen EC-Klassen [12], so kann zusätzlich sichergestellt werden, dass der Designalgorithmus gut generalisiert.

Das Vorgehen beim Zusammenstellen und Aufbereiten der Strukturdatensätze wird im Folgenden detailliert beschrieben.

3.10.3.1 Auswahl der Enzyme

Der Datensatz für die Rekapitulationsrechnungen (vgl. 3.10.1) stammt aus drei unterschiedlichen Quellen. Zum einen wurden die 27 Enzyme des TransCent-Testdatensatzes aus [76] übernommen, welche bei der Evaluation der ersten TransCent-Version zum Einsatz kamen (siehe Tabelle 3.4).

Diese wurden erweitert um Vertreter der zehn Enzymgruppen, welche in einer Studie von Weng et al. [141] als Enzyme mit den am wenigsten „flexiblen“ aktiven Zentren identifiziert wurden (siehe Tabelle 3.4). Flexibilität bedeutet in diesem Zusammenhang, dass sich die aktiven Zentren homologer Enzyme strukturell relativ stark unterscheiden können. Im Gegensatz dazu sind bei den „starren“ Enzymen die Anforderungen an die aktiven Zentren derart, dass nur wenige Freiheitsgrade bei der strukturellen Umsetzung existieren. Für Rekapitulationsrechnungen sind diese Enzyme daher besonders geeignet, da das optimale Designergebnis wohldefiniert ist.

Quelle	**PDB-Codes**
Fischer et al.	1ajs(A), 1b8o(A), 1dbt(A), 1dqx(A), 1f74(A), 1f8e(A), 1g6s(A), 1h4g(A), 1jub(A), 1km4(A), 1kqp(A), 1lbm(A), 1m40(A), 1o08(A), 1o8b(B), 1obo(A), 1po5(A), 1qop(A), 1u7g(A), 1ucd(A), 1ujp(A), 1v2x(A), 1vyr(A), 1wui(L), 1y0y(A), 2aeb(A), 2bkx(A)
Weng et al.	1aop(A), 1dj1(A), 1dod(A), 1dve(A), 1fcb(A), 1idt(A), 1n2c(A), 2cpo(A), 3nos(A), 7atj(A)

Tabelle 3.4: PDB-Codes der Datensätze aus Fischer et al. und Weng et al. Zur eindeutigen Identifizierung der Enzymstruktur ist jeweils in Klammern der Name der Kette angegeben.

Der dritte und größte Teil des Datensatzes stammt aus dem „*general set*“ der PDBBind-Datenbank (Version 2010, [142, 143]). Die Datenbank umfasst alle Komplexstrukturen aus der Protein Data Bank [144, 145] (Stand 1. Januar 2010), für die ein experimentell bestimmter Wert für die Bindungsaffinität (K_i, K_d oder IC_{50}) vorliegt. Bei diesen 6772 Strukturen handelt es sich um Protein-Protein-, Protein-DNA-, DNA-Ligand- und Protein-Ligand-Komplexe, wobei nur letztere für den Rekapitulationsdatensatz in Frage kommen. Zusätzlich enthält der PDBBind-Datensatz weitere Strukturen, die aus verschiedenen Gründen ungeeignet für Rekapitulationen sind, weswegen die Auswahl anhand verschiedener Filterkriterien verfeinert wurde:

1. Es werden nur Strukturen akzeptiert, die durch Röntgenkristallographie aufgeklärt wurden.
2. Die Auflösung der Struktur darf maximal 2,0 Å betragen und der R-Faktor den Wert 0,2 nicht übersteigen.
3. Die Anzahl der Residuen des Proteins ist auf 550 beschränkt.
4. Der an der Katalyse beteiligte Ligand muss zwischen 10 und 50 Schweratome umfassen.
5. Einer Struktur muss eine EC-Nummer zugeordnet werden können.

6. Im HSSP-MSA [146], welches zur Struktur gehört, müssen mindestens 80 homologe Sequenzen vorhanden sein.

Die Informationen für die Filterung wurden, wo notwendig, direkt von der PDB bezogen. Durch obige Kriterien wird sichergestellt, dass es sich um Enzymstrukturen hoher Qualität handelt und ausreichend viele homologe Sequenzen bekannt sind, welche für das Erstellen der Strukturbibliothek (vgl. 3.4.1) verwendet werden können. Die so verbliebenen 754 Strukturen wurden anschließend mit Hilfe des PISCES-Servers [147] auf Redundanzfreiheit gefiltert, wobei eine maximale paarweise Sequenzidentität von 25% erlaubt und der Vergleich „kettenweise“ durchgeführt wurde (Standardoptionen mit *cull by chains*), d.h. falls eine PDB-Struktur mehrere Ketten enthielt, so wurden diese unabhängig voneinander behandelt.

Dadurch wurde die auf der PDBBind-Datenbank basierende Auswahl auf 103 Enzyme eingegrenzt, welche zusammen mit den 37 Enzymen aus den anderen beiden Quellen erneut wie oben mit dem PISCES-Server gefiltert wurden. Die so erhaltenen 109 Strukturen wurden dann in Zusammenarbeit mit Dr. Marco Bocola einzeln begutachtet und nach folgenden Kriterien auf ihre Eignung für Rekapitulationsrechnungen überprüft:

1. Bei dem Liganden muss es sich um ein dem Substrat oder Produkt der Enzymreaktion ähnliches Molekül handeln.
2. Der Ligand darf nicht kovalent im aktiven Zentrum gebunden sein.
3. Die Struktur muss im Bereich des aktiven Zentrums vollständig aufgelöst sein, d.h. es dürfen keine Residuen in und um das aktive Zentrum fehlen.
4. Die Aminosäuren an den Positionen im aktiven Zentrum müssen wildtypisch sein, d.h. Strukturen mit einer oder mehreren Mutationen in diesem Bereich werden ausgeschlossen.
5. Strukturen, bei denen ausschließlich Cofaktoren gebunden sind, werden verworfen.
6. Bei Strukturen von Dimeren oder höheren Oligomeren werden solche ausgeschlossen, bei denen Residuen verschiedener Protomere am Aufbau eines aktiven Zentrums beteiligt sind.

Als Informationsgrundlage dienten hierbei die entsprechenden Einträge in der PDB-Datenbank und der darauf aufbauenden PDBSum-Datenbank [148, 149]. Bei Strukturen, welche mindestens eines der obigen Kriterien nicht erfüllten, wurde versucht, passende Alternativstrukturen äquivalenter Enzyme zu identifizieren. Diese wurden dann ggf. als Ersatz in den Datensatz aufgenommen. Dabei wurden die Schwellwerte für die Auflösung ($\leq$ 2,2 Å) und den R-Faktor ($\leq$ 0,23) etwas angehoben und die Beschränkung hinsichtlich der Anzahl von Schweratomen des Liganden fallen gelassen. Insgesamt erfüllen 73 Enzymstrukturen die geforderten Kriterien.

3.10.3.2 Aufbereitung der Strukturen

Um die Rekapitulationsrechnungen für die ausgewählten Enzyme durchführen zu können, mussten zunächst die Strukturdatensätze aufbereitet werden. Dazu waren zwei Schritte notwendig. Zum einen wurden mit dem Programm S2MSAAA (vgl. 3.4.1.1) gefilterte MSAs homologer Sequenzen für die einzelnen Enzyme erstellt (paarweise Sequenzidentität

zwischen 20% und 90%, maximale Sequenzlängenabweichung 30%). Da nur für 53 der 73 Enzyme hinreichend viele Sequenzen gefunden werden konnten, mussten weitere 20 Strukturen aus dem Datensatz entfernt werden.

Für die verbliebenen Enzyme, deren PDB-Codes in Tabelle 3.5 aufgeführt sind, wurden im zweiten Schritt die Strukturdatensätze von der PDB heruntergeladen und folgendermaßen überarbeitet:

1. Überzählige Ketten, nicht relevante Heteroatome, Wassermoleküle und Wasserstoffatome wurden aus den Datensätzen gelöscht.
2. Selenomethionin-Residuen wurden zu Methionin-Residuen „mutiert“, d.h. Selenatome wurden in Schwefelatome umbenannt, da TransCent auf die 20 kanonischen Aminosäuren beschränkt ist.
3. Waren alternative Konformationen für Seitenketten in der Struktur angegeben, so wurde jeweils nur die Konformation mit der größten Besetzungszahl beibehalten (i.A. Konformation A), vorausgesetzt diese führte nicht zu einer sterischen Kollision zwischen Seitenkette und Ligand.
4. Mit Hilfe des Programms Profix aus dem Softwarepaket Jackal [150] wurden zum einen fehlende Atome bzw. fehlende Residuen ergänzt und zum anderen Punktmutationen in den Strukturen auf die wildtypische Aminosäure zurückgesetzt. Fehlende Residuen an den Termini der Proteinketten wurden nicht modelliert.
5. Die Atome der Ligandmoleküle wurden ggf. umbenannt, um mit der DSX-Notation (siehe 3.3.1) in Einklang zu sein.
6. Abschließend wurden die Residuen bei eins beginnend durchgehend nummeriert.

Obgleich Strukturen mit nicht aufgelösten Residuen im Bereich des aktiven Zentrums generell vom Datensatz ausgeschlossen waren, befanden sich dennoch unter den ausgewählten 53 Enzymstrukturen einige mit fehlenden Loop-Residuen oder nicht aufgelösten Seitenketten. Diese mussten ergänzt werden, um eine näherungsweise korrekte Berechnung der Proteinenergien sicherzustellen. Zu diesem Zweck wurden drei Verfahren zur Modellierung von Proteinstrukturen getestet, MODELLER (siehe 3.4.1.2), I-Tasser [151] und Profix aus dem Softwarepaket Jackal [150].

Letzteres hat dabei den Vorteil, dass es die fehlenden Atome ergänzt und den Rest der Struktur unverändert lässt. Nur an den Übergängen zwischen dem existierendem Rückgratverlauf und den zu modellierenden Fragmenten werden, wenn nötig, kleine Änderungen eingeführt. Die beiden anderen Programme führen eine komplette Homologiemodellierung durch, was dazu führt, dass auch die aufgelösten Teile der Proteinstruktur mitunter kleine Änderungen erfahren und von ihrer ursprünglichen Position abweichen. Diese Veränderungen wirken sich jedoch potentiell auf das Ergebnis des Rekapitulationsdesigns aus, insbesondere wenn sie das aktive Zentrum betreffen, und verzerren somit die Vorhersagekraft von TransCent. Daher fiel die Wahl auf Profix als Methode für die Modellierung fehlender Atome.

3.10.3.3 TransCent-Rekapitulationsdatensatz

Der fertige TransCent-Rekapitulationsdatensatz umfasst 53 Enzyme aus 32 unterschiedlichen Spezies, wobei *Escherichia coli* mit 10 Enzymen am häufigsten vertreten ist. Die

Länge der Proteinketten liegt zwischen 124 und 547 Residuen mit einer durchschnittlichen Proteinlänge von 316 Aminosäuren. Die Struktur der humanen Chitinase (1hkk) enthält zwei Moleküle des 43 Schweratome umfassenden Liganden Allosamidin, welche beide als designrelevant angesehen werden und daher auch beide in der überarbeiteten Struktur enthalten sind. Insgesamt enthält der Datensatz Ligandmoleküle bestehend aus 9 bis 86 Schweratomen mit einer mittleren Größe von 27 Atomen. Die MSAs für die Berechnung der Strukturbibliothek enthalten zwischen 88 und 636 Sequenzen homologer Enzyme. Eine detaillierte Auflistung der einzelnen Werte und Angaben findet sich in Tabelle 3.5. Bemerkenswerterweise sind Enzyme aus allen sechs EC-Klassen im Datensatz enthalten. Somit ist sichergestellt, dass TransCent für ein breites Spektrum an Enzymreaktionen optimiert und bewertet wird.

PDBID	Kette	N_R	N_S	N_L	EC	Å	R	Organismus
16pk	A	415	513	39	2.7.2.3	1,60	0,19	*Trypanosoma brucei*
1a4m	A	349	293	19	3.5.4.4	1,95	0,19	*Mus musculus*
1ajs	A	412	536	25	2.6.1.1	1,60	0,17	*Sus scrofa*
1b8o	A	280	613	19	2.4.2.1	1,50	0,18	*Bos taurus*
1bxo	A	323	143	44	3.4.23.20	0,95	0,10	*Penicillium janthinellum*
1c1h	A	306	119	43	4.99.1.1	1,90	0,18	*Bacillus subtilis*
1c21	A	262	473	9	3.4.11.18	1,80	0,16	*Escherichia coli*
1dg9	A	157	207	15	3.1.3.48	1,90	0,18	*Bos taurus*
1dj1	A	291	289	43	1.11.1.5	1,93	0,19	*Saccharomyces cerevisiae*
1e6q	M	499	591	14	3.2.3.1	1,35	0,12	*Sinapis alba*
1ec9	A	443	255	13	4.2.1.40	2,00	0,18	*Escherichia coli*
1f74	A	293	123	20	4.1.3.3	1,60	0,17	*Haemophilus influenzae*
1fsg	A	230	129	33	2.4.2.8	1,05	0,12	*Toxoplasma gondii*
1g6s	A	427	334	26	2.5.1.19	1,50	0,15	*Escherichia coli*
1gar	A	209	567	48	2.1.2.2	1,96	0,17	*Escherichia coli*
1gvk	B	240	337	26	3.4.21.36	0,94	0,12	*Sus scrofa*
1gx4	A	287	131	51	2.4.1.151	1,46	0,15	*Bos taurus*
1h46	X	430	242	19	3.2.1.91	1,52	0,17	*Phanerochaete chrysosporium*
1h4g	A	205	211	18	3.2.1.8	1,10	0,16	*Bacillus agaradhaerens*
1hkk	A	364	201	86	3.2.1.14	1,85	0,18	*Homo sapiens*
1jak	A	499	236	14	3.2.1.52	1,75	0,18	*Streptomyces plicatus*
1km5	A	212	96	21	4.1.1.23	1,50	0,17	*Methanothermobacter thermautotrophicus*
1n2b	A	286	615	41	2.7.4.6	2,20	0,23	*Virgibacillus halodenitrificans*
1o08	A	221	278	20	5.4.2.6	1,20	0,14	*Lactococcus lactis*
1o0h	A	124	194	27	3.1.27.5	1,20	0,19	*Bos taurus*
1ocn	A	360	104	42	3.2.1.91	1,31	0,13	*Humicola insolens*
1pfu	A	547	292	9	6.1.1.10	1,91	0,20	*Escherichia coli*
1q0n	A	158	558	45	2.7.6.3	1,25	0,10	*Escherichia coli*
1q6c	A	490	131	46	3.2.1.2	1,86	0,17	*Glycine max*
1qop	A	267	411	17	4.2.1.20	1,40	0,15	*Salmonella enterica*
1r5y	A	372	425	13	2.4.2.29	1,20	0,17	*Zymomonas mobilis*
1u2y	A	495	279	13	3.2.1.1	1,95	0,17	*Homo sapiens*
1ucd	A	190	303	21	3.1.27.1	1,30	0,20	*Momordica charantia*
1uj6	A	225	592	14	5.3.1.6	1,74	0,20	*Thermus thermophilus*
1uod	A	347	390	10	2.7.1.29	1,90	0,17	*Escherichia coli*
1v0l	A	302	226	17	3.2.1.8	0,98	0,12	*Streptomyces lividans*
1vyr	A	363	515	47	1.7.99.-	0,90	0,12	*Enterobacter cloacae*
1yjq	A	293	168	48	1.1.1.169	2,09	0,16	*Escherichia coli*
2aeb	A	314	418	13	3.5.3.1	1,29	0,17	*Homo sapiens*
2afw	A	329	174	11	2.3.2.5	1,56	0,19	*Homo sapiens*
2bkx	A	242	474	16	3.5.99.6	1,40	0,12	*Bacillus subtilis*
2c4w	A	159	636	24	4.2.1.10	1,55	0,16	*Helicobacter pylori*
2ctc	A	307	90	12	3.4.17.1	1,40	0,16	*Bos taurus*
2cyh	A	164	457	13	5.2.1.8	1,64	0,19	*Homo sapiens*
2jf4	A	511	158	23	3.2.1.28	2,20	0,17	*Escherichia coli*
2pu1	A	429	387	10	4.2.1.11	1,80	0,17	*Trypanosoma brucei*
2tmn	E	316	88	13	3.4.24.27	1,60	0,18	*Bacillus thermoproteolyticus*
2zcr	A	284	221	28	2.5.1.-	1,92	0,18	*Staphylococcus aureus*
3b7p	A	281	523	34	2.5.1.16	2,00	0,22	*Plasmodium falciparum*
3e3u	A	196	617	26	3.5.1.88	1,56	0,17	*Mycobacterium tuberculosis*
3g8c	A	444	396	47	6.3.4.14	2,00	0,18	*Escherichia coli*
7atj	A	305	540	59	1.11.1.7	1,47	0,16	*Armoracia rusticana*
8a3h	A	300	93	25	3.2.1.4	0,97	0,10	*Bacillus agaradhaerens*

Tabelle 3.5: Auflistung der 53 Enzyme des TransCent-Rekapitulationsdatensatzes
Angegeben sind: die PDBID der Struktur, der Name der verwendeten Kette, die Anzahl der Residuen ($\mathbf{N_R}$), die Anzahl der homologen Sequenzen zur Erstellung der Strukturbibliothek ($\mathbf{N_S}$), die Anzahl der Ligandatome ($\mathbf{N_L}$), die zugewiesene EC-Nummer [12] (**EC**), die Auflösung (**Å**) und der R-Wert (**R**) der Struktur und der Organismus aus dem das Enzym stammt. Mit Ausname von $\mathbf{N_S}$ wurde sämtliche Information aus dem zugehörigen PDB-Eintrag entnommen. Verfügt ein Eintrag über mehr als eine EC-Nummer, so ist jeweils nur die erste angegeben. Die Liste ist alphabetisch nach PDBID geordnet.

4 Ergebnisse

Im Rahmen dieser Arbeit wurde das Enzymdesignprogramm TransCent zur Übertragung enzymatischer Funktionen zwischen zwei Proteinstrukturen überarbeitet und um die Fähigkeit zur flexiblen Ligandpositionierung erweitert. Zwei der vier Module wurden durch neuere Versionen ersetzt und ein drittes punktuell überarbeitet. Die dadurch notwendige Neueinstellung der Modulgewichte wurde in Form einer Gridsuche auf einem erweiterten Testdatensatz durchgeführt. Zur Lösung des Problems der Ligandpositionierung wurde das Programm TransLig entwickelt und entsprechende Änderungen am TransCent-Algorithmus vorgenommen.

In diesem Kapitel werden zunächst die überarbeiteten Module vorgestellt und deren Beitrag zur Performanz von TransCent untersucht. Anhand der Ergebnisse der Gridsuche wird eine ausführliche *in silico* Analyse der Rekapitulationsrechnungen durchgeführt. Die Evaluation von TransLig erfolgt sowohl basierend auf den Resultaten der Ligandpositionssuche als auch für Rekapitulationsdesigns in Kombination mit TransCent. Abschließend werden die Ergebnisse einiger Funktionsübertragungsdesigns zusammen mit den Resultaten der experimentellen Überprüfung präsentiert.

4.1 Überarbeitung der TransCent-Module

TransCent ist aufgebaut aus den Modulen für Proteinstabilität, Ligandenbindung, pK_a-Wert-Optimierung und dem Fingerprint-Modul. Mit Ausnahme des PROPKA-Moduls wurden im Rahmen dieser Arbeit alle Module ausgetauscht bzw. überarbeitet. Die Änderungen sind im Folgenden kurz beschrieben.

4.1.1 Proteinstabilität

Für die Proteinstabilität ist in TransCent das Rosetta-Modul zuständig. Dieses greift auf die Modellierungseinheit und Energiefunktion des Designalgorithmus für ein starres Proteinrückgrat (*fixbb*, siehe 3.2) der Rosetta Software Suite zurück. Rosetta wird von mehreren Arbeitsgruppen weltweit ständig weiterentwickelt und hat seit der letzten Version von TransCent [85] einen Versionssprung von Rosetta++ auf Rosetta3 vollzogen. Um auf den erweiterten Funktionsumfang und die verbesserte Energiefunktion zurückgreifen zu können, wurde wie schon bei Rosetta++ eine neue Schnittstelle (siehe 3.2.3) implementiert. Bei der Wahl der Energiefunktion fiel die Entscheidung zugunsten von *score12* (siehe 3.2.1), welche bereits erfolgreich beim Design von Proteinen eingesetzt wurde (siehe z.B. [92, 152]).

4.1.2 Ligandenbindung

Mit DSX (DrugScore eXtended, siehe 3.3.1) wurde 2011 die Nachfolgeversion des Programms DrugScore veröffentlicht, auf dem das Modul für Ligandenbindung von TransCent

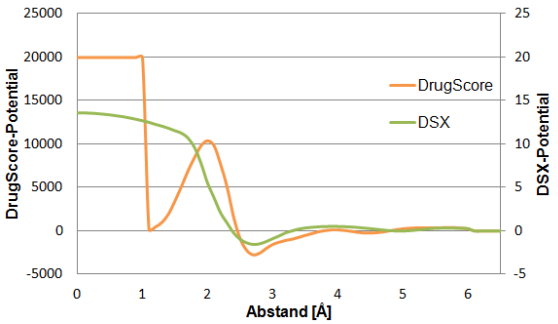


Abbildung 4.1: Vergleich zweier Potentialverläufe von DrugScore und DSX
Die beiden Kurven zeigen das Potential für die Wechselwirkung zweier Sauerstoffatome im Abstand von 0 und 6,5 Å. Während bei DSX aufgrund der Gauß-Korrektur kleine Abstände unter 2 Å immer deutlich bestraft werden, existiert bei DrugScore ein fast neutrales Minimum bei ca. 1 Å.

basiert. Gerd Neudert hat auch für DSX eine rotamerbasierte Variante entwickelt, welche in der neuen TransCent-Version zum Einsatz kommt. DSX unterscheidet mehr Atomtypen als DrugScore und hat eine größere Anzahl von Atompaar-Potentialen, so dass die Wechselwirkungen zwischen Ligand und Protein noch präziser beschrieben werden können.

Für die Umstellung auf DSX gibt es einen weiteren Grund. Wie in Abbildung 4.1 anhand von zwei repräsentativen Beispielen gezeigt wird, differieren die Potentialverläufe der beiden Programmversionen im Bereich um das Abstandsoptimum (hier bei ca. 3 Å) nur wenig. Für Abstände kleiner 2 Å treten hingegen gravierende Unterschiede auf. Bei den DrugScore-Potentialen gehen die Werte nach dem Maximum wieder fast bis auf 0 zurück und nur Abstände unter einem Ångström werden maximal bestraft. Für TransCent bedeutet dies, dass Distanzwerte zwischen Ligand- und Proteinatom von einem Ångström und ca. 4 Å qualitativ gleichwertig sind. Aus physikalischer Sicht ist dieses Verhalten jedoch unsinnig und wird bei den DSX-Potentialen durch Einführung einer Gauß-Korrektur für kleine Distanzwerte ausgeglichen.

Da bei der ersten TransCent-Version die Ligandposition fest vorgegeben war, war das Verhalten von DrugScore bei kurzen Atomabständen unkritisch für das Design. Da das Modul für Ligandenbindung allerdings entscheidenden Einfluss bei der Wahl der Ligandposition hat (vgl. 3.6), muss in der neuen Version sehr wohl Rücksicht darauf genommen werden. Durch die Umstellung auf DSX können daher Artefakte bei der Bewertung von Protein-Ligand-Interaktionen vermieden werden.

4.1.3 Fingerprint

Zusätzlich zum Austausch der Module für Proteinstabilität und Ligandenbindung wurde auch das Fingerprint-Modul überarbeitet, um Schwächen bei der Modellierung der Protein-Ligand-Interaktionen zu beseitigen. Zum einen wurde das Referenzmodell zur Berechnung der Aminosäurewahrscheinlichkeiten (siehe 3.4.2.2) verfeinert. Anstelle wie bisher alle Aminosäuren als gleich wahrscheinlich zu betrachten, werden die Aminosäurewahrscheinlichkeiten durch ihre Häufigkeiten in der Swiss-Prot Datenbank geschätzt.

Die zweite Änderung betrifft die Ableitung der Ellipsoid-Definitionen aus den Punktwolken der wechselwirkenden Schweratome (siehe 3.4.2.1). Statt ihrer Projektionen auf die Koordinatenachsen werden nun die „echten“ Halbachsen der Ellipsoide verwendet. Diese entsprechen den Richtungen der drei Hauptkomponenten der Punktwolken, wodurch die Größe und die Orientierung der Wolken präziser erfasst werden kann (siehe Abb. 4.2).

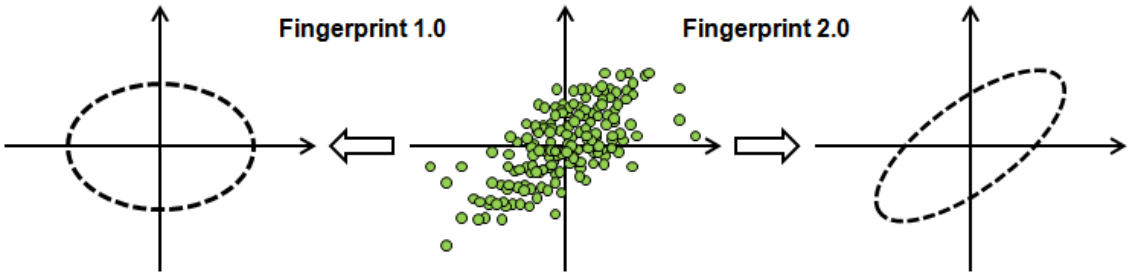


Abbildung 4.2: Vergleich der aus den Punktwolken abgeleiteten Ellipsoid-Definitionen
Die zweidimensionale Darstellung zeigt schematisch den Unterschied der Ellipsoid-Definitionen, welche aus den Punktwolken abgeleitet werden. In der neuen Version wird der Ellipsoid, welcher durch die Kovarianzmatrix der Punktkoordinaten definiert wird direkt übernommen. Auf eine Projektion der Halbachsen auf die Koordinatenachsen wird verzichtet, wodurch die Form der Punktwolken besser erfasst wird.

Der bedeutendste Eingriff erfolgte bei der Berechnung der Fingerprint-Energien der Rotamere für die einzelnen Potentiale. Die Seitenketten einiger Aminosäuren haben potentiell mehrere Möglichkeiten, Wechselwirkungen auszubilden. Bei Arginin kommen z.B. die drei Stickstoffatome der Seitenkette als Wasserstoffbrücken-Donor in Frage [153, 154]. Bislang wurden jeweils für alle Möglichkeiten die Fingerprint-Energien gemäß Gleichung (3.12) berechnet und das entsprechende Rotamer mit dem Mittelwert der Energien bewertet. Dies kann allerdings dazu führen, dass Seitenketten, welche eigentlich ein Potential erfüllen würden, mit einer ungünstigen Energie belegt werden.

Das Beispiel in Abbildung 4.3 zeigt ein Arginin-Rotamer, welches mit einem der beiden Stickstoffe der Guadiniumgruppe innerhalb des Ellipsoids liegt und eine Wechselwirkung mit dem Liganden eingehen kann. Dies wird mit einem günstigen, negativen Energiewert von -1,76 belohnt. Da die beiden anderen potentiellen Interaktionspartner außerhalb des Potentials liegen, werden diese mit der maximalen Strafenergie von 9,21 belegt. Daraus ergibt sich für das Rotamer ein gemittelter Energiewert von 5,55. Weil nicht alle prinzipiell möglichen Wechselwirkungen gleichzeitig ausgebildet werden, erhält das Arginin-Rotamer also einen vergleichsweise schlechten Energiewert, ungeachtet der Tatsache, dass es die gewünschte Wechselwirkung eingehen kann.

Grundsätzlich fordert das Fingerprint-Modul, dass genau eine Wechselwirkung je Potential modelliert wird und nicht möglichst viele. Daher wird in der neuen TransCent-Version ein Rotamer mit der kleinsten und somit besten aller möglichen Energien bewertet. Die Seitenkette aus dem Beispiel erhält somit den günstigen Energiewert von -1,76.

Das PROPKA-Modul zur Optimierung der Protonierungszustände wurde von [85] unverändert übernommen. Insgesamt erfordern die eingeführten Änderungen jedoch eine neuerliche Optimierung der Modulgewichte für die Energiefunktion (vgl. 3.1). Beispielsweise führt

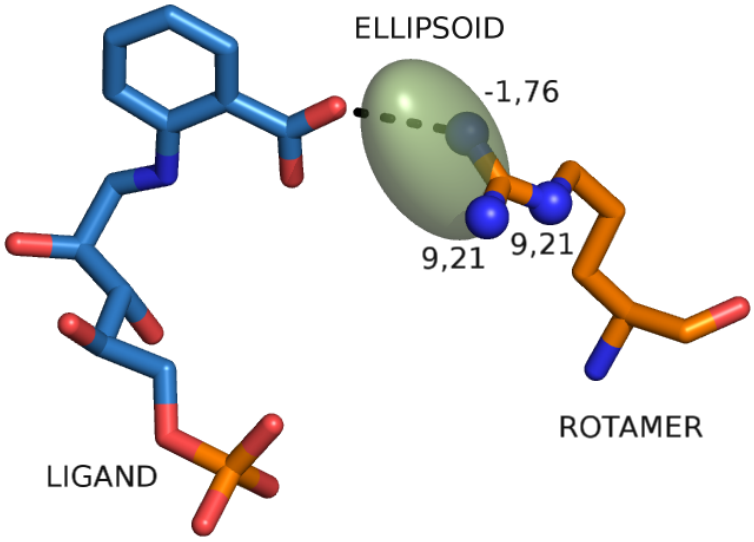


Abbildung 4.3: Energetische Bewertung eines Rotamers für ein Fingerprint-Potential
Die Seitenkette eines Arginin-Residuums hat grundsätzlich drei Möglichkeiten, eine Wechselwirkung mit dem Ligandmolekül einzugehen. Eines der beiden η-Stickstoffatome kann passend im Ellipsoid platziert werden und erhält dafür eine Fingerprint-Energie von -1,76. Die Wechselwirkung mit dem Liganden wird durch gestrichelte schwarze Linie angedeutet. Die anderen Stickstoffatome liegen außerhalb des Ellipsoids und werden mit der maximalen Strafenergie von 9,21 belegt.

die Neuskalierung der DSX-Potentiale dazu, dass der Wertebereich der DSX-Energien um ca. drei Größenordnungen kleiner ist als der von DrugScore (vgl. Abb. 4.1). Die Beiträge der Module müssen also neu aufeinander abgestimmt werden, um eine optimale Performanz zu gewährleisten.

4.2 Optimierung und *in silico* Bewertung

Wie bei der Optimierung von Designprogrammen üblich und auch schon bei der Entwicklung der ersten TransCent-Version geschehen, werden auch in dieser Arbeit Rekapitulationen (siehe 3.10.1) als Grundlage für die Gewichtsoptimierung und die Evaluation der Performanz verwendet. Der Designerfolg bei Rekapitulationsrechnungen ist ein bewährtes Maß zur *in silico* Bewertung von Designalgorithmen, da die erwarteten Designergebnisse *a priori* bekannt sind. Bezogen auf das Funktionsübertragungsdesign bedeutet dies, dass das gleiche Enzym sowohl als Vorlage wie auch als Zielstruktur dient. Es wird also versucht, die wildtypische Aktivität auf dem Enzymgerüst zu etablieren.

Sowohl bei Rekapitulationsrechnungen als auch bei „echten“ Funktionsübertragungen wird für gewöhnlich nie das gesamte Enzym neu designt. Durch die verfügbare Hardware sind dem Enzymdesign noch Grenzen gesetzt, denn das vollständige Design eines Enzyms mit 300 Aminosäuren würde sowohl was die Rechenzeit als auch den Speicherbedarf angeht

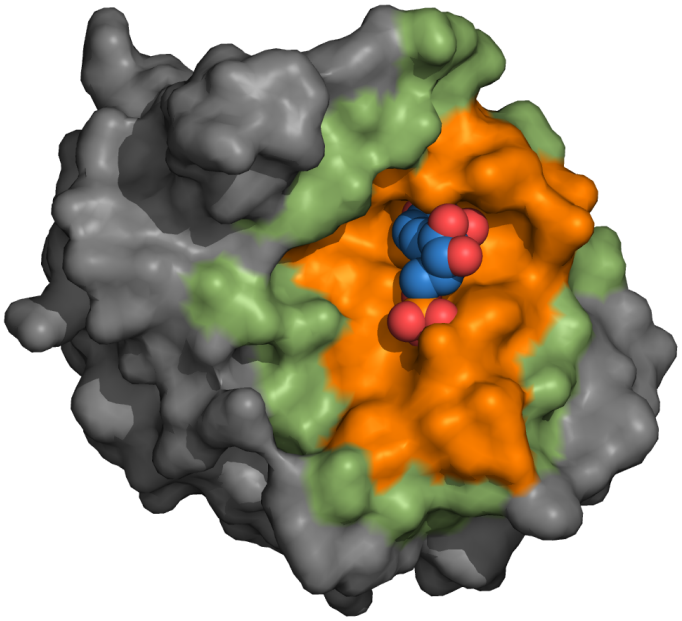


Abbildung 4.4: Definition der designbaren Bereiche
Beim Design werden die Proteinpositionen in drei Klassen unterteilt. Die erste Klasse umfasst die mutierbaren Positionen (oranger Bereich), an denen tatsächlich Mutationen eingeführt werden können. Dieser Bereich beschränkt sich i.A. auf die direkte Umgebung des Ligandmoleküls. Bei den Positionen der zweiten Klasse (grüner Bereich) sind zwar Rotamerwechsel erlaubt, allerdings muss die native Aminosäure beibehalten werden. Die Positionen der dritten Klasse (grauer Bereich) sind fest und werden beim Design weder verändert noch berücksichtigt.

die Möglichkeiten heutiger Rechner übersteigen. Darüber hinaus ist meist nur ein kleiner Teil eines Enzyms direkt an der Katalyse beteiligt. Der weitaus größere Teil unterstützt diese indirekt, indem er die Faltung des Proteins definiert und stabilisiert.

Da als Zielprotein bei einer Funktionsübertragung i.A. ein natürlich vorkommendes stabiles Enzym gewählt wird, ist es sinnvoll dessen Grundgerüst nur so weit abzuändern, wie es die Etablierung der gewünschten Enzymfunktion erfordert. Der Bereich in dem Mutationen erlaubt sind, kann also auf das zu modellierende aktive Zentrum und dessen direkte Nachbarschaft beschränkt werden.

4.2.1 Mutierbare Positionen

In TransCent gibt es zwei Optionen, wie die Auswahl der mutierbaren Positionen erfolgen kann. Zum einen hat der Benutzer die Möglichkeit, über eine Textdatei die Auswahl der Positionen inklusive der erlaubten Aminosäuren je Position festzulegen. TransCent verwendet dabei sowohl die Verarbeitungsroutinen als auch den Syntax der sogenannten *resfiles* von Rosetta [155]. Alternativ kann die Auswahl auch dem Programm überlassen werden. TransCent verwendet dabei Abstandskriterien um das Protein in drei Bereiche einzuteilen (vgl. Abb. 4.4).

Die innerste Schale umfasst jene Positionen, an denen alle 20 kanonischen Aminosäuren erlaubt sind. Dazu zählen alle Residuen, welche sich in einem Radius von 5 Å um den Liganden befinden. Ergänzt werden diese um die Positionen, welche mit einem Lysin (der Aminosäure mit der längsten Seitenkette) besetzt bis auf 4 Å an den Liganden heranreichen können. Dazu werden an allen Proteinpositionen die passenden Lysin-Rotamere aus der Rotamerbibliothek (siehe 3.2.2) modelliert und getestet.

Die Positionen der zweiten Schale werden als „rotierbar“ definiert. Das bedeutet, dass die Art der Aminosäure aus der Zielstruktur erhalten bleibt und der Designalgorithmus lediglich zwischen verschiedenen Rotameren der gleichen Aminosäure wählen kann. Mutationen sind an diesen Positionen nicht gestattet. Dennoch ist dieser Bereich wichtig für das Design, da manche Mutationen in der inneren Schale erst durch Umlagerungen der rotierbaren Positionen ermöglicht werden. Der Bereich umfasst alle Positionen mit einem maximalen Abstand zum Liganden von 15 Å, welche nicht zur ersten Schale gehören.Alle verbliebenen Proteinpositionen werden zur dritten Schale zusammengefasst und als starr angenommen, d.h. sie sind weder mutierbar noch rotierbar und werden beim Design ignoriert.

Bei der Verwendung einer Ligandpositionsbibliothek (vgl. 3.6) werden für jede Ligandposition die drei Schalen bestimmt. Der Designbarkeitsstatus einer Position ergibt sich dann aus der Überlagerung aller Bereiche, wobei für jede Position der Status mit der maximalen Anzahl von Freiheitsgraden übernommen wird.

Bei den angegebenen Werten für die Abstandskriterien handelt es sich um die TransCent-Standardeinstellungen, die beliebig angepasst werden können.

4.2.2 Festlegung der Modul-Gewichte

Über die Wahl der Gewichte für die Energiefunktion (siehe Gl. (3.1)) kann der Einfluss der Module beim Design aufeinander abgestimmt werden. Dies geschieht in völliger Analogie zur Situation in natürlich entstandenen Enzymen, da auch hier verschiedene Einflüsse bei der „Wahl“ der Aminosäure für eine Position miteinander konkurrieren.

Die Gewichte der TransCent-Module werden so festgelegt, dass sie die Leistungsfähigkeit des Designalgorithmus bei Rekapitulationsrechnungen auf einem Testdatensatz maximieren. Diese Form der Parametrisierung ist zwar die einzige praktikable Lösung des Optimierungsproblems, birgt allerdings auch Risiken. Es besteht dabei nämlich die Gefahr der „Überanpassung“ (engl. *overfitting*). Dies bedeutet, dass die frei wählbaren Parameter einer Methode so gewählt werden, dass die Ergebnisse für die präsentierten Fälle optimal sind, ohne jedoch die zugrunde liegenden Gesetzmäßigkeiten abzubilden. Um dem vorzubeugen sollte der Testdatensatz möglichst groß sein und in einen Trainings- bzw. Evaluierungsdatensatz unterteilt werden. Falls nicht ausreichend Beispielfälle verfügbar sind, können auch Methoden wie z.B. die Leave-One-Out-Kreuzvalidierung zum Einsatz kommen. Damit eine Methode gut generalisiert, sollte darüber hinaus der Testdatensatz variantenreich sein und möglichst die gesamte Bandbreite denkbarer Fälle abdecken. Der in dieser Arbeit neu zusammengestellte Datensatz für TransCent umfasst 53 Enzyme aus allen sechs EC-Klassen (siehe Tabelle 3.5) und ist daher gut geeignet für die Bestimmung der Gewichte.

Da die TransCent-Energieeinheiten keiner physikalischen Größe entsprechen, kann die Energiefunktion beliebig skaliert werden. Folglich darf eines der Gewichte frei gewählt

werden. Nachdem das Modul für Proteinstabilität die Basis des Designalgorithmus bildet, wird das Rosetta-Gewicht in Gleichung (3.1) auf $\omega_{\text{Rosetta}} = 1$ gesetzt.

Die restlichen drei Gewichte ω_{DSX}, $\omega_{\text{Fingerprint}}$ und ω_{PROPKA} werden mittels einer Gridsuche bestimmt. Bei dieser Art von Parameteroptimierung werden die Gewichte als Freiheitsgrade betrachtet, welche den Raum möglicher Lösungen aufspannen. Um die optimale Gewichtskombination zu finden, wird der Lösungsraum abgetastet und die einzelnen „Punkte“ bewertet. Dies geschieht in Form eines Gitters, welches über den Lösungsraum gelegt wird. Der Vorteil dieser Methode gegenüber einer Einzeloptimierung ist, dass Abhängigkeiten zwischen den Modulen implizit berücksichtigt werden.

In dieser Arbeit wurden für jedes Gewicht fünf Werte zwischen 2^{-4} und 2^3 gewählt (siehe Tabelle 4.1) und für die daraus resultierenden $5^3 = 125$ Kombinationen Rekapitulationsdesigns der 53 Enzyme des Testdatensatzes durchgeführt. Für jede Gewichtskombination wurden zehn Designoptimierungen berechnet und das energetisch beste Ergebnis verwendet. Insgesamt wurden damit 1250 Modelle je Enzym berechnet. Die mutierbaren Positionen für die Designs wurden, wie in 4.2.1 beschrieben, automatisch ermittelt.

Modul	**Gewichte**				
Fingerprint	2^{-4}	2^{-1}	2^{0}	2^{1}	2^{3}
DSX	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^{0}
PROPKA	2^{-4}	2^{-1}	2^{0}	2^{1}	2^{3}

Tabelle 4.1: Wertebereich der Modulgewichte bei der ersten Gridsuche

Um den Designerfolg bei einer Gewichtskombination zu bewerten, werden für jedes Modell die drei Ähnlichkeitsmaße μ_{SID}, μ_{BLOSUM} und μ_{PSSM} berechnet (siehe 3.10.2) und über die Ergebnisse der 53 Enzyme gemittelt. Die Rekapitulationsrate μ_{SID} der erfolgreich „wiedergefundenen“ Positionen ist zwar als Maß intuitiv leicht zu erfassen, ignoriert aber das Designergebnis an allen anderen Positionen. Daher ist der sensitivere Blosum-Score besser geeignet, um die optimale Gewichtskombination zu finden. Der PSSM-Score μ_{PSSM} bewertet die Ähnlichkeit eines Designmodells zu einer ganzen Gruppe homologer Enzyme. Deswegen unterscheidet sich dessen Bedeutung leicht von der des Blosum-Scores. Mit ihm kann eine Aussage darüber getroffen werden, wie stark ein Modell natürlich vorkommenden Enzymen mit einer bestimmten Funktion ähnelt. Allerdings werden dadurch keine Korrelationen zwischen einzelnen Proteinpositionen erfasst, weswegen die Entscheidung bei der Wahl der Gewichte anhand des Blosum-Scores getroffen wurde. An dieser Stelle sei betont, dass für die Bestimmung aller Ähnlichkeitsmaße ausschließlich die mutierbaren Positionen betrachtet werden.

Für die Berechnung der Ergebnisse wurde eine Leave-One-Out-Kreuzvalidierung durchgeführt, d.h. es wurde zunächst die beste Gewichtskombination für alle, jeweils 52 Enzyme umfassenden Teilmengen des Testdatensatzes ermittelt und über die Werte für das jeweils verbleibende 53-ste Enzym gemittelt. Da in allen Fällen dieselbe Gewichtskombination die besten Resultate lieferte, sind die Ergebnisse identisch mit den arithmetischen Mittelwerten über den gesamten Datensatz. Diese sind in Abbildung 4.5 zusammengefasst dargestellt. Jede Kugel repräsentiert einen Satz von Gewichten. Färbung und Radius der Kugeln richten sich nach den Mittelwerten der Ähnlichkeitsmaße. Die Kugel, welche die

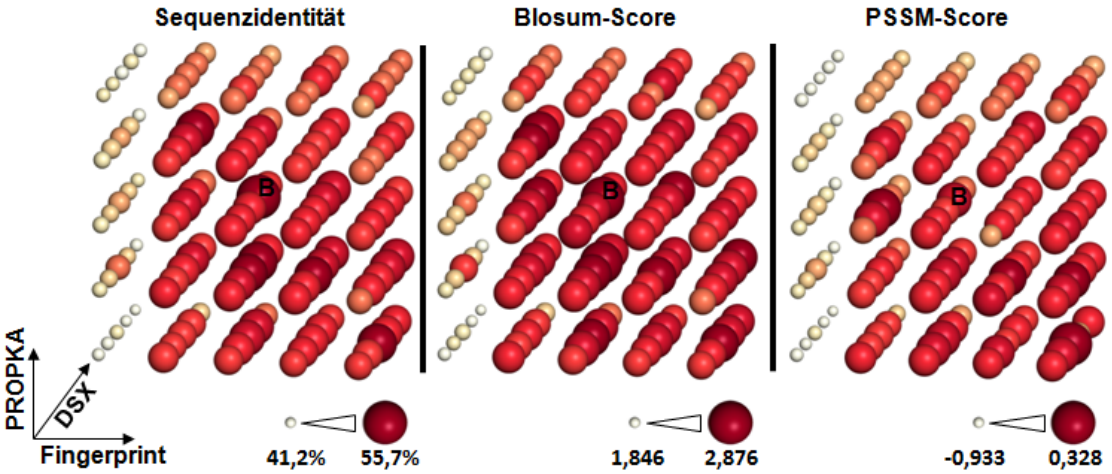


Abbildung 4.5: Ergebnisse der ersten Gridsuche
Dargestellt sind die Resultate der ersten Gridsuche basierend auf den Designergebnissen für die 53 Enzyme des Rekapitulationsdatensatzes. Jede Kugel entspricht einer der 125 Gewichtskombinationen für die Werte aus Tabelle 4.1. Die Koordinatenachsen deuten die Richtung an, in der die Größe des jeweiligen Modulgewichts zunimmt. Sowohl die Farbe aus auch der Durchmesser der Kugeln kodieren für den Wert der mittleren Sequenzidentität, den Blosum-Score bzw. den PSSM-Score gemäß der angegebenen Wertebereiche. Die Skalierung der Werte ist nicht linear, um die Anschaulichkeit zu verbessern. Die insgesamt beste Gewichtskombination ist jeweils mit einem „B“ markiert.

beste Gewichtskombination verkörpert, ist in allen Teildiagrammen mit einem „B“ gekennzeichnet. Mit dem Gewichtssatz:

$$\begin{aligned} \omega_{\text{DSX}} &= 0,5 \\ \omega_{\text{Fingerprint}} &= 1,0 \\ \omega_{\text{PROPKA}} &= 1,0 \end{aligned}$$

erreicht TransCent einen mittleren Blosum-Score von 2,876, was einer Rekapitulationsrate von durchschnittlich 55,7% entspricht. Betrachtet man die Einzelergebnisse für die Enzyme des Datensatzes so fällt auf, dass die Streuung der Werte zwischen 1,700 und 4,440 bzw. zwischen 37% und 80% relativ stark ist. Dabei sind die Unterschiede zwischen den Enzymen oft sogar größer als die Abweichungen innerhalb der Werte für ein Enzym. Dies zeigt, dass die Leistungsfähigkeit von TransCent stärker von der Art des präsentierten Problems abhängt als von der Wahl der Gewichte.

Der mittlere PSSM-Wert von 0,216 ist ebenfalls überdurchschnittlich, wird allerdings von einer anderen Gewichtskombination ($\omega_{\text{DSX}} = 0,25$, $\omega_{\text{Fingerprint}} = 8,0$, $\omega_{\text{PROPKA}} = 0,0625$) bei einer Rekapitulationsrate von 55,0% und einem ebenfalls guten Blosum-Score von 2,819 mit 0,328 klar übertroffen. Möglich macht dies das verhältnismäßig hohe Gewicht des Fingerprint-Moduls, welches die anderen drei Module dadurch jederzeit „überstimmen“ kann. Es erzwingt somit, das alle Potentiale optimal erfüllt werden unabhängig von möglichen negativen Auswirkungen auf die Proteinstabilität und die pK_a-Wert-Optimierung.

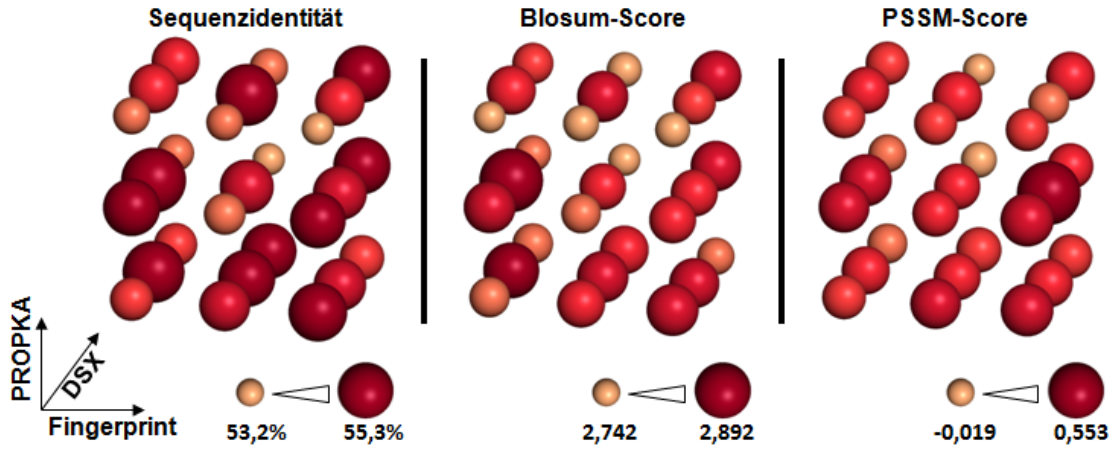


Abbildung 4.6: Ergebnisse der zweiten Gridsuche
Dargestellt sind die Ergebnisse der zweiten Gridsuche mit feinerer Auflösung um das Optimum der ersten Gridsuche basierend auf Rekapitulationsrechnungen für die 53 Enzyme des Rekapitulationsdatensatzes. Jede Kugel entspricht einer der 27 Gewichtskombinationen für die Werte aus Tabelle 4.2. Die Koordinatenachsen deuten die Richtung an, in der die Größe des jeweiligen Modulgewichts zunimmt. Sowohl die Farbe aus auch der Durchmesser der Kugeln kodieren für den Wert der mittleren Sequenzidentität, den Blosum-Score bzw. den PSSM-Score gemäß der angegebenen Wertebereiche. Die Skalierung der Werte ist nicht linear, um die Anschaulichkeit zu verbessern.

Auch um solcherlei Artefakte zu vermeiden wurde die oben beschriebene Wahl der Gewichte getroffen.

Zur Kontrolle der Ergebnisse wurde in einer zweiten Gridsuche der Lösungsraum um das gefundene Optimum mit feinerer Auflösung nochmals abgetastet (siehe Tabelle 4.2). Für die $3^3 = 27$ Gewichtskombinationen wurden wiederum je zehn Designoptimierungen durchgeführt.

Modul	**Gewichte**		
Fingerprint	0,75	1,00	1,25
DSX	0,25	0,50	0,75
PROPKA	0,75	1,00	1,25

Tabelle 4.2: Wertebereich der Modulgewichte bei der zweiten Gridsuche

Der beste gefundene mittlere Blosum-Score beträgt 2,892 mit einer Rekapitulationsrate von 55,3% für die Gewichtskombination $\omega_{\text{DSX}} = 0,5$, $\omega_{\text{Fingerprint}} = 0,75$ und $\omega_{\text{PROPKA}} = 1,0$, was in guter Näherung die Ergebnisse der ersten Gridsuche bestätigt.

4.2.3 Performanzanalyse

Im letzten Kapitel wurden mithilfe einer Gridsuche optimale Gewichte für die TransCent-Module bestimmt. Die folgende *in silico* Analyse der Rekapitulationsrechnungen gibt Auf-

schluss darüber, wie erfolgversprechend damit Designs von Funktionsübertragungen sein können und wo auf mögliche Schwächen des Programms geachtet werden muss.

4.2.3.1 Vergleich mit TransCent1.0

Zunächst soll, wiederum anhand von Rekapitulationsrechnungen, gezeigt werden, welche Änderungen sich durch den Austausch bzw. die Überarbeitung der Module für die Vorhersagequalität gegenüber der ersten Version von TransCent ergeben. Für einen fairen Vergleich der beiden Versionen muss dieser auf einer gemeinsamen Datenbasis durchgeführt werden. Dazu wurde der 27 Enzyme umfassende Testdatensatz (siehe Tabelle 3.4) aus [85] herangezogen, für den die Ergebnisse der Rekapitulationsrechnungen in [85] angegeben sind. Allerdings mussten zwei Enzyme ausgeschlossen werden (PDB-Code: 1f8e, 1po5) bei denen die Aminosäurehäufigkeiten für die Energiefunktion des PROPKA-Moduls (vgl. Gl. (3.23)) nicht bestimmt werden konnten.

Für die restlichen 25 Enzyme wurden je 100 Designoptimierungen durchgeführt und jeweils das Ergebnis mit der niedrigsten Energie ausgewertet. Dabei wurden die Schalendefinitionen (vgl. 4.2.1) aus [85] wiederverwendet, so dass dieselben 678 Positionen als mutierbar gekennzeichnet waren. Für die Modul-Gewichte wurden die Werte verwendet, welche sich bei der Gridsuche als optimal erwiesen haben ($\omega_{\mathrm{DSX}} = 0,5$, $\omega_{\mathrm{Fingerprint}} = 1,0$, $\omega_{\mathrm{PROPKA}} = 1,0$).

Die Auswertung der Ergebnisse ergibt einen mittleren Blosum-Score von 2,7 und eine Rekapitulationsrate von 55,7%. Dies entspricht einer moderaten Verbesserung gegenüber der alten TransCent-Version, die mit Werten von 2,6 bzw. 54,3% für Blosum-Score und Rekapitulationsrate zu Buche steht [85].

4.2.3.2 Beitrag der Module

Um die Leistungssteigerung besser den einzelnen Modulen zuordnen zu können, wurden Rekapitulationsrechnungen mit unterschiedlichen Modulkombinationen durchgeführt. Das Rosetta-Modul bildet die Grundlage des TransCent-Algorithmus und muss daher immer aktiviert bleiben. Die anderen Module können beliebig hinzugenommen werden, was insgesamt zu acht verschiedenen Kombinationen führt. Die Modulgewichte wurden unabhängig von der gewählten Kombination auf die optimalen Werte aus der Gridsuche gesetzt (siehe 4.2.2). Für jedes der 53 Enzyme des Rekapitulationsdatensatzes (siehe 3.10.3) wurden 400 Modelle berechnet, 50 je Modulkombination, und jeweils dasjenige mit dem besten Energiewert ausgewertet.

Die Ergebnisse sind zusammen mit den Vergleichswerten aus [85] in Abbildung 4.7 dargestellt. Die Kennzahlen für die alte TransCent-Version wurden zwar auf einem anderen Datensatz berechnet, die gute Übereinstimmung der Werte für den vollen Modulsatz mit den Ergebnissen aus 4.2.3.1 legt jedoch nahe, dass die Zahlen durchaus miteinander vergleichbar sind.

Auffällig ist, dass alleine durch den Wechsel von Rosetta++ zur *score12*-Energiefunktion von Rosetta3 eine Verbesserung des Blosum-Scores um 0,5 erreicht werden kann. Die größte Steigerung weist die Modulkombination von Rosetta und PROPKA auf. Der Blosum-Score nimmt um 0,6 zu bei gleichzeitiger Verbesserung der Rekapitulationsrate um 5,8 Prozentpunkte. Betrachtet man die Kombination aller Module, so fallen die Unterschiede

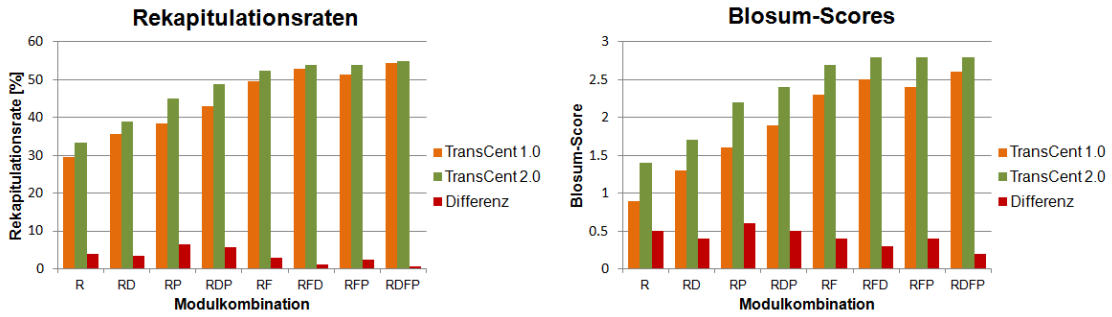


Abbildung 4.7: Vergleich von Rekapitulationsrate und Blosum-Score für unterschiedliche Modulkombinationen
Die Werte für TransCent 1.0 wurden aus [85] übernommen, stammen allerdings von einem anderen Datensatz. Die Angabe Modulkombination bezeichnet jeweils, welche Module bei der Berechnung aktiviert waren. (R=Rosetta, D=DSX, F=Fingerprint, P=PROPKA)

weit geringer aus, nämlich nur 0,6 Prozentpunkte bei der Rekapitulationsrate und 0,2 Punkte beim Blosum-Score.

Die Werte belegen zum einen, dass auch mit den Änderungen, welche an TransCent vorgenommen wurden, alle Module signifikant zur Performanz des Programms beitragen. Andererseits zeigt sich, dass die Effekte der einzelnen Module in hohem Maße nicht additiv sind. Betrachtet man nur die Zwei-Modul-Kombinationen, so erhält man gegenüber den Ergebnissen, welche lediglich mit dem Rosetta-Modul erreicht werden, eine Ratensteigerung von 5,4, 11,5 bzw. 19,0 Prozentpunkten für die Hinzunahme des DSX-, PROPKA- bzw. Fingerprint-Moduls. Die tatsächliche Steigerung von 21,5 Prozentpunkten für die Kombination aller Module liegt somit weit unter dem theoretischen Wert von 35,9, der sich aus der Summe der Einzelbeiträge ergeben würde.

Während die Beiträge des DSX- und des PROPKA-Moduls näherungsweise als unabhängig betrachtet werden könnten (15,4 statt theoretisch 16,9 Prozentpunkte), sind sich die Effekte des Fingerprint- und des DSX-Moduls offenbar sehr ähnlich (20,6 statt 30,5 Prozentpunkte). Dies ist nicht weiter verwunderlich, da durch das Einführen spezifischer Wechselwirkung durch das Fingerprint-Modul gleichzeitig auch die Ligandenbindung optimiert wird. Dies unterstreicht die Notwendigkeit der gemeinsamen Optimierung aller Modulgewichte.

Ein ähnliches Bild zeigt sich, wenn man die Auswirkungen der Moduländerungen (siehe 4.1) betrachtet. Obwohl alleine der Wechsel des Rosetta-Moduls eine Verbesserung des Blosum-Score um 0,5 Punkte bewirkt, kann für die Gesamtkombination nur eine Zunahme um 0,2 Punkte festgestellt werden.

Vergleicht man die Ergebnisse der Gridsuche mit den Resultaten für die Kombination aller Module, so fällt auf, dass sowohl die mittlere Rekapitulationsrate (54,9% bzw. 55,7%) als auch der gemittelte Blosum-Score (2,8 bzw. 2,7) voneinander abweichen, obwohl die identische Datenbasis und der gleiche Gewichtssatz verwendet wurden. Dies kommt dadurch zustande, dass TransCent ein heuristisches Verfahren als Optimierungsroutine verwendet (siehe 3.1.1). Das *Simulated Annealing* Protokoll kann nicht garantieren, dass bei jeder Designoptimierung die Lösung gefunden wird, die dem globalen Energieminimum entspricht. Stattdessen liefert der Algorithmus „nur“ Modelle mit hinreichend guten Energiewerten,

welche sich mehr oder minder stark unterscheiden können. Dies führt zu Schwankungen bei den TransCent-Leistungsdaten. Erschwerend kommt hinzu, dass keine perfekte (Anti-) Korrelation zwischen TransCent-Energie und Blosum-Score bzw. Rekapitulationsrate zu erwarten ist.

4.2.3.3 PROPKA-Auswertung

In [85] sowie in Unterkapitel 4.2.3.2 wurde gezeigt, dass durch Hinzunahme des PROPKA-Moduls die Performanz von TransCent bei Rekapitulationsrechnungen hinsichtlich Blosum-Score und Rekapitulationsrate erhöht werden kann. Zusätzlich soll hier nun untersucht werden, welchen Einfluss die Verwendung von PROPKA auf die pK_a-Werte der titrierbaren Gruppen hat.

Grundlage der Auswertung bilden Rekapitulationsdesigns des Testdatensatzes (vgl. 3.10.3), wobei für jedes Enzym 1000 Designoptimierungen durchgeführt wurden. Zusätzlich wurden je Enzym 100 Modelle ohne aktiviertes PROPKA-Modul erstellt. Der Vergleich erfolgt anhand der Ergebnisse für die Modelle mit den jeweils besten TransCent-Energiewerten.

Insgesamt enthält der Datensatz 366 mutierbare Positionen, welche vom Fingerprint-Modul erfasst werden und wildtypisch mit titrierbaren Gruppen besetzt sind. Unter Verwendung von PROPKA werden 333 dieser Positionen korrekt modelliert, was einer beeindruckenden Rekapitulationsrate von rund 91% entspricht. Dies kann jedoch ursächlich nur bedingt dem PROPKA-Modul zugeschrieben werden, da TransCent auch ohne dessen Einsatz eine Rate von 88% (321 Positionen) erreicht. Aus einem Vergleich aller mutierbaren Positionen der 53 Enzyme folgt, dass eine Steigerung der Rekapitulationsrate von 54,8% auf 56,0% erreicht wurde.

Das eigentliche Ziel des PROPKA-Moduls ist die Optimierung der pK_a-Werte, d.h. die pK_a-Werte der katalytisch wichtigen Residuen sollen möglichst gut mit den nativen Werten übereinstimmen (siehe 3.5.3). Hier zeigt sich, dass mit PROPKA eine mittlere betragsmäßige Abweichung von 0,77 erreicht wird, wohingegen bei inaktivem Modul der Unterschied im Mittel 1,75 pH-Einheiten beträgt. Für den Vergleich werden nur Werte der Positionen herangezogen, welche sowohl mit als auch ohne PROPKA-Modul korrekt modelliert wurden. Für diese 314 Residuen werden die gemittelten Beträge der pK_a-Wert-Differenzen angegeben. Insgesamt kann also die mittlere Abweichung der pK_a-Werte um rund eine pH-Einheit reduziert werden.

Die Histogramme in Abbildung 4.8 zeigen die Verteilungen der pK_a-Wert-Differenzen für beide Fälle. Mit PROPKA liegt bei weit über der Hälfte aller Residuen der Unterschied zum Zielwert unter 0,5. Außerdem kann damit der Anteil von Positionen mit starken Abweichungen (>3 pH-Einheiten) von 21% auf 4% reduziert werden.

Betrachtet man den Wertebereich der berechneten pK_a-Werte so liegt dieser zwischen -5,33 und 20,15 mit PROPKA-Modul bzw. zwischen -5,04 und 23,01 ohne. Diese teilweise sehr extrem liegenden Werte kommen allerdings weniger durch Fehler im Designalgorithmus zustande, sondern sind auf Artefakte bei der pK_a-Wert-Berechnung selbst zurückzuführen, denn auch die berechneten wildtypischen Werte reichen von -4,94 bis 21,44. Nachdem es sich hierbei um systematische Fehler handelt, welche sowohl bei der Berechnung der Referenzwerte als auch beim Design selbst auftreten, ist dieses Verhalten unkritisch für den Designerfolg, da in die Energiefunktion (siehe Gl. (3.23)) nur pK_a-Wert-Differenzen eingehen.

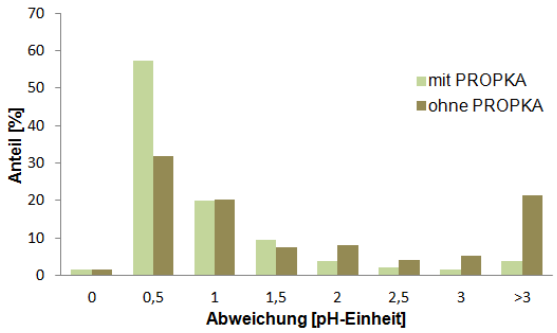


Abbildung 4.8: Relative Verteilung der pK_a-Wert-Abweichungen
Die beiden Histogramme zeigen die anteilsmäßige Verteilung der betragsmäßigen Differenzen zu den vorgegebenen Referenz-pK_a-Werten für 333 (mit PROPKA) bzw. 321 (ohne PROPKA) korrekt modellierte Positionen mit titrierbaren Gruppen. Die Balken bei 0 repräsentieren die Fälle, bei denen die Vorgabe exakt erreicht wurde.

4.2.3.4 Variationsbreite der TransCent-Ergebnisse

Die Performanz von TransCent hängt nicht nur von der Art des Problems, d.h. von der zu übertragenden Enzymfunktion, und den Werten der Modulgewichte ab. Auch die Anzahl der durchgeführten Designoptimierungen hat einen Einfluss auf die Qualität der Ergebnisse. Der Grund dafür ist, dass es sich bei der verwendeten Optimierungsheuristik um einen Monte Carlo Algorithmus handelt. Dessen stochastische Natur führt dazu, dass sich die Ergebnisse mehrerer Designoptimierungen mehr oder weniger stark unterscheiden können. Einerseits wird durch das *Simulated Annealing* Verfahren sicher gestellt, dass ein Designergebnis einem lokalen Optimum in der Energielandschaft entspricht, andererseits hängt es aber von der Form dieser Landschaft ab, wie ähnlich die Lösung dem globalen Minimum ist. Die übliche Strategie bei Verfahren dieser Art ist, dass durch wiederholte Ausführung der Optimierung eine Vielzahl von Ergebnissen berechnet wird. Wird hierbei die energetisch beste Lösung mehrfach gefunden, so ist dies ein starker Hinweis dafür, dass es sich dabei um das globale Optimum handelt.

Für TransCent folgt daraus, dass es im Allgemeinen nicht ausreicht, eine oder wenige Designoptimierungen durchzuführen, um das bestmögliche Modell zu berechnen. Stattdessen sollten möglichst viele Modelle erstellt werden, um eine gründliche Abtastung des Lösungsraum zu gewährleisten. Eine einzige Designoptimierung kann allerdings mehrere Minuten in Anspruch nehmen, so dass aufgrund der begrenzt zur Verfügung stehenden Rechenzeit nicht beliebig viele Modelle berechnet werden können. Es muss also ein Richtwert ermittelt werden, ab dem sichergestellt ist, dass die optimale Lösung mit hinreichend hoher Wahrscheinlichkeit gefunden wird. Da bei einer Funktionsübertragung, anders als bei einer Rekapitulationsrechnung, ausschließlich der TransCent-Energiewert zur Bewertung eines Modells dient, ist die entscheidende Frage, wie viele Designoptimierungen notwendig sind, bis eine Lösung mit minimaler Energie gefunden wird.

Da es nicht möglich ist, das tatsächliche globale Minimum der TransCent-Energiefunktion zu ermitteln, wird dieses für die folgende Auswertung durch den niedrigsten Energiewert angenähert, der nach 1000 Designoptimierungen erreicht werden konnte (nachfolgend „Re-

ferenzoptimum“ genannt). Alle angegebenen Ergebnisse sind Mittelwerte für die 53 Enzyme des Testdatensatzes (siehe 3.10.3).

Um nun herauszufinden, wie viele Modelle mindestens erstellt werden müssen, wurde untersucht, wie weit das beste Ergebnis nach 10, 50, 100, 300 und 500 Designoptimierungen vom Referenzoptimum entfernt ist. Abbildung 4.9 zeigt die Kurven der kumulativen Histogramme für die prozentuale Abweichung der niedrigsten Energie vom Referenzoptimum. Wie man sieht reichen zehn Designoptimierungen bei keinem der Enzyme aus, um das Optimum zu erreichen und bei ca. 30% aller Fälle liegt der beste Energiewert um mehr als 3% davon entfernt. Im Vergleich dazu wird für 60% der Enzyme das optimale Ergebnis bereits nach 500 Optimierungen gefunden. Akzeptiert man eine Abweichung von bis zu einem Prozent so wird schon nach 300 Designoptimierungen mit einer Wahrscheinlichkeit von 90% ein hinreichend guter Energiewert erreicht.

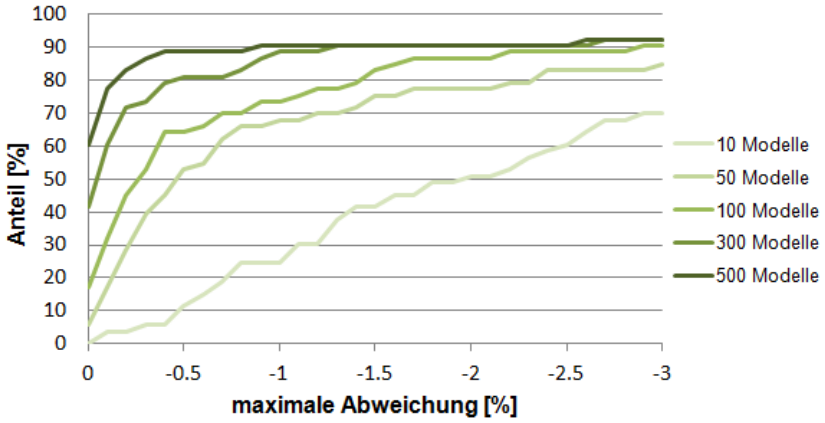


Abbildung 4.9: Kumulative Verteilung der besten Designmodelle mit Abstand zum Referenzoptimum abhängig von der Anzahl der Designoptimierungen
Die Kurven zeigen den Anteil der 53 Enzyme des Rekapitulationsdatensatzes, für die nach 10, 50, 100, 300 und 500 Designoptimierungen ein Modell erzeugt wurde, dessen Energie maximal um den jeweiligen Prozentsatz unter dem Referenzoptimum liegt. Die Werte bei einer Abweichung von 0% entsprechen dem Anteil der Enzyme, für die das Referenzoptimum nach der entsprechenden Anzahl von Designoptimierungen erreicht wurde.

Betrachtet man statt der Energie die Entwicklung der Rekapitulationsrate, so ergibt sich ein ähnliches Bild. Abbildung 4.10 zeigt deren Verlauf in Abhängigkeit von der Anzahl N der durchgeführten Designoptimierungen. Jeder Wert ist dabei gemittelt über die 53 Enzyme des Testdatensatzes, wobei zur Ermittlung der Rekapitulationsrate jeweils das energetische beste der ersten N Modelle verwendet wurde. Der dargestellte Kurvenverlauf verdeutlicht, dass der Anteil wildtypisch besetzter Residuen durch Berechnung weiterer Modelle tendenziell erhöht werden kann, wobei bei rund 300 Designoptimierungen eine Art Sättigungseffekt eintritt. Dies bestätigt die oben angeführten Beobachtungen hinsichtlich der Anzahl notwendiger Designoptimierungen.

Zudem fällt auf, dass der relative Gewinn, der durch die Erhöhung der Anzahl von Designoptimierungen erzielt wird, verhältnismäßig gering ist, denn selbst mit einer einzigen Designoptimierung können Modelle mit einer durchschnittlichen Rekapitulationsrate von

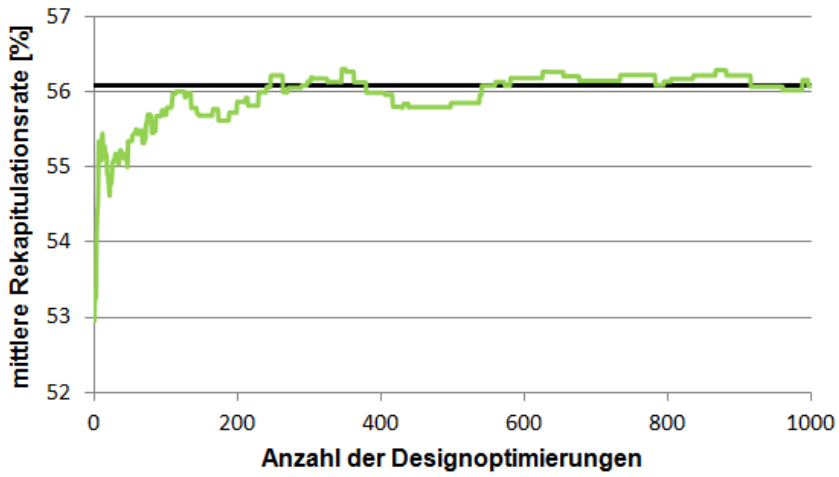


Abbildung 4.10: Entwicklung der mittleren Rekapitulationsrate mit zunehmender Anzahl von Designoptimierungen
Die angegebene Rekapitulationsrate ist das Mittel über die 53 Enzyme des Rekapitulationsdatensatzes. Insgesamt wurden jeweils 1000 Designmodelle berechnet. Zur Berechnung der Rekapitulationsraten wird jeweils das energetisch beste Modell der ersten N Ergebnisse verwendet. Die schwarze Linie repräsentiert den mittleren Wert, der sich für die Gesamtheit der jeweils 1000 Designmodelle ergibt.

53% erzeugt werden. Da jedoch bereits eine ungünstig gewählte Aminosäure die enzymatische Aktivität eines Proteins entscheidend zu beeinträchtigen vermag, kann die Anzahl der berechneten Modelle trotzdem maßgeblich für den Erfolg einer Funktionsübertragung sein.

Anders als bei der Betrachtung der Energie verläuft die Kurve in Abbildung 4.10 nicht monoton steigend. Dies bedeutet, dass Modelle trotz eines niedrigeren Energiewertes eine schlechtere Rekapitulationsrate haben können, und ist ein Indiz dafür, dass keine perfekte Antikorrelation zwischen den beiden Werten existiert.

Die Korrelation zwischen TransCent-Energie und Blosum-Score bzw. TransCent-Energie und Rekapitulationsrate ist aber eine wichtige Größe. Sie entscheidet im Einzelfall darüber, wie viele Designmodelle tatsächlich umgesetzt und getestet werden müssen, um mit hoher Wahrscheinlichkeit das bestmögliche Modell zu erfassen. Zur Untersuchung von Korrelationen zwischen zwei Größen, können unterschiedliche Maße verwendet werden:

Der Pearson-Korrelationskoeffizient ρ_{Pearson} beschreibt allgemein den linearen Zusammenhang zweier Merkmale. Die Spearman'sche Rangkorrelation ρ_{Spearman} gibt ebenfalls an, inwieweit ein Zusammenhang zwischen den Werten zweier Größen besteht, setzt aber keine lineare Abhängigkeit voraus. Der Wertebereich beider Maße liegt zwischen -1 und +1, wobei ein Wert von +1 bzw. -1 einer perfekten Korrelation bzw. Antikorrelation entspricht. Ist der Korrelationskoeffizient gleich 0, so kann kein (linearer) Zusammenhang festgestellt werden.

Für die Evaluation der TransCent-Energiefunktion wurden basierend auf den Ergebnissen von 1000 Designoptimierungen die enzymspezifischen Korrelationskoeffizienten zwischen dem Energiewert und dem Blosum-Score bzw. der Rekapitulationsrate berechnet.

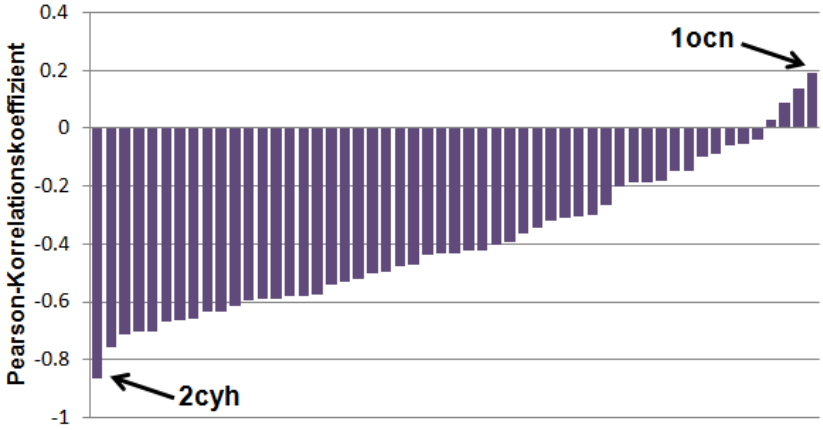


Abbildung 4.11: Verteilung der Pearson-Korrelationskoeffizienten zwischen TransCent-Energie und Blosum-Score für den Rekapitulationsdatensatz
Grundlage für die Berechnung der Korrelation sind die Wertepaare TransCent-Energie und Blosum-Score von jeweils 1000 Designmodellen.

Im Mittel erhält man für die Enzyme des Testdatensatzes Werte von $\rho_{\text{Spearman}} = -0,37$ und $\rho_{\text{Pearson}} = -0,36$ für die Korrelation zwischen Rekapitulationsrate und Energiewert und mit $\rho_{\text{Spearman}} = \rho_{\text{Pearson}} = -0,39$ minimal bessere Werte für den Blosum-Score. Die Standardabweichung aller berechneten Koeffizienten ist mit 0,25 jedoch ungewöhnlich hoch und zeigt, dass im Einzelfall die Werte relativ stark vom Mittelwert abweichen können. Um dies zu verdeutlichen ist in Abbildung 4.11 exemplarisch die Verteilung der 53 Pearson-Korrelationskoeffizienten für den Blosum-Score aufgetragen. Der Wertebereich reicht von einer annähernd perfekten Antikorrelation ($\rho_{\text{Pearson}} = -0,87$) für das humane Enzym Cyclophilin A (PDB-Code: 2cyh) bis hin zu einem positiven Korrelationswert von $\rho_{\text{Pearson}} = 0,19$ für die Cellobiohydrolase-II aus *Humicola insolens* (PDB-Code: 1ocn).

Da bei 1000 Designoptimierungen eine hinreichend genaue Abtastung des Lösungsraumes gesichert sein sollte, kann die Ursache für die mangelnde Antikorrelation in einigen Fällen nur auf die Form der Energielandschaft selbst zurückzuführen sein. Um dies zu illustrieren sind in Abbildung 4.12 für die beiden oben genannten Enzyme die Energiewerte der 1000 Modelle gegen ihre Blosum-Scores aufgetragen. Diese stark vereinfachte topologische Darstellung der Energielandschaften beider Enzyme verdeutlicht, dass im Fall von *2cyh* ein breites globales Energieminimum existiert, welches von zahlreichen höherenergetischen lokalen Minima umgeben ist. Dies äußert sich in einem annähernd linearen Zusammenhang zwischen Energiewert und Blosum-Score. Ein anderes Bild ergibt sich im Fall von *1ocn*. Hier existiert neben dem eigentlich angestrebten Minimum noch ein weiteres, tieferes Minimum (angedeutet durch die beiden Ovale), welches zwar weiter von der wildtypischen Sequenz entfernt ist, von der TransCent-Energiefunktion allerdings besser bewertet wird.

Die entscheidende Größe für die experimentelle Überprüfung der TransCent-Ergebnisse ist die Anzahl der Designs die umgesetzt werden müssen, um das bestmögliche Modell zu finden und hängt natürlich unmittelbar mit der beobachteten Korrelation zusammen. Die in Abbildung 4.13 dargestellte Verteilung gibt an, wie viele Modelle von TransCent besser bewertet werden, als dasjenige mit dem höchsten Blosum-Score, welches mit der größten Wahrscheinlichkeit einem aktiven Enzym entspricht. Die Reihenfolge der Modelle

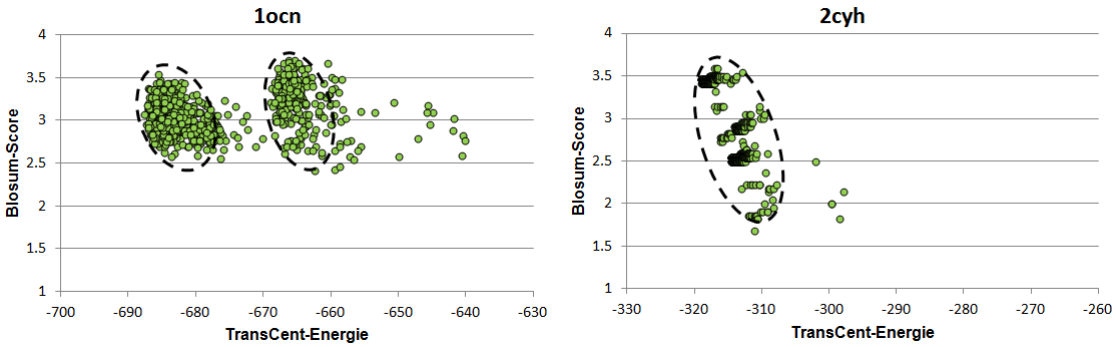


Abbildung 4.12: Vereinfachte Darstellung der Energielandschaft für 1ocn und 2cyh
Die beiden Abbildungen zeigen die Wertepaare von TransCent-Energie und Blosum-Score für jeweils 1000 Designmodelle. Die Ovale deuten die „Energie-Trichter“ an, in denen lokale Minima gehäuft auftreten. Bei 1ocn existiert neben dem Trichter, welcher der wildtypischen Sequenz am nächsten kommt noch ein weiterer, tieferer Trichter.

wird dabei durch den TransCent-Energiewert festgelegt. Da zum einen identische Ergebnisse mehrfach von der Optimierungsroutine erzeugt werden können und andererseits unterschiedliche Rotamerkombinationen dieselbe Aminosäuresequenz repräsentieren können, werden Modelle mit identischer Sequenz zusammengefasst und die „bereinigten“ Ränge angegeben.

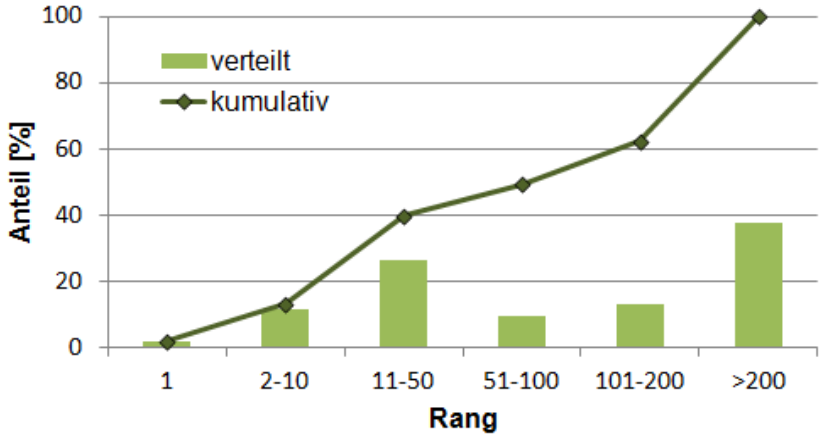


Abbildung 4.13: Rang-Verteilung für die Modelle mit dem größten Blosum-Score
Die Liste der Designmodelle ist geordnet nach TransCent-Energiewert. Identische Modelle, die mehrfach erzeugt wurden, wurden zusammengefasst. Angegeben ist sowohl der Anteil der einzelnen Intervalle als auch die kumulative Verteilung.

Für die Enzyme des Rekapitulationsdatensatzes liegt der Rang-Median des bestmöglichen Modells bei 105. Bei sieben von 53 Enzymen reichen zwischen einem und zehn Designs aus, um das bestmögliche zu erfassen. Im Allgemeinen müssten aber rund 100 Designs tatsächlich umgesetzt werden, um auch nur bei der Hälfte aller Enzyme das potentiell

beste Designergebnis zu erhalten. Dies ist jedoch nicht praktikabel und erfordert, dass weitere Maßnahmen ergriffen werden, um die notwendige Mindestanzahl zu verringern. Dies kann z.B. durch die Verwendung von Clusteralgorithmen geschehen [156], welche die Designmodelle in Gruppen zusammenfassen, von denen dann jeweils ein Repräsentant getestet wird. Eine sinnvolle Vorgehensweise ist auch der Einsatz von Expertenwissen. Die Designmodelle können z.B. anhand der vorhergesagten Struktur einzeln begutachtet werden, wobei auch die Möglichkeit besteht, Kriterien anzulegen, welche im Designalgorithmus nicht realisiert sind. Außerdem könnte auf eventuell vorhandenes Wissen aus anderen Quellen wie z.B. Mutationsstudien für das Zielenzym zurückgegriffen werden.

In den Bereich Expertenwissen fällt auch das Wissen über grundlegende Eigenschaften und Präferenzen des Programms TransCent bzw. dessen Stärken und Schwächen, die im Folgenden näher untersucht werden sollen.

4.2.3.5 Vergleich der aminosäurespezifischen Rekapitulationsraten

Die Ergebnisse der Rekapitulationsrechnungen zeigen, dass TransCent in der Lage ist, im Mittel bei über der Hälfte aller mutierbaren Positionen die wildtypische und somit höchstwahrscheinlich beste Aminosäure zu identifizieren (siehe 4.2.3.2). Dies bedeutet aber gleichzeitig, dass in vielen Fällen nicht die wildtypische Aminosäure gewählt wird. Um die Ursachen dafür zu analysieren, werden zunächst die Aminosäureverteilungen der TransCent-Modelle und der wildtypischen Enzyme miteinander verglichen. Für die Auswertung wird im Folgenden jeweils die Besetzung der insgesamt 2233 mutierbaren Positionen der 53 Enzyme des Rekapitulationsdatensatzes betrachtet.

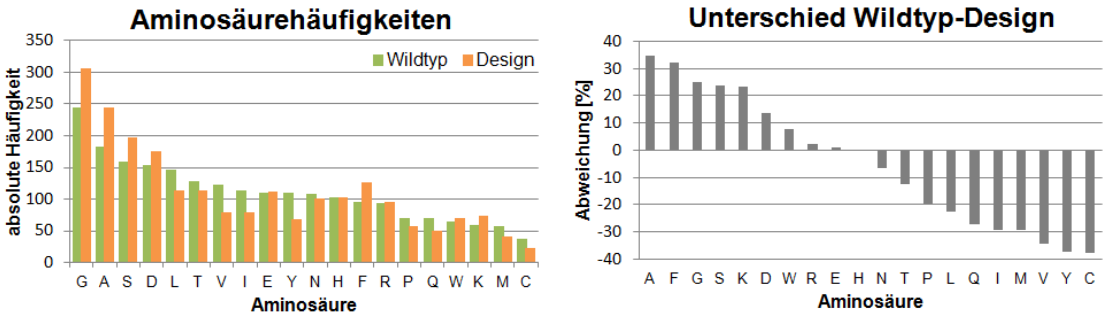


Abbildung 4.14: Aminosäurehäufigkeiten bei Designmodellen und wildtypischen Enzymen
In der linken Abbildung sind die Verteilungen der absoluten Aminosäurehäufigkeiten auf den 2233 mutierbaren Positionen des Rekapitulationsdatensatzes für die Designmodelle und die Wildtypsequenzen dargestellt. Das rechte Diagramm zeigt die relative Abweichung der beiden Verteilungen je nach Art der Aminosäure.

Aus dem Histogramm der absoluten Aminosäurehäufigkeiten in Abbildung 4.14(a) können einige generelle Trends abgelesen werden. Zum einen nimmt der Anteil der sehr kleinen Aminosäuren Glycin, Alanin und Serin, welche auch wildtypisch am häufigsten vorkommen, nochmals von 26% auf 34% der Gesamtverteilung zu. Im Gegenzug fällt auf, dass das Vorkommen der drei aliphatischen Aminosäuren Valin, Leucin und Isoleucin relativ zum wildtypischen Niveau um jeweils mehr als 20% abnimmt (vgl. Abb. 4.14(b)). Noch größere Abweichungen werden nur für Cystein (-38%) und Tyrosin (-37%) beobachtet.

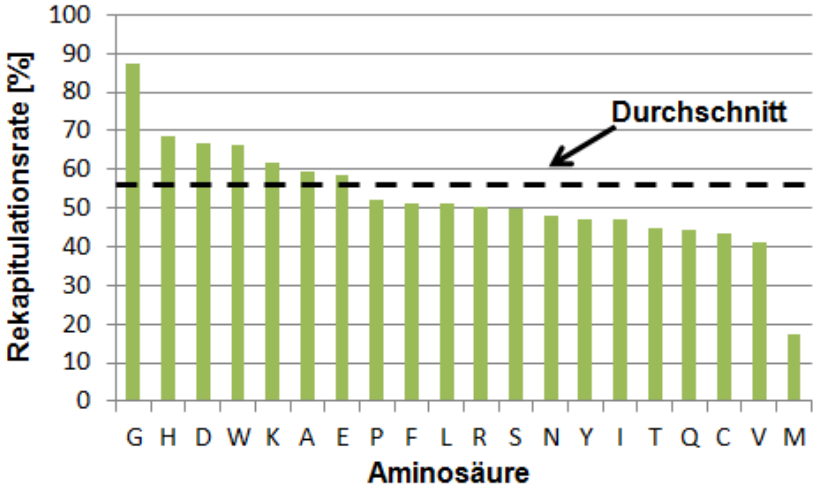


Abbildung 4.15: Rekapitulationsrate aufgeschlüsselt nach Art der Aminosäure
Angegeben ist der Prozentsatz der Positionen, die beim Rekapitulationsdesign korrekt mit der Aminosäure des Wildtyp-Enzyms besetzt werden, aufgeschlüsselt nach Art der wildtypischen Aminosäure. Die durchschnittliche Rekapitulationsrate für alle Positionen liegt bei 56%.

Der Rückgang bei Tyrosin kann hingegen zumindest teilweise anhand der „Ersetzungsmatrix“ in Abbildung 4.16 erklärt werden. Darin angegeben sind die absoluten Häufigkeiten dafür, wie oft eine wildtypische Aminosäure durch eine andere (=Mutation) oder durch sich selbst (=Rekapitulation) beim Design ersetzt wurde. Die Hauptdiagonale der Matrix enthält somit alle Fälle, in denen die modellierte mit der wildtypischen Aminosäure übereinstimmt, aufgeschlüsselt nach Art der Aminosäure. Aus diesen Werten lassen sich dann auch die Rekapitulationsraten der einzelnen Aminosäuren berechnen, welche in Abbildung 4.15 aufgeführt sind. Während man für Methionin einen weit unterdurchschnittlichen Wert von nur 17,2% erhält, werden Positionen die wildtypisch mit einem Glycin besetzt sind zu 87,3% korrekt besetzt. Letzteres ist partiell auch auf die Überrepräsentation der Aminosäure in den Designmodellen zurückzuführen.

Betrachtet man nun in der „Ersetzungsmatrix“ die Zeile für Tyrosin, so fällt auf, dass neben dem Maximalwert von 52 erfolgreichen Rekapitulationen ein weiterer verhältnismäßig hoher Wert von 17 Mutationen hin zu Phenylalanin zu Buche steht. Dabei handelt es sich um eine sehr ähnliche Aminosäure, die sich von Tyrosin nur durch eine OH-Gruppe unterscheidet, was auch durch einen Blosum62-Score von 1 zum Ausdruck kommt. Da nur vier Mutationen in entgegengesetzter Richtung beobachtet wurden, erklärt dieser Effekt zumindest teilweise die oben beschriebene Abnahme der Tyrosin-Häufigkeit.

Diese Art von „Verwechslung“ ähnlicher Aminosäuren tritt beim Enzymdesign mit TransCent häufiger auf, wie anhand der Tabelle in Abbildung 4.16 zu erkennen ist. So kommt es relativ oft vor, dass beim Design für eine Position, die wildtypisch mit einem Valin besetzt ist, ein Isoleucin gewählt wird und umgekehrt (14 bzw. 17 Fälle). Gleiches gilt für das Paar Serin-Alanin (30 bzw. 19 Fälle). Darüber hinaus ersetzt TransCent wiederholt Threonin durch Serin (25 Fälle) und verwendet bevorzugt Aspartat statt Glutamat und Asparagin

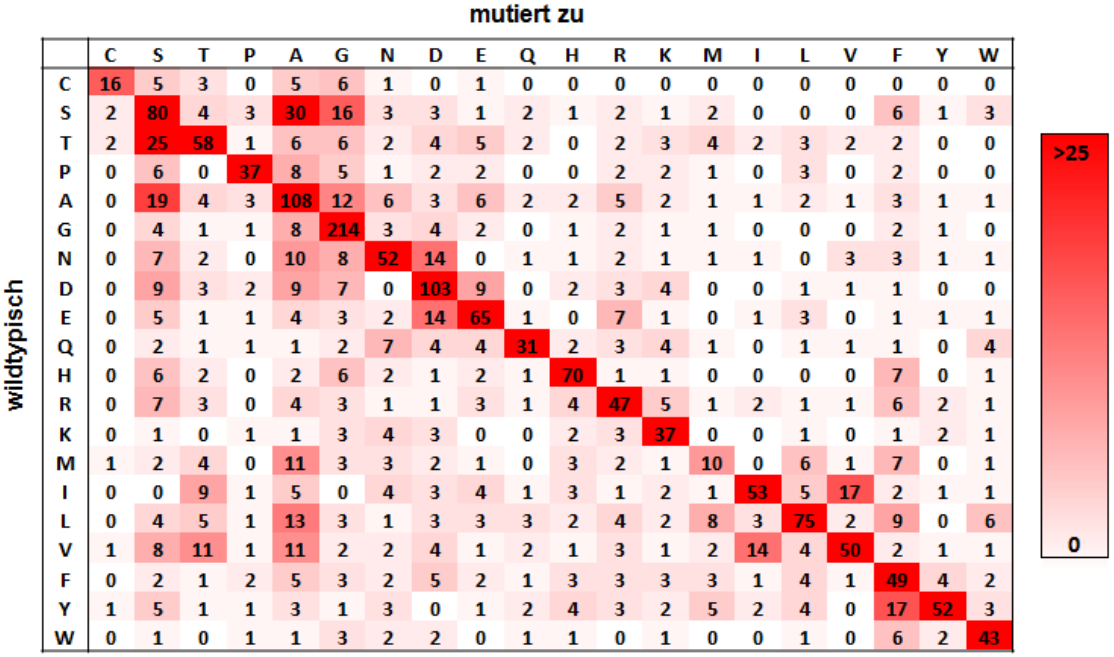


Abbildung 4.16: Ersetzungsmatrix der Mutationen von Wildtyp-Enzym zu Designmodell
Die Matrix enthält die absoluten Häufigkeiten der „Mutationen“, welche notwendig wären, um die wildtypischen Enzyme in die entsprechenden Designmodelle umzuwandeln, für insgesamt 2233 mutierbare Positionen. Auf der Hauptdiagonalen der Matrix liegen die korrekt besetzten Positionen. Die Matrix ist nicht symmetrisch, da bei der Betrachtung der Aminosäurewechsel die Richtung (Wildtyp zu Designmodell) berücksichtigt wird, z.B. wird ein wildtypisches Serin in 30 Fällen durch ein Alanin ersetzt, umgekehrt sind es nur 19 Fälle. Die Reihenfolge der Zeilen bzw. Spalten ist so gewählt, dass Gruppen ähnlicher Aminosäuren blockweise angeordnet sind.

(beides 14 Fälle). Die Aminosäurepaare aller aufgeführten Fälle haben einen Score ≥ 1 in der Blosum62-Matrix.

Die Zusammenfassung der Ergebnisse legt nahe, dass die Energiefunktion von TransCent prinzipiell dafür geeignet ist, die physikochemischen Anforderungen der meisten Proteinpositionen zu beschreiben, im Detail jedoch gelegentlich falsche Präferenzen setzt. Insgesamt bestätigt die Auswertung die Tendenz hin zur Wahl sehr kleiner Aminosäuren und die Schwierigkeiten bei der Rekapitulation von Cystein-Residuen.

Die Energiefunktion (vgl. 3.1) ist, neben der Optimierungsroutine (vgl. 3.1.1) und der Modellierungseinheit (vgl. 3.2.2), nur ein Teil des Designalgorithmus, welcher über die Qualität der erzeugten Enzymmodelle bestimmt. Dass die Optimierungsroutine in der Lage ist, bei hinreichend hoher Anzahl von Designoptimierungen zumindest in guter Näherung das globale Optimum zu finden, wurde bereits gezeigt (siehe 4.2.3.4). Ebenso wurden einige Schwächen der TransCent-Energiefunktion offengelegt. Als weitere potentielle Ursache für die unvollkommene Performanz von TransCent bei den Rekapitulationsrechnungen kommen aber auch mögliche Ungenauigkeiten bei der Modellierungseinheit in Frage.

4.2.3.6 Evaluation der Modellierungseinheit

Die Modellierungseinheit eines Designalgorithmus ist dafür verantwortlich, durch die Generierung einer Vielzahl von Modellen eine möglichst umfassende Abdeckung der Energielandschaft zu gewährleisten, um der Optimierungsroutine eine erfolgversprechende Suche nach dem globalen Optimum zu ermöglichen. Bei TransCent handelt es sich um einen rotamerbasierten Enzymdesignalgorithmus mit festem Proteinrückgrat, d.h. die Aufgabe der Modellierungseinheit beschränkt sich auf das Erstellen der Rotamere und Rotamervariationen (siehe 3.2.2) für die Aminosäureseitenketten. Da die tatsächlich beobachteten Seitenkettenkonformationen mehr oder minder stark von den idealisierten Rotameren abweichen können, ist auch diese Aufgabe nicht von trivialer Natur.

Um die Leistungsfähigkeit der Modellierungseinheit anhand von Rekapitulationsdesigns zu testen, bietet sich bei Rosetta und somit auch bei TransCent die Option `-use_input_sc` an. Diese bewirkt, dass für jede Proteinposition zusätzlich zu den berechneten Rotameren die Konformation der nativen Aminosäure aus der Kristallstruktur als zusätzliches „Rotamer“ beim Design zur Auswahl steht. Mit dieser Strategie können eventuelle Lücken bzw. Mängel bei der Modellierung der Rotamere aufgeklärt werden.

Die Analyse erfolgt auf der Basis von Rekapitulationsrechnungen des Testdatensatzes, wobei 100 Designoptimierungen für jedes der 53 Enzyme durchgeführt wurden. Die Modulgewichte entsprechen den Werten, die sich bei der Gridsuche (siehe 4.2.2) als optimal erwiesen haben. Zur Berechnung der Ergebnisse wird jeweils das Modell mit dem niedrigsten TransCent-Energiewert verwendet.

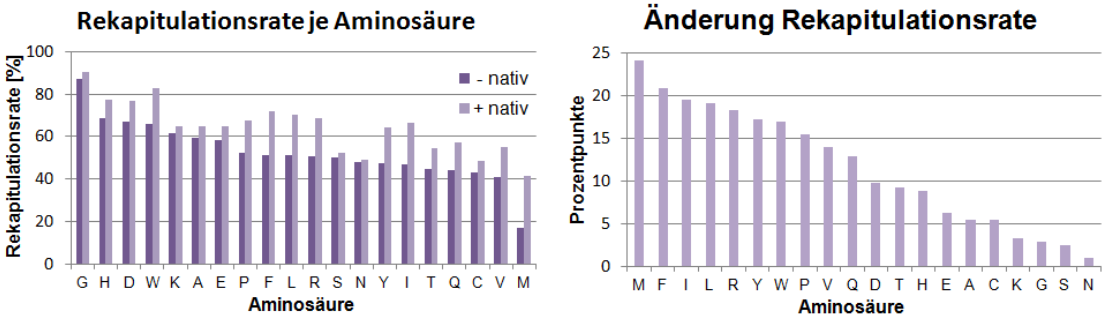


Abbildung 4.17: Vergleich der Rekapitulationsraten mit und ohne Verwendung der nativen Seitenkettenkonformation
Das linke Diagramm zeigt die Aminosäure-spezifischen Rekapitulationsraten für die 2233 mutierbaren Positionen des Rekapitulationsdatensatzes mit und ohne Verwendung der nativen Seitenkettenkonformation beim Design (Option `-use_input_sc`). In der rechten Abbildung ist dargestellt, um wie viele Prozentpunkte die Rekapitulationsrate durch Verwendung der nativen Seitenkette gesteigert werden kann.

Der Anteil erfolgreich designter Residuen steigt dabei von 56,0% auf 66,6% bei gleichzeitiger Erhöhung des Blosum-Scores um 0,7 Punkte auf einen Wert von 3,6. Die Vergleichswerte stammen dabei von Rekapitulationsdesigns ohne native Seitenketten (1000 Designoptimierungen, identische Modulgewichte). Wie man sieht, kann alleine durch eine genauere Abtastung des Lösungsraumes die Performanz von TransCent um über 10 Prozentpunkte gesteigert werden. Allerdings fällt die Änderung je nach Art der betrachteten Aminosäure unterschiedlich stark aus (siehe Abb. 4.17).

Erwartungsgemäß profitieren große Aminosäuren (R,F,W,Y) überproportional von der Hinzunahme des nativen Rotamers, da sich in diesen Fällen selbst kleine Abweichungen von den erforderlichen χ-Winkeln aufgrund des Lennard-Jones-Potentials extrem ungünstig auf die Energie auswirken können (siehe Abb. 4.18). Einen überdurchschnittlichen Gewinn erzielen auch die aliphatischen Aminosäuren (V,I,L) und das ebenfalls hydrophobe Methionin. Nachdem hydrophobe Aminosäuren bevorzugt im Inneren von Proteinen liegen, herrschen dort aufgrund der hohen Packungsdichte besonders hohe Ansprüche bezüglich der Exaktheit bei der Platzierung der Seitenketten. Dieser Umstand liefert eine plausible Erklärung für die beobachtete Steigerung der Rekapitulationsraten hydrophober Aminosäuren.

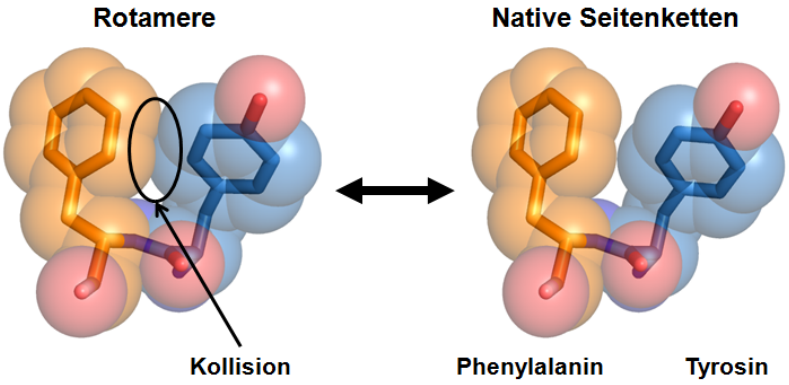


Abbildung 4.18: Abweichungen zwischen Rotameren und nativen Seitenketten
Die Abbildung zeigt zwei benachbarte Residuen, Phenylalanin (orange) und Tyrosin (blau), einmal mit den nativen Seitenketten, wie sie in der Kristallstruktur vorliegen (rechts) und einmal mit den Rotameren, welche den nativen Seitenketten am nächsten kommen. Die minimale Abweichung zwischen beiden Fällen, welche bei den Rotamer-Seitenketten zu einer Kollision zweier Atome führt, hat aufgrund der Form des Lennard-Jones-Potentials einen dramatischen Einfluss auf die energetische Bewertung des Rotamerpaares.

Interessanterweise kann selbst für Glycin und Alanin eine Erhöhung des Anteils korrekt modellierter Residuen erreicht werden. Da in diesen beiden Fällen die Berücksichtigung zusätzlicher „Rotamere“ nicht möglich ist, kommen aber nur indirekte Effekte als Ursachen in Frage. Um diesen auf den Grund zu gehen wurde für die Ergebnisse der Designs mit nativer Seitenkette analog zu Tabelle in Abbildung 4.16 die „Ersetzungsmatrix“ berechnet (siehe Abb. 4.19).

Der Vergleich der beiden Tabellen zeigt, dass die Mehrheit der nun korrekt modellierten Alanin-Residuen ursprünglich zu Glycin mutiert wurde, möglicherweise aufgrund sterischer Zwänge, welche von benachbarten Residuen ausgeübt wurden.

Insgesamt zeigt sich, dass mit der Option `-use_input_sc` weit weniger sehr kleine Aminosäuren (Glycin, Alanin, Serin) eingeführt werden (662 statt 749). Ihr Anteil an der Gesamtverteilung sinkt damit von 34% auf 30% (vgl. Abb. 4.20). Darüber hinaus dokumentiert der Vergleich der beiden Tabellen, dass die Anzahl der „Verwechslungen“ ähnlicher Aminosäuren (vgl. 4.2.3.5) abnimmt. Dies gilt allerdings nur, was die Ersetzung einer

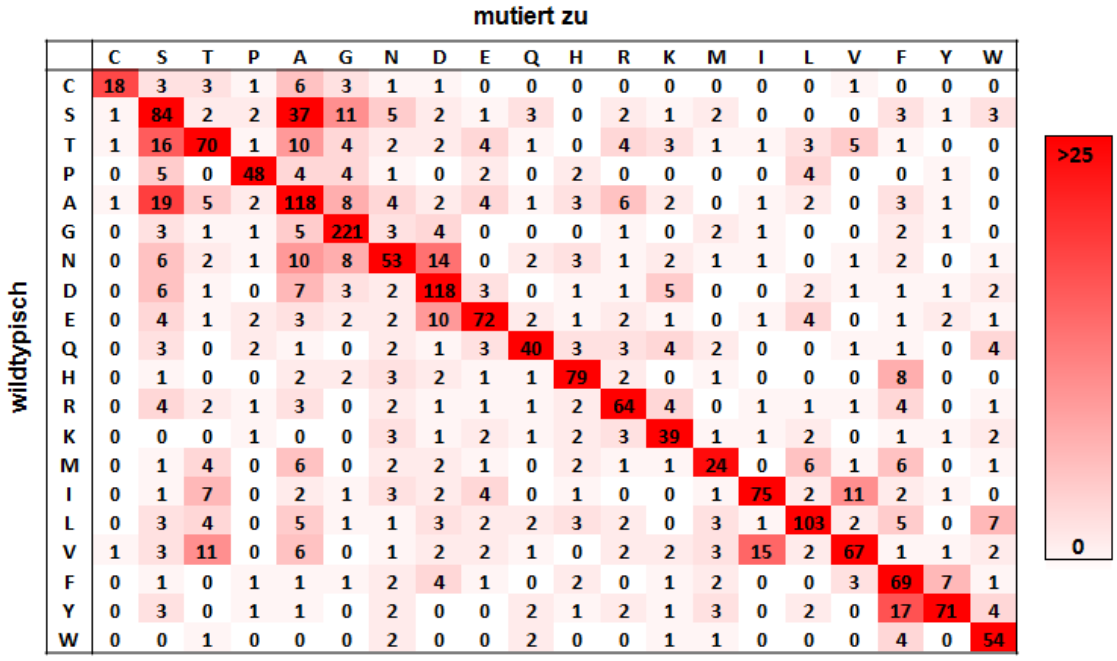


Abbildung 4.19: Ersetzungsmatrix der Mutationen von Wildtyp-Enzym zu Designmodell unter Verwendung der nativen Seitenkettenkonformation
Die Matrix enthält die absoluten Häufigkeiten der „Mutationen“, welche notwendig wären, um die wildtypischen Enzyme in die entsprechenden Designmodelle umzuwandeln, für insgesamt 2233 mutierbare Positionen. Auf der Hauptdiagonalen der Matrix liegen die korrekt besetzten Positionen. Die Matrix ist nicht symmetrisch, da bei der Betrachtung der Aminosäurewechsel die Richtung (Wildtyp zu Designmodell) berücksichtigt wird. Die Reihenfolge der Zeilen bzw. Spalten ist so gewählt, dass Gruppen ähnlicher Aminosäuren blockweise angeordnet sind.

Aminosäure durch eine kleinere mit ähnlichen physikochemischen Eigenschaften betrifft (-9 Mutationen Threonin zu Serin, -6 Mutationen Isoleucin zu Valin). Beim umgekehrten Fall (klein zu groß) wird diese Tendenz nicht beobachtet. Außerdem nimmt, entgegen dem allgemeinen Trend, die Anzahl der Mutationen von Serin zu Alanin zu.

Zusammenfassend kann festgestellt werden, dass in einigen Fällen eine perfekte Rekapitulation auch deswegen nicht gelingt, weil mit den von der Modellierungseinheit erstellten Rotameren die Energielandschaft nicht mit der erforderlichen Genauigkeit abgetastet wird. Dies kommt auch dadurch zum Ausdruck, dass der TransCent-Energiewert der besten Designmodelle unter Verwendung der nativen Seitenketten durchschnittlich um 353 Energieeinheiten niedriger ist als ohne.

Die Evaluation der Modellierungseinheit hat gezeigt, dass mittels genauerer Abtastung der Energielandschaft die Performanz von TransCent verbessert werden kann. Allerdings ist selbst bei Verwendung der nativen Seitenketten als zusätzliche Rotamere keine perfekte Rekapitulation möglich, wie die enzymspezifischen Rekapitulationsraten zwischen 47% und 86% belegen. Zum einen liegt dies an den Defiziten der Energiefunktion, kann aber auch auf Fehler in den Kristallstrukturen zurückzuführen sein [157, 158].

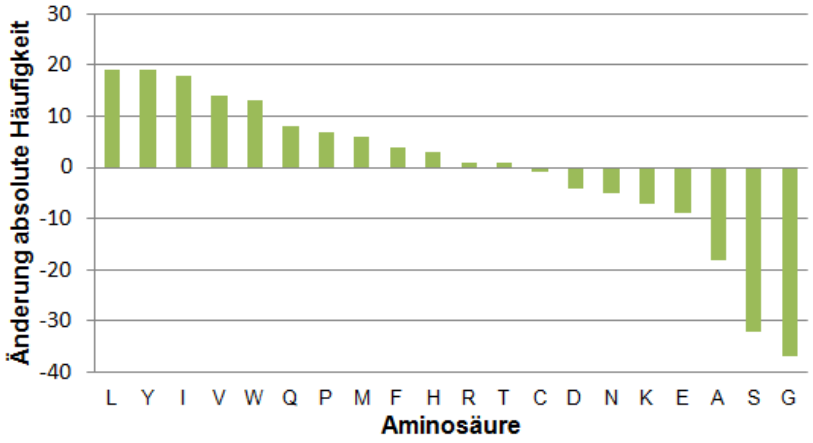


Abbildung 4.20: Änderung der absoluten Aminosäurehäufigkeiten durch Hinzunahme der nativen Seitenkettenkonformationen beim Design
Die Abbildung zeigt die Differenzen der absoluten Aminosäurehäufigkeiten für die besten Designmodelle des Rekapitulationsdatensatzes, welche mit bzw. ohne Verwendung der nativen Seitenkettenkonformationen erzeugt wurden.

Die Option `-use_input_sc` ist zwar auch beim Design von Funktionsübertragungen verfügbar, besitzt dabei aber eine andere Qualität, denn die wildtypische Aminosäure ist für eine Proteinposition dann nicht mehr unbedingt die beste Wahl. Sie bewirkt aber weiterhin, dass für jede mutierbare Position ein zusätzliches Rotamer zur Verfügung steht. Ferner kann die Verwendung der nativen Seitenkette zu einer leichten Verzerrung der energetischen Verhältnisse führen, so dass die wildtypische Aminosäure etwas bevorzugt wird. Eine mögliche Folge ist, dass Mutationen, welche energetisch nur minimal besser als die wildtypische Aminosäure abschneiden würden, nicht zustande kommen. Diesen Effekt kann man sich in ähnlicher Form beim Design von Funktionsübertragung zunutze machen, wie im nachfolgenden Abschnitt erläutert wird.

4.2.3.7 Bewertung einzelner Mutationen

Die bei einer Funktionsübertragung vorgeschlagenen Mutationen im Zielprotein werden von TransCent nicht alle mit dem gleichen Energiegewinn bewertet. Auch bei natürlich vorkommenden Proteinen ist der Effekt einer Mutation z.B. auf die Proteinstabilität nicht an allen Positionen gleich. In völliger Analogie dazu variiert auch der Gewinn an TransCent-Energie, welcher durch den Austausch einer Aminosäure an einer Position i erzielt werden kann. Dieser wird wie folgt berechnet:

$$\Delta E\left(A_i^{WT} \rightarrow A_i^D\right) = E\left(A_i^D\right) - E\left(A_i^{WT}\right) \tag{4.1}$$

Dabei ist $E(A_i^D)$ die Energie des Designmodells, wie es von der Optimierungsroutine erzeugt worden ist. Der Energiewert $E(A_i^{WT})$ des Modells ohne die Mutation an Position i wird mit Hilfe einer eingeschränkten Designoptimierung berechnet. Wie beim eigentlichen Design wird dabei die Rotamerkombination mit der niedrigsten Energie gesucht (vgl. 3.1.1), mit dem Unterschied, dass es keine mutierbaren sondern lediglich rotierbare

Positionen gibt (siehe 4.2.1). Die Art der Aminosäure ist mit einer Ausnahme auf die des Designmodells festgelegt. Nur an Positionen i wird A_i^D durch die wildtypische Aminosäure A_i^{WT} ersetzt.

Um die (energetische) Bedeutung einer Mutation zu erfassen, wird also die Energiedifferenz zwischen Designmodell und der Variante ohne die entsprechende Mutation betrachtet. Als Referenz dient somit das Designergebnis und nicht das wildtypische Enzym, was den entscheidenden Vorteil bietet, dass dadurch auch Abhängigkeiten zwischen mehreren Mutationen berücksichtigt werden können. Ist die Differenz positiv, so bevorzugt TransCent die Aminosäure des Wildtyps. Je negativer der Wert ausfällt, umso mehr Bedeutung misst das Programm der Einführung einer Mutation bei.

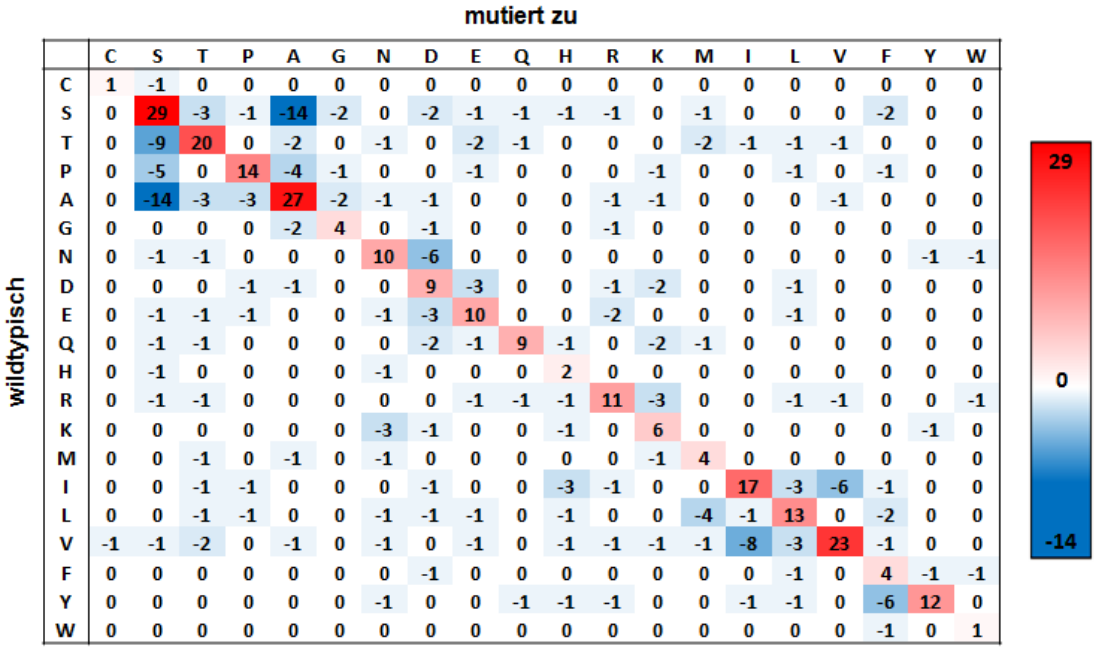
Die Berechnung der Energiedifferenzen für alle mutierten Positionen ist damit ein möglicher Teil der Nachbereitung eines Designprozesses. Tritt dabei ein positiver Wert auf, so sollte diese Mutation zurückgenommen werden. In Anlehnung an den Begriff der sinnverändernden Punktmutation, welcher den Austausch einer einzelnen Aminosäure bezeichnet, wird die Rücknahme eines Austauschs Punktrückmutation genannt.

Punktrückmutationen können auch bei leicht negativen Energiedifferenzen sinnvoll sein. Wird für den Wechsel einer Aminosäure ein kleiner Energiegewinn verzeichnet, so bedeutet dies, dass TransCent dadurch eine minimale Verbesserung erwartet, sei es hinsichtlich Proteinstabilität, Ligandenbindung oder Protonierungswahrscheinlichkeit essentieller Residuen. Ob dies auf einem tatsächlich zu erwartenden Effekt beruht oder nur aufgrund von Ungenauigkeiten bei der Energiefunktion oder der Modellierungseinheit zustande kommt (vgl. 4.2.3.5 und 4.2.3.6), ist ohne experimentelle Überprüfung nicht festzustellen. Andererseits belegt das Vorkommen der wildtypischen Aminosäure, dass sie zu einem stabilen und funktionalen Enzym beiträgt. Folglich erscheint es beim Design von Funktionsübertragungen vernünftig, Punktrückmutationen an Positionen mit geringem Energiegewinn zuzulassen.

Zur *in silico* Überprüfung dieser Strategie wurden die Energiedifferenzen aller Mutationen berechnet, die sich beim Rekapitulationsdesign der Enzyme des Testdatensatzes (siehe 3.10.3) ergeben. Die Basis der Rechnungen bildete jeweils das Modell mit der niedrigsten Energie, welches sich nach 1000 Designoptimierungen ergab. Aufgrund der nachfolgenden Ergebnisse wurde ein Schwellwert von -1 für Punktrückmutationen eingeführt, d.h. Mutationen mit einem Energiegewinn $\Delta E\left(A_i^{WT} \rightarrow A_i^D\right) \geq -1$ wurden zurückgenommen.

Zunächst kann festgehalten werden, dass für zwölf der 983 Mutationen eine positive Energiedifferenz ermittelt wurde. Dies bedeutet, dass in diesen Fällen die *Simulated Annealing* Routine es nicht geschafft hat, das (lokale) Energieminimum zu erreichen. Für weitere 214 Mutationen liegt der Energieunterschied zwischen 0 und dem Schwellwert, so dass insgesamt 226 Punktrückmutationen in Frage kommen. Zählt man nun diese Positionen zu den erfolgreich designten Residuen hinzu, so erhielte man eine Steigerung der Rekapitulationsrate um 10,1 Prozentpunkte auf 66,1%, was ziemlich genau dem Niveau entspricht, welches auch unter Verwendung der nativen Seitenketten als zusätzliche Rotamere (siehe 4.2.3.6) erreicht wird.

Eine detaillierte Untersuchung der Punktrückmutationen zeigt, dass sich die betroffenen Aminosäureaustausche kaum von denen bei der Verwendung der nativen Seitenkettenkonformation unterscheidet. Zu einem Großteil sind wiederum Austausche zwischen sehr ähnlichen Aminosäuren betroffen, vor allem jene, bei denen TransCent zu „Verwechslungen“ neigt (vgl. 4.2.3.5), wie anhand der Tabelle in Abbildung 4.21 nachvollzogen werden



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	29	-3	-1	-14	-2	0	-2	-1	-1	-1	-1	0	-1	0	0	0	-2	0	0
T	0	-9	20	0	-2	0	-1	0	-2	-1	0	0	0	-2	-1	-1	-1	0	0	0
P	0	-5	0	14	-4	-1	0	0	-1	0	0	0	-1	0	0	-1	0	-1	0	0
A	0	-14	-3	-3	27	-2	-1	-1	0	0	0	-1	-1	0	0	0	-1	0	0	0
G	0	0	0	0	-2	4	0	-1	0	0	0	-1	0	0	0	0	0	0	0	0
N	0	-1	-1	0	0	0	10	-6	0	0	0	0	0	0	0	0	0	0	-1	-1
D	0	0	0	-1	-1	0	0	9	-3	0	0	-1	-2	0	0	-1	0	0	0	0
E	0	-1	-1	-1	0	0	-1	-3	10	0	0	-2	0	0	0	-1	0	0	0	0
Q	0	-1	-1	0	0	0	0	-2	-1	9	-1	0	-2	-1	0	0	0	0	0	0
H	0	-1	0	0	0	0	-1	0	0	0	2	0	0	0	0	0	0	0	0	0
R	0	-1	-1	0	0	0	0	0	-1	-1	-1	11	-3	0	0	-1	-1	0	0	-1
K	0	0	0	0	0	0	-3	-1	0	0	-1	0	6	0	0	0	0	0	-1	0
M	0	0	-1	0	-1	0	-1	0	0	0	0	0	-1	4	0	0	0	0	0	0
I	0	0	-1	-1	0	0	0	-1	0	0	-3	-1	0	0	17	-3	-6	-1	0	0
L	0	0	-1	-1	0	0	-1	-1	-1	0	-1	0	0	-4	-1	13	0	-2	0	0
V	-1	-1	-2	0	-1	0	-1	0	-1	0	-1	-1	-1	-1	-8	-3	23	-1	0	0
F	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	-1	0	4	-1	-1
Y	0	0	0	0	0	0	-1	0	0	-1	-1	-1	0	0	-1	-1	0	-6	12	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	1

Abbildung 4.21: Verteilungsmatrix der Punktrückmutationen
Die Matrix enthält die absoluten Häufigkeiten der Punktrückmutationen (blau) und der dadurch zusätzlich korrekt besetzten Proteinpositionen (rot) für insgesamt 226 Punktrückmutationen. Die Summe der Werte einer Zeile ist immer gleich 0, da sich für jede Punktrückmutation (negativer Wert) der entsprechende Eintrag in der Hauptdiagonale um 1 erhöht.Die Reihenfolge der Zeilen bzw. Spalten ist so gewählt, dass Gruppen ähnlicher Aminosäuren blockweise angeordnet sind.

kann. Insbesondere fällt auf, dass mehr als drei Viertel der Mutationen von Alanin zu Serin mit einem Energiegewinn von weniger als einer Energieeinheit verbunden sind (14 von 19 Fällen). Gleiches gilt für eine Vielzahl von Austauschen in entgegengesetzter Richtung (14 von 30 Fällen) und zahlreiche Mutationen des Aminosäurepaares Valin-Isoleucin (8 bzw. 6 Fälle).

Wird der triviale Fall der identischen Ersetzung ausgeschlossen, so sind bei 20 Aminosäurearten insgesamt $20 \cdot 19 = 380$ Typen von Mutationen denkbar. Die Anwendung der Punktrückmutationen führt dazu, dass 25 der ursprünglich 293 beobachteten Mutationstypen verschwinden. Daraus folgt, dass diese mit jeweils ein bis drei Fällen vertretenen Austausche durchgehend mit Energiegewinnen unterhalb der gesetzten Schwelle bewertet wurden.

Zusammenfassend kann festgestellt werden, dass die Analyse der Punktrückmutationen die bisherigen Ergebnisse hinsichtlich der Genauigkeit der Energiefunktion bestätigt und dass die Anwendung dieses Ansatzes hilft, die Anzahl der Mutationen bei Funktionsübertragungen zu reduzieren.

Schließt man die beiden Punktrückmutationen mit ein, so erreicht TransCent beim Design der Ribonuklease A aus *Bos taurus* (PDB-Code: 1o0h) eine Rekapitulationsrate von 88%. Bei 25 mutierbaren Positionen bedeutet dies, dass sich Designmodell und wildtypisches Enzym nur noch an drei Positionen unterscheiden. Somit ist die Wahrscheinlichkeit relativ

hoch, dass es sich bei dem Designmodell um ein stabiles Enzym mit der gewünschten Aktivität handelt, obwohl Designergebnis und Wildtyp nicht zu 100% identisch sind. Wie die Natur am Beispiel homologer Enzyme gezeigt hat, gibt es oft mehrere Möglichkeiten, um ein „Designziel“ zu erreichen.

Aufgrund der in der Natur beobachteten Bandbreite unterschiedlicher Sequenzen, die alle dieselbe Enzymfunktion realisieren, lag es nahe, ein Maß zum Vergleich von Sequenzmengen zu entwickeln.

4.2.3.8 PSSM-Auswertung

Das Enzym, das einem Rekapitulationsdesign unterworfen wird, wird nachfolgend als Referenzenzym bezeichnet. Um herauszufinden, inwieweit die von TransCent erzeugten Modelle natürlich vorkommenden Proteinen mit der betrachteten enzymatischen Aktivität entsprechen, wird wie beim Vergleich mit dem Referenzenzym die Besetzung der aktiven Zentren untersucht. Grundlage für die Auswertung sind die multiplen Sequenzalignments homologer Proteine der 53 Enzyme des Rekapitulationsdatensatzes (vgl. 3.10.3), welche jeweils auch zum Erstellen der Strukturbibliothek verwendet wurden (siehe 3.4.1). Da i.A. nur für das Referenzenzym die Tertiärstruktur bekannt ist, wird für die folgende Analyse die vereinfachende Annahme gemacht, dass die funktionsbestimmenden Residuen bei allen Proteinen identisch sind. Der Vergleich kann sich somit auf die Spalten beschränken, welche den mutierbaren Positionen des Designs entsprechen.

Berechnet man zunächst die minimale Sequenzidentität zwischen einer Sequenz aus dem MSA und dem Referenzenzym, so erhält man einen Wert von durchschnittlich 55,9%, was fast exakt mit der mittleren Rekapitulationsrate von TransCent (56,0%) übereinstimmt. In 23 von 53 Fällen ist die erzielte Rekapitulationsrate sogar höher als der minimale Sequenzidentitätswert. Im Mittel stimmt die Besetzung der mutierbaren Positionen von MSA- und Referenzsequenz zu 75,8% überein.

Bei der gleichen Analyse auf Grundlage des Blosum-Scores, welcher auch die nicht identischen Residuen bewertet, erreichen nur noch 18 Designmodelle einen Score der im Wertebereich der Sequenzen des MSAs liegt. Der Durchschnitt des mittleren Blosum-Scores der MSA-Sequenzen ist mit 4,2 relativ nahe am Maximum von 5,5, das man für den Vergleich der Referenzenzyme mit sich selbst erhält. Im Vergleich dazu erhält man für die TransCent-Ergebnisse lediglich einen Wert von 2,9. Die Zusammensetzung der aktiven Zentren homologer Enzyme ähnelt sich also in hohem Maße. Folglich müssen auch für die nicht identisch besetzten Positionen Beschränkungen existieren, was die Wahl der Aminosäure betrifft.

Um diese mit zu erfassen eignet sich ein weiteres Maß für den Sequenzvergleich, der PSSM-Score (vgl. 3.10.2.3). Die zugrunde liegende Scoringmatrix beschreibt für jede Position welche Aminosäuren dort mit welcher Präferenz vorkommen. Der PSSM-Score ist somit das Maß, mit dem am besten bewertet werden kann, inwieweit ein von TransCent generiertes Modell einem natürlich vorkommenden Enzym entspricht.

Er berücksichtigt dabei auch implizit die Variabilität innerhalb der homologen Sequenzen, denn die (bekannte) Anzahl möglicher Varianten kann sich je nach Enzymfunktion zum Teil deutlich unterscheiden. Beispielsweise liegt die mittlere Sequenzidentität für die 275 homologen Enzyme der Methionyl-tRNA Synthetase aus *Escherichia coli*

(PDB-Code: 1pfu) bei 98,5%, wenn man nur die 27 mutierbaren Positionen des aktiven Zentrums betrachtet. Am anderen Ende der Skala rangiert die Hypoxanthin-Guanin-Phosphoribosyltransferase aus *Toxoplasma gondii* (PDB-Code: 1fsg), bei der die 51 mutierbaren Residuen der 81 Homologen im Durchschnitt nur zu 55,7% identisch sind.

Beim Vergleich der PSSM-Scores wird nur ein einziges mit TransCent berechnetes Enzymmodell besser als eine dazu homologe Sequenz bewertet. Im Mittel ist der Wert für die Designmodelle bei 0,25 und liegt somit beträchtlich unter der Schwelle von 1,51, dem Durchschnittswert für die jeweils „untypischsten“ Vertreter einer Enzymfunktion (vgl. Abb. 4.22). Dieser Wert kann als Grenze angesehen werden, ab der es sich bei einer Modellsequenz mit hoher Wahrscheinlichkeit um ein aktives Enzym handelt. Der mittlere PSSM-Wert der 53 Referenzenzyme liegt bei 2,52 und somit nur knapp unter dem Maximum von 2,70, das man erhält, wenn man in allen Fällen die Sequenz mit dem höchsten PSSM-Score betrachtet. Dieses Ergebnis bestätigt die Annahme, dass der PSSM-Score ein geeignetes Maß zur Beschreibung einer Enzymfunktion ist.

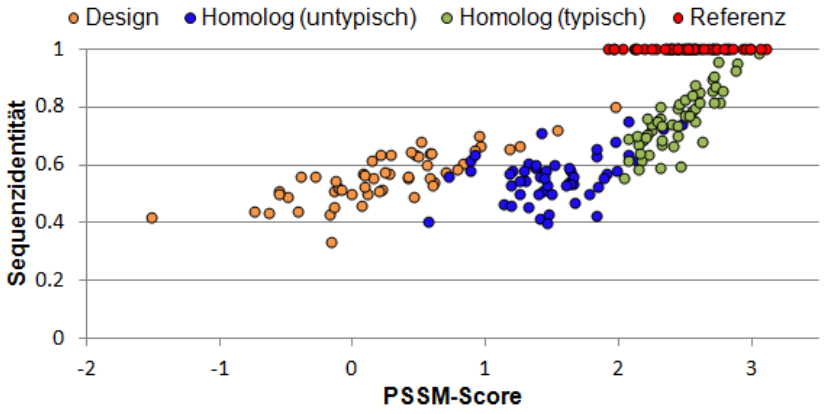


Abbildung 4.22: PSSM-Score-Verteilung für modellierte und natürliche Sequenzen
Die Abbildung zeigt die Verteilung der Wertepaare von PSSM-Score und Sequenzidentität für jeweils vier Repräsentanten der 53 Enzyme des Rekapitulationsdatensatzes. Dies sind zum einen die Sequenz des besten Designmodells, des Referenzenzyms, das die Grundlage des Rekapitulationsdesigns bildet und die Sequenz des „untypischten“ homologen Enzyms, welches den kleinsten PSSM-Score aufweist. Hinzu kommt eine „typische“ Sequenz mit dem mittleren PSSM-Score und der durchschnittlichen Sequenzidentität aller Enzyme des MSAs. Als Wert für die Sequenzidentität wird jeweils der Anteil identischer Residuen mit dem Referenzenzym berechnet, wobei nur die mutierbaren Positionen berücksichtigt werden.

Zusammengefasst kann festgestellt werden, dass die Designergebnisse von TransCent zwar, was die Anzahl identischer Residuen betrifft, durchaus im Bereich wildtypischer Proteine liegen, bei näherer Betrachtung aber doch merklich von diesen abweichen.

Diese Beobachtung wirft die Frage auf, ob TransCent allgemein in der Lage ist, die wichtigen Wechselwirkungen zu modellieren und nur bei der Ausgestaltung der Details Schwierigkeiten hat. Um dieser Frage nachzugehen wurde die Konserviertheit der von TransCent korrekt modellierten Positionen untersucht, da sich die Wichtigkeit eines Residuums meist

dadurch ausdrückt, wie stark es konserviert ist. Vor allem bei strikt konservierten Aminosäuren kann davon ausgegangen werden, dass sie von essentieller Bedeutung für das Funktionieren eines Enzyms sind. Konserviertheit ist hier definiert als die relative Häufigkeit $f(as_i, k)$ der Aminosäure as_i an Position k im Referenzenzym in der korrespondierenden Spalte des MSAs.

Wie der Verlauf der Kurven in Abbildung 4.23 zeigt, nimmt die Rekapitulationsrate kontinuierlich mit fallender Konserviertheit ab. Während bei den strikt konservierten Residuen zwei Drittel korrekt besetzt werden, ist es bei den nicht konservierten Positionen ($\leq 10\%$) nur noch jede vierte.

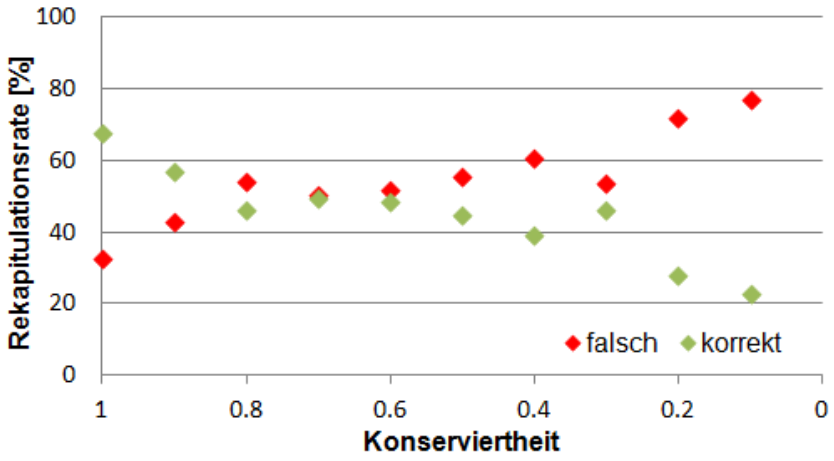


Abbildung 4.23: Abhängigkeit der Rekapitulationsrate von der Konserviertheit eines Residuums
Die beiden Kurven zeigen den Anteil der korrekt bzw. falsch modellierten Positionen in Abhängigkeit von deren Konserviertheit. Unter Konserviertheit ist in diesem Zusammenhang definiert als die relative Häufigkeit der Aminosäure aus dem Referenzenzym in der entsprechenden Spalte des MSAs homologer Sequenzen.

Die Ergebnisse der Analyse legen nahe, dass die TransCent-Energiefunktion zwar im Detail stellenweise an Präzision vermissen lässt, die wichtigen Wechselwirkungen aber relativ zuverlässig abbildet.

Anhand der bisher präsentierten Ergebnisse konnte gezeigt werden, dass die Performanz von TransCent durch die Überarbeitung bzw. den Austausch der Module gegenüber der ersten Version verbessert wurde. Zudem wurde der wesentliche Beitrag des PROPKA-Moduls zur Optimierung der Protonierungszustände nachgewiesen. Die Resultate der Rekapitulationsrechnungen mit den neu bestimmten Modul-Gewichten weisen auf Verbesserungsmöglichkeiten sowohl bei der Energiefunktion als auch bei der Modellierungseinheit hin. Eine Strategie, um diese Defizite auszugleichen, wurde mit der Einführung von Punktrückmutationen vorgestellt. Insgesamt ergibt die Auswertung der Rekapitulationen, dass im Mittel mehr als die Hälfte aller mutierbaren Positionen korrekt besetzt wird. Insbesondere die für die Katalyse unbedingt notwendigen Seitenketten sind i. A. in den mit TransCent berechneten Designmodellen enthalten.

4.2.4 Geschwindigkeitsgewinn durch Parallelisierung

Die erweiterte Funktionalität von TransCent erfordert auch zusätzliche Rechnerleistung. Zum einen müssen die Energietabellen des DSX- und Fingerprint-Moduls für jede Ligandposition unabhängig berechnet und im Arbeitsspeicher vorgehalten werden. Viel gravierender ist aber, dass ein Wechsel der Ligandposition bei der Designoptimierung inhärent mit den aufwändig zu berechnenden Mehr-Körper-Energien verbunden ist und somit im Vergleich zu einem normalen Rotamertausch überproportional stark ins Gewicht fällt.

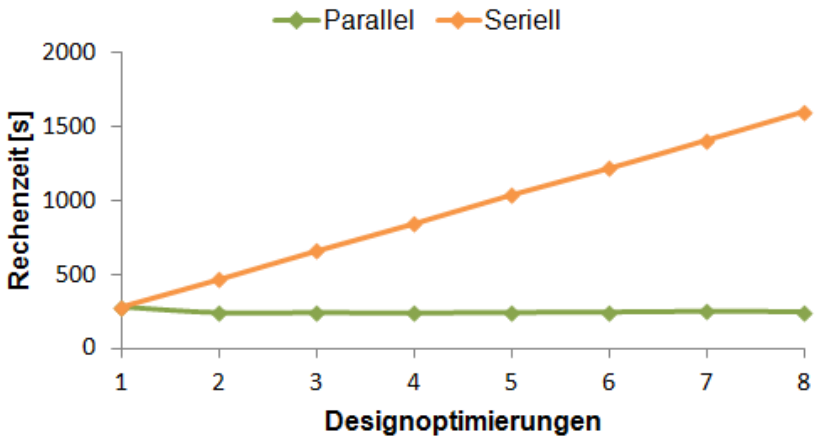


Abbildung 4.24: Vergleich der Rechenzeit bei serieller und paralleler Ausführung
Die Kurven zeigen die benötigte Rechenzeit für 1 bis 8 Designoptimierungen zur Rekapitulation der Ribonuclease MC1 aus *Momordica charantia* (PDB-Code: 1ucd). Im seriellen Fall muss ein Thread die Optimierungen nacheinander ausführen. Bei der parallelen Berechnung existieren jeweils acht Threads, um die Aufgaben parallel abzuarbeiten. Die Ergebnisse sind Mittelwerte für 10 Wiederholungen.

Durch die parallele Abarbeitung der Designoptimierungen kann die benötigte Rechenzeit für den Designprozess je nach Anzahl der vorhandenen Prozessorkerne reduziert werden. Eine tatsächliche Beschleunigung kann aber nur erreicht werden, wenn mehr als ein Designmodell erstellt werden soll, was i.A. auch der Fall ist. Wie Abbildung 4.24 zu entnehmen ist, hält sich der durch die Parallelisierung verursachte Overhead in Grenzen. Um völlig identische Bedingungen zu garantieren wurden in allen Fällen jeweils acht Prozessorkerne für die Ausführung des Programms reserviert.

4.3 Flexible Ligandpositionierung

Die bisherige Version von TransCent ging von einer fest vorgegebenen Position des Liganden aus. Beim Rekapitulationsdesign ist ein solches Vorgehen keine Einschränkung, da die vorgegebene Ligandposition bereits optimal ist. Die Situation ändert sich aber drastisch beim Design einer Funktionsübertragung. In diesem Fall kann *a priori* nicht ermittelt werden, wie der Ligand positioniert werden muss, um das beste Designergebnis zu erhalten. Auch normale Docking-Algorithmen versagen hier, da das aktive Zentrum, in welchem der

Ligand platziert werden soll, vor Abschluss des Designprozesses noch gar nicht eindeutig definiert ist. Es war daher ein zentrales Thema der vorliegenden Arbeit, diese massive Einschränkung aufzuheben.

Zur Lösung dieses Problems hat Friedemann Paulini im Rahmen seiner Diplomarbeit das Programm TransLig entwickelt [123], welches anhand vorgegebener Abstandskriterien im designierten aktiven Zentrum des Zielproteins geeignete Ligandpositionen sucht (siehe 3.8). Diese gestatten dem Designalgorithmus die größtmögliche Zahl von Freiheitsgraden bei der Wahl der Aminosäureseitenketten und ermöglichen so das Design aller wichtigen Protein-Ligand-Wechselwirkungen (vgl. Abb. 4.25).

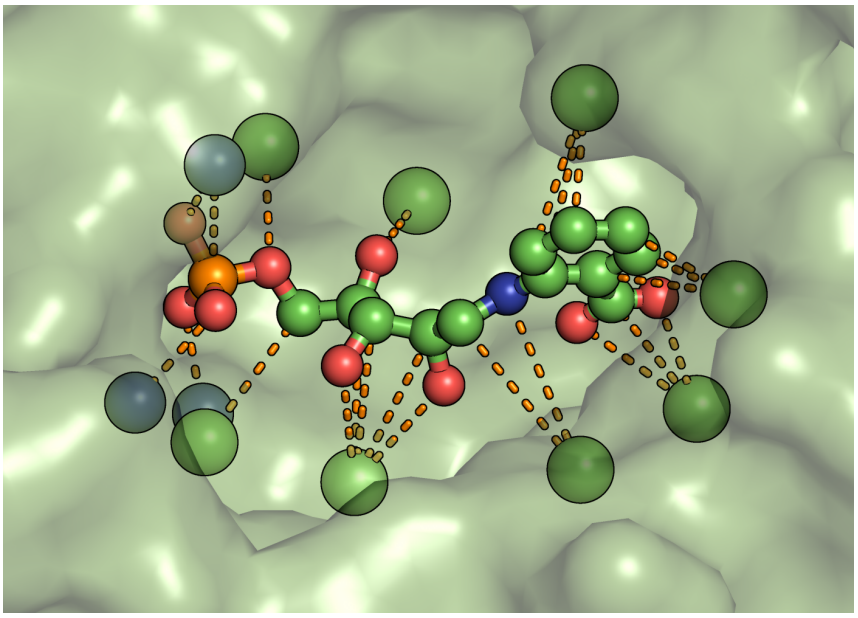


Abbildung 4.25: Suche nach geeigneten Ligandpositionen mit TransLig
Der Suchalgorithmus startet mit einem zentral im „leergeräumten“ aktiven Zentrum platzierten Liganden (*Ball&Stick*-Darstellung). Gemäß dem Kräftemodell wird dessen Position iterativ optimiert, bis die Abstände (gestrichelte Linien) zu den Proteinatomen (blaue und grüne Kugeln) möglichst gut den Vorgaben entsprechen.

4.3.1 Evaluation des TransLig-Algorithmus

Die Performanz von TransLig bei der Suche nach passenden Ligandpositionen wurde bereits in [123] mit Hilfe von Rekapitulationsrechnungen untersucht. Wie beim Enzymdesign wird dabei angenommen, dass die in der Kristallstruktur beobachtete Pose der optimalen Lösung entspricht. Folglich wird als Kriterium für die Qualität der TransLig-Vorhersagen der RMSD-Wert zwischen vorhergesagter und nativer Ligandgeometrie verwendet. Die TransLig-interne Bewertung der Ligandpositionen erfolgt anhand des cRMSD-Wertes, welcher die Abweichung der erreichten von den geforderten Atomabständen misst (vgl. 3.8.4).

Zur Bestätigung der Resultate aus [123] wurden TransLig-Ergebnisse für die Proteine des TransCent-Testdatensatzes (siehe 3.10.3) analysiert. Diese Ligandpositionen bilden auch die Grundlage für die nachfolgende Evaluation der TransCent-TransLig-Kombination.

Für jedes Enzym wurden zunächst automatisiert aus den Strukturen der Protein-Ligand-Komplexe die einzuhaltenden Abstandskriterien abgeleitet. Es erfolgte dabei keine gesonderte Gewichtung einzelner Ligandatome und auch auf die Spezifikation individueller Interaktionspartner wurde verzichtet (vgl. 3.8.2.2 bzw. 3.8.3.1). Für die Berechnung der Ligandpositionen wurde als Kraftfunktion die Option „quadratisch“ gewählt (siehe 3.8.2.2), welche bei den Testrechnungen in [123] im Mittel am besten abgeschnitten hat. Die Anzahl der Startpunkte, Variationen je Startpunkt und Iterationen pro Startpunktvariation wurden auf 3, 30 bzw. 50 festgelegt (vgl. 3.8.2.3), so dass insgesamt 4500 Iterationsschritte für jedes Protein berechnet wurden. Außerdem wurde der Suchraum für die TransLig-Optimierung auf das „leergeräumte“ aktiven Zentrum beschränkt, d.h. alle Seitenketten in einem Radius von 5 Å um die native Ligandposition wurden auf das C_β-Atom reduziert und die verbliebenen Atome als Partner markiert.

Für 52 der 53 Enzyme konnte TransLig durchschnittlich 744 Ligandpositionen mit cRMSD-Werten zwischen 0,02 Å und 2,40 Å ermitteln. Die Platzierung von Adenosindiphosphat im aktiven Zentrum der Ribonuklease A aus *Bos taurus* (PDB-Code: 1o0h) schlug allerdings fehl, da die Berechnung des Voronoi-Diagramms nicht möglich war. Wie in Abbildung 4.26 zu sehen ist, existiert bei diesem Enzym keine ausgeprägte Ligandenbindetasche, was allerdings Grundvoraussetzung für die Anwendung von TransLig ist.

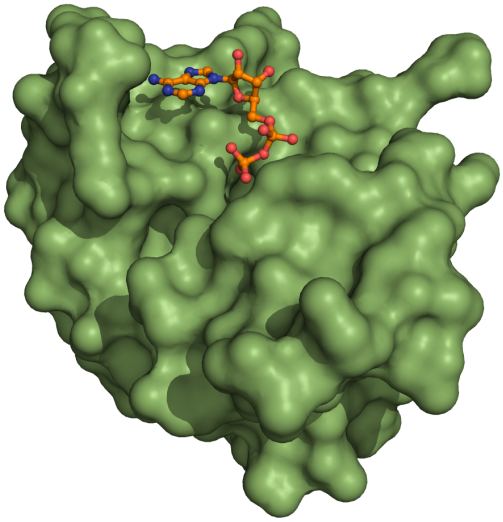


Abbildung 4.26: Ligandenbindestelle der Ribonuklease A aus *Bos taurus* (PDB-Code: 1o0h)
Der Ligand (orange, *Ball & Stick*-Darstellung) bindet auf der Oberfläche des Enzyms (grün, *Surface*-Darstellung). Mit Ausnahme der C_β-Atome wurden die Seitenkettenatome des aktiven Zentrums entfernt analog zur Positionssuche mit TransLig.

Betrachtet man die jeweils niedrigsten erreichten cRMSD-Werte, so liegt deren arithmetisches Mittel bei 0,18 Å, im besten Fall wird sogar ein Wert von 0,02 Å erzielt. Dies belegt, dass TransLig in Lage ist, Ligandpositionen zu finden, welche die vorgegebenen Abstandsbedingungen fast perfekt erfüllen. Der mittlere RMSD-Wert dieser Positionen ist 0,44 Å, was bedeutet, dass die besten TransLig-Ergebnisse sehr gut mit den nativen Ligandgeometrien übereinstimmen. Nur in einem von 52 Fällen liegt der RMSD-Wert der

besten Position über der Schwelle von 2 Å, die bei Docking-Verfahren üblicherweise als Grenze für „nativ-ähnliche“ Positionen verwendet wird.

Berechnet man den Spearman-Korrelationskoeffizienten für die RMSD- und cRMSD-Werte aller TransLig-Ergebnisse, so erhält man im Mittel über den Testdatensatz einen eher mäßigen Wert von 0,24. Viel wichtiger ist in diesem Zusammenhang allerdings, wie die Ligandpositionen, die nahe der nativen Position liegen relativ zu den anderen TransLig-Vorhersagen bewertet werden. Hier zeigt sich, dass bei rund der Hälfte aller Fälle die Position mit dem kleinsten cRMSD-Wert auch diejenige mit dem niedrigsten RMSD-Wert ist. Für 38 von 52 Enzymen liegt die nativ-ähnlichste Ligandposition im obersten Perzentil der Gesamtheit aller Ergebnisse und nur in einem einzigen Fall nicht unter den besten 100 Vorhersagen.

Aufgrund dieser Ergebnisse werden bei den nachfolgenden Designrechnungen jeweils die besten 100 TransLig-Vorhersagen als alternative Positionen für das Design verwendet. Darüber hinaus führt diese Beschränkung zu einer handhabbare Menge an Ligandpositionen sowohl was die Rechenzeit als auch den Speicherbedarf betrifft.

4.3.2 *In silico* Evaluation der TransCent-TransLig-Kombination

Die Bewertung der Performanz der TransCent-TransLig-Kombination gestaltet sich ähnlich schwierig wie bei TransCent selbst. Auch in diesem Fall ist die Überprüfung der Vorhersagen mittels biochemischer Experimente die einzige Möglichkeit, die Korrektheit der Modelle nachzuweisen. Dennoch kann eine *in silico* Evaluation wertvolle Hinweise dafür liefern, ob sich die von TransLig generierten Ligandpositionen tatsächlich zum Design von Enzymen eignen.

Als Maß für den Erfolg eines Designs wird wie schon bei der Bewertung von TransLig der RMSD-Wert der besten Ligandposition verwendet. Die beste Position ist in diesem Zusammenhang diejenige, die im Designmodell mit dem niedrigsten TransCent-Energiewert gewählt wurde. Der Testdatensatz bestand aus dem TransCent-Rekapitulationsdatensatz aus dem 1o0h entfernt worden war. Für dieses Protein war TransLig nicht in der Lage, geeignete Ligandpositionen vorherzusagen. Für jedes der Enzyme wurden Rekapitulationsrechnungen mit 160 Designoptimierungen durchgeführt und die besten 32 Modelle ausgewertet. Zur Auswahl standen dabei jeweils die 100 TransLig-Vorhersagen mit den besten cRMSD-Werten. In drei Fällen konnten allerdings mit TransLig keine 100 Ligandpositionen ermittelt werden, so dass nur 5, 11 bzw. 63 Ligandpositionen berücksichtigt wurden. In Analogie zur Verwendung der nativen Seitenketten (vgl. 4.2.3.6) wurden die Designrechnungen einmal mit und einmal ohne native Ligandposition durchgeführt.

Abbildung 4.27 zeigt die Verteilung der RMSD-Werte für die Ergebnisse, die unter Verwendung der nativen Position zustande kamen. Bei 34 Enzymen ist der Wert gleich 0 Å, was bedeutet, dass für das energetisch beste Design die native Ligandposition gewählt wurde. In weiteren zehn Fällen wurden ebenfalls Modelle mit dieser Position erstellt, erhielten dabei allerdings nicht den niedrigsten Energiewert. Insgesamt wurde bei einer maximalen Abweichung von 0,30 Å ein mittlerer RMSD-Wert von nur 0,02 Å erreicht.

Streicht man die native Position aus der Liste der wählbaren Ligandpositionen so erhält man einen immer noch sehr guten Mittelwert von 0,22 Å, der nur knapp über dem absoluten Minimum von 0,18 Å liegt. 18 der 52 besten Designmodelle enthalten dabei

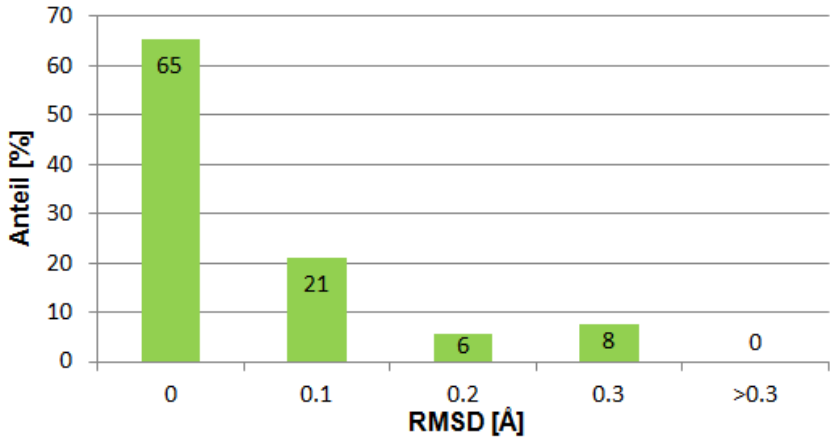


Abbildung 4.27: RMSD-Wert-Verteilung der Ligandpositionen beim Rekapitulationsdesign
Das Histogramm beschreibt die Verteilung der RMSD-Werte der Ligandpositionen, welche bei Rekapitulationsrechnungen für 52 Enzyme des Rekapitulationsdatensatzes im jeweils besten Modell gewählt wurden. Die Säule für 0 Å repräsentiert die Fälle, bei denen die native Position im Designmodell verwendet wird.

die nativ-ähnlichste Position. Der größte beobachtete RMSD-Wert beträgt in diesem Fall 1,29 Å.

Diese Ergebnisse belegen, dass TransCent in der Lage ist, bei Rekapitulationsrechnungen unter allen zur Auswahl stehenden Ligandpositionen die native und somit vermutlich beste Position zu erkennen. Das Beispiel 1q6c belegt, dass TransCent Ligandpositionen zuverlässig unterscheiden kann, die nur um 0,02 Å voneinander abweichen. Ist die native Ligandposition nicht erlaubt, so wählt TransCent eine sehr ähnliche Position und bewertet konsequenterweise diese dann i.A. energetisch etwas schlechter.

Betrachtet man die Sequenz-Rekapitulationsraten der Designmodelle so erhält man unter Verwendung der nativen Positionen einen durchschnittlichen Wert von 54,5%. Dieser liegt, in Anbetracht der geringen Zahl von Designoptimierungen, im Bereich dessen, was für Designs mit vorgegebener Ligandposition erwartet wird. Im Fall ohne native Position fällt der Wert auf 52,1%. Obwohl die Ligandpositionen im Mittel nur um 0,22 Å abweichen, schlägt sich dies also in einer deutlichen Abnahme um 2,4 Prozentpunkte nieder. Dies belegt den grundlegenden Einfluss der Ligandposition auf das Designergebnis und unterstreicht die Notwendigkeit, durch flexible Ligandpositionierung die bestmögliche Position für den Liganden zu finden.

4.3.3 DSX vs DrugScore

Die Module, welche bei TransCent maßgeblich zur Wahl der Ligandposition beitragen, sind das Modul für Ligandenbindung und das Fingerprint-Modul. Beide bewirken, dass die essentiellen Interaktionen zwischen Protein und Ligand beim Design berücksichtigt bzw. realisiert werden, und beeinflussen dadurch die energetische Bewertung einzelner Ligandpositionen. Während die Energiefunktion des PROPKA-Moduls zumindest indirekt

Modul	native Position [%]	RMSD (+nativ) [Å]	RMSD (-nativ) [Å]
Fingerprint	56	0,10	-
DSX	33	0,18	0,30
DrugScore	37	0,16	0,28

Tabelle 4.3: Beitrag der Module zur korrekten Ligandpositionierung
Angegeben sind der Anteil an Designmodellen mit nativer Ligandposition und der durchschnittliche RMSD-Wert der Ligandpositionen mit und ohne Verwendung der nativen Position beim Design.

über die Verknüpfung mit dem Fingerprint-Modul von der Position des Liganden abhängt, ignoriert das Rosetta-Modul die Ligandposition, da es nur für die Proteinstabilität verantwortlich ist.

Um den Beitrag der Module abschätzen zu können, wurden die Rekapitulationsrechnungen aus 4.3.2 wiederholt, wobei zusätzlich zu Rosetta jeweils nur entweder DSX oder das Fingerprint-Modul aktiviert war. Für jedes Enzym wurden pro Modulkombination 100 Designoptimierungen mit den jeweils besten 100 TransLig-Vorhersagen durchgeführt unter Verwendung der Standard-Modulgewichte aus der Gridsuche (siehe 4.2.2). Die Ergebnisse der Berechnungen zeigen, dass sowohl das DSX- als auch das Fingerprint-Modul als Erweiterung von Rosetta ausreichen, um nativ-ähnliche Ligandpositionen energetisch so stark zu bevorzugen, dass sie für die Designmodelle gewählt werden. Der mittlere RMSD-Wert der Ligandpositionen in den besten Modellen liegt für DSX bei 0,18 Å und 0,10 Å für das Fingerprint-Modul. Die native Position des Liganden wurde in 17 bzw. 29 von 52 Fällen verwendet.

Im Vergleich dazu erreicht TransCent für die Modulkombination Rosetta-DrugScore Werte von 0,16 Å respektive 19 erkannte nativen Positionen. Die Performanz von DrugScore ist somit annähernd identisch mit der von DSX. Dies gilt auch, wenn die native Ligandposition beim Design nicht erlaubt ist. In diesem Fall werden durchschnittliche RMSD-Werte von 0,30 Å (DSX) und 0,28 Å (DrugScore) erzielt. Alle Ergebnisse sind in Tabelle 4.3 zusammengefasst.

Die Qualität der erzielten Ergebnisse ist für die überarbeiteten Potentiale von DSX praktisch identisch mit der für DrugScore. Da die DSX-Potentiale weniger anfällig für Artefakte sind (vgl. 4.1.2), ist der Wechsel von DrugScore zu DSX gerechtfertigt.

Insgesamt zeigen die vorgestellten Ergebnisse, dass TransLig dazu in der Lage ist, nativ-ähnliche Ligandpositionen im aktiven Zentrum zu finden. Diese werden von TransCent auch als solche erkannt und als Grundlage für die Berechnung von Designmodellen verwendet. Durch den Wechsel von DrugScore zu DSX wurde außerdem die Gefahr potenialbedingter Artefakte bei der Wahl der Ligandposition ausgeschlossen. Somit stellt die Kombination von TransCent und TransLig eine geeignete Methode zur flexiblen Ligandpositionierung beim Enzymdesign dar.

4.4 Experimentelle Überprüfung

Die *in silico* Evaluation von Ergebnissen eines Enzymdesignprogramms ist wichtig, um dessen Leistungsfähigkeit einschätzen zu können, und dazu geeignet, seine Eigenschaften, Stärken und Schwächen kennenzulernen. Die Überprüfung der Vorhersagen mittels biochemischer Experimente ist allerdings die einzige Möglichkeit, die Korrektheit der Modelle tatsächlich nachzuweisen. Daher wurde eine Enzymfunktion ausgewählt und insgesamt fünf Designmodelle auf Stabilität, Bindung und Aktivität hin untersucht. Die Vorgehensweise bei der Berechnung der Designmodelle wird im Folgenden detailliert anhand eines Beispiels beschrieben und eine Zusammenstellung der Designergebnisse sowie der experimentellen Resultate angegeben.

4.4.1 PRA-Isomerase Aktivität

Die Vorhersagequalität von TransCent wurde am Beispiel der Phosphoribosylanthranilat-Isomerase Aktivität [159] überprüft. Es handelt sich dabei um eine relativ einfache Ringöffnungsreaktion, bei der Phosphoribosylanthranilat (PRA) zu 1-(o-Carboxyphenylamino)-1-desoxyribulose-5-phosphat (CdRP) umgesetzt wird (siehe Abb. 4.28).

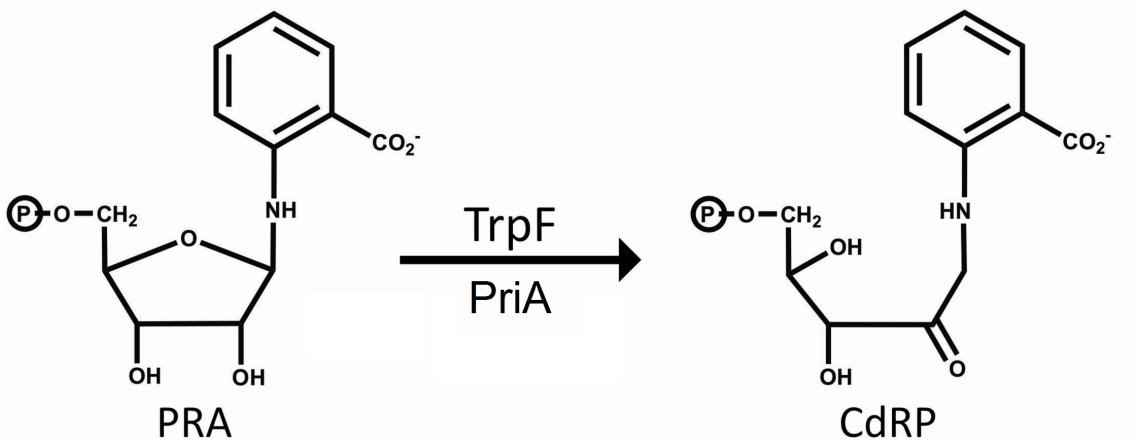


Abbildung 4.28: Reaktionsschema für die Isomerisierung von Phosphoribosylanthranilat

Die Reaktion ist ein Beispiel für eine Säure-Base-Katalyse bei der nur zwei Residuen des Enzyms direkt an der Reaktion beteiligt sind. Sie ist außerdem ein wichtiger Schritt in der Tryptophan-Biosynthese und wird von den Enzymen Phosphoribosylanthranilat Isomerase A (PriA) und Phosphoribosylanthranilat-Isomerase (TrpF) katalysiert. Beide Enzyme weisen den gleichen Faltungstyp des $(\beta\alpha)_8$-Fasses auf. Die zwei Vertreter, welche als Vorlagestrukturen bei der Funktionsübertragung dienen, sind TrpF aus *Thermotoga maritima* (PDB-Code: 1lbm) und PriA aus *Mycobacterium tuberculosis* (PDB-Code: 2y85). In beiden Strukturen ist als Ligandmolekül das Produktanalogon rCdRP kokristallisiert.

Als Zielstrukturen wurden die Proteine *tm*HisA (PDB-Code: 1qo2) und *tm*HisF (PDB-Code: 1thf) gewählt. Dabei handelt es sich um zwei Enzyme aus dem Histidin-Biosynthese-Stoffwechsel des hyperthermophilen Bakteriums *T. maritima*, d.h. beide Enzyme weisen

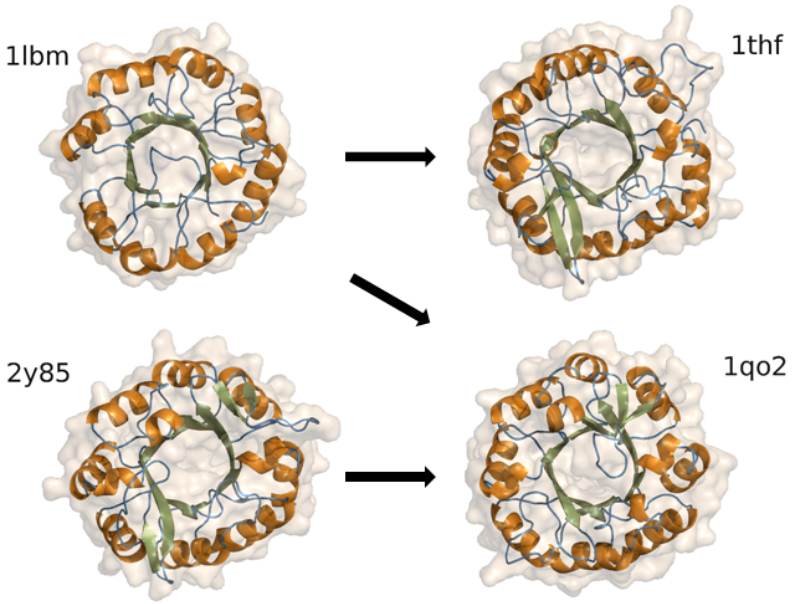


Abbildung 4.29: Überblick über die Funktionsübertragungsdesigns
Die PRA-Isomerase-Aktivität wird von *mt*PriA (2y85) und *tm*TrpF (1lbm) auf *tm*HisA (1qo2) bzw. *tm*HisA und *tm*HisF (1thf) übertragen.

eine sehr hohe natürliche Stabilität auf und bilden somit eine gute Basis für Umwandlungsdesigns.

Für drei der vier möglichen Kombinationen zwischen Vorlage- und Zielprotein wurden Funktionsübertragungsdesigns mit TransCent berechnet und umgesetzt (vgl. Abb. 4.29). Die evolutionäre Verwandtschaft von PriA und HisA macht das Enzympaar zu einem vielversprechenden Kandidaten für eine erfolgreiche Umwandlung [160]. Daher wurde versucht, den PriA-Mechanismus auf *tm*HisA zu etablieren. Außerdem wurden Designs zur Übertragung der TrpF-Funktion auf *tm*HisA und *tm*HisF berechnet.

Alle vier Enzyme gehören zur Familie der $(\beta\alpha)_8$-Fässer. Somit herrscht eine hohe strukturelle Ähnlichkeit zwischen Vorlage- und Zielproteinen, wie die paarweisen C_α-RMSD-Werte der Strukturen in Tabelle 4.4 belegen. Dadurch erhöht sich die Wahrscheinlichkeit, dass im aktiven Zentrum des Zielproteins die katalytisch wichtigen Seitenketten in ähnlicher Weise wie im Vorlageenzym platziert werden können.

	1qo2	1thf
2y85	2,66 Å	2,36 Å
1lbm	3,10 Å	2,82 Å

Tabelle 4.4: Vergleich der paarweisen C_α-RMSD-Werte
Zur Berechnung der RMSD-Werte wurde TM-Align [113] verwendet.

Darüber hinaus erlaubt die räumliche Trennung von Stabilitäts- und Katalysepol bei $(\beta\alpha)_8$-Fässern (siehe 2.1.2) die Einführung von Mutationen im aktiven Zentrum, ohne dass die Stabilität des Proteins entscheidend beeinträchtigt wird.

4.4.2 Berechnung der Designmodelle

Die Vorgehensweise bei der Berechnung der Designmodelle wird im Folgenden exemplarisch am Beispiel der Übertragung der TrpF-Aktivität auf das Proteingerüst von *tm*HisF veranschaulicht.

Als Datenbasis für das Design dienen die Kristallstrukturen von Vorlage- und Zielenzym, sowie ein MSA homologer Proteine für das Vorlageenzym. Die Strukturen (PDB-Code: 1lbm, 1thf) stammen aus der Protein Data Bank [144, 145] und wurden wie in Abschnitt 3.10.3.2 beschrieben aufbereitet. Zusätzlich wurden bei 1thf sämtliche Heteroatome entfernt. Das MSA homologer Sequenzen wurde mit Hilfe der modifizierten Version von S2MSAAA (siehe 3.4.1.1) unter Verwendung der Standardparameter (paarweise Sequenzidentität zwischen 20% und 90%, maximale Sequenzlängenabweichung 30%) erstellt und enthält 549 Sequenzen.

Zunächst musste festgelegt werden, in welchem Teil der Zielstruktur das neue aktive Zentrum etabliert werden sollte. Um die strukturelle Ähnlichkeit der beiden Enzyme auszunutzen, wurde das aktive Zentrum von HisF als mutierbarer Bereich gewählt. In konsequenter Fortführung dieser Strategie sollten auch die bereits existierenden Phosphatbindestellen „wiederverwertet“ werden, denn rCdRP, der Ligand aus TrpF, besitzt wie das natürliche Substrat von HisF (PRFAR), eine Phosphat-Gruppe, mit der der Ligand bei der Bindung in der Bindetasche „verankert“ wird. Da PRFAR zwei Phosphat-Gruppen besitzt, existieren dementsprechend auch zwei Phosphatbindestellen im aktiven Zentrum des Enzyms, eine in der N-terminalen und eine in der C-terminalen Hälfte. In der Natur wird eine Präferenz für die Bindestelle in der C-terminale Hälfte des Proteins beobachtet [20, 17]. Auch bei den bislang erfolgreich durchgeführten Designstudien [28, 161, 162, 163] konnte Bindung nur für die C-terminale Phosphatbindestelle beobachtet werden. Dennoch spricht prinzipiell nichts gegen ein Design unter Verwendung der N-terminalen Bindestelle. Aus diesem Grund wurden in der ersten Designphase Berechnungen für beide Varianten angestellt.

Die Ligandpositionen für das Design wurden mit TransLig ermittelt. Dazu wurde wie bei den Rekapitulationsrechnungen die Konformation des Liganden aus der Vorlagestruktur unverändert übernommen. Um geeignete Positionen zu berechnen, wurden die Abstandsbedingungen für die Ligandatome überwiegend aus der Vorlagestruktur abgeleitet. Nur bei den Atomen der Phosphat-Gruppe wurden die Abstände aus der Zielstruktur übernommen. Zusätzlich wurden für diese Ligandatome individuelle Interaktionspartner definiert (siehe 3.8.3.1). Dies sind jeweils zwei Residuen in der N-terminalen (Position 103, 104) bzw. in der C-terminalen Phosphatbindestelle (Position 203, 225), zu denen die Abstandsvorgaben eingehalten werden müssen. Mit dieser Vorgehensweise wird sichergestellt, dass TransLig die Phosphat-Gruppe optimal platzieren kann.

Unter Verwendung der quadratischen Kraftfunktion konnten nach jeweils 7500 Iterationsschritten (5 Startpunkte, 30 Startpunktvariationen, 50 Iterationen je Startpunktvariation) 1629 (C-terminal) bzw. 1981 (N-terminal) Ligandpositionen ermittelt werden. Diese wurden anschließend gefiltert anhand der Abweichung des Ligand-Phosphoratoms von der angestrebten Position des äquivalenten Phosphoratoms in der HisF-Kristallstruktur. Alle Positionen mit einer Abweichung kleiner als 1,0 Å wurden akzeptiert.

Aufgrund der restriktiven Filterung standen für die nachfolgenden Energieoptimierungen nur 23 bzw. 130 Ligandpositionen beim N- bzw. C-terminalen Design zur Auswahl. In

beiden Fällen wurde das gesamte aktive Zentrum von HisF (53 Proteinpositionen) als mutierbar definiert und der Rest des Enzyms auf rotierbar gesetzt (siehe Abb. 4.30). Um Artefakte bei den rotierbaren Positionen zu vermeiden wurde die Option `-use_input_sc` benutzt, mit der die Rotamerbibliotheken um die jeweils nativen Seitenkettenrotamere erweitert werden (vgl. 4.2.3.6).

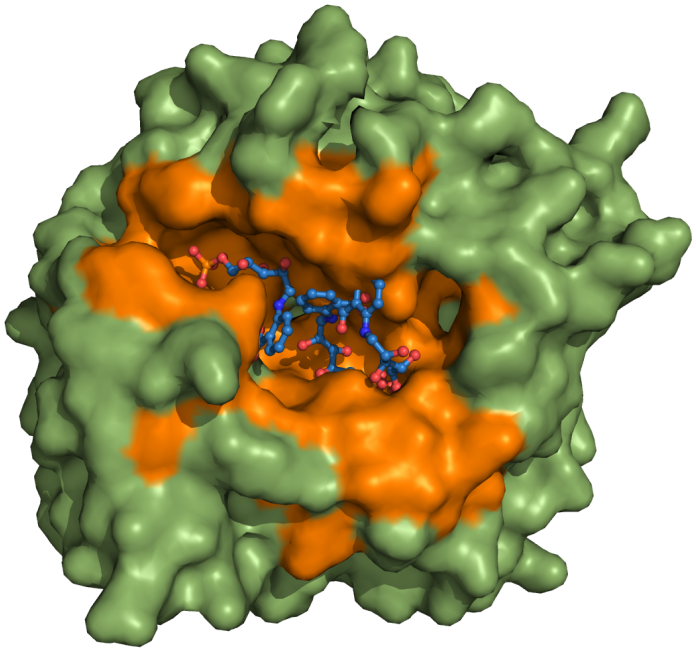


Abbildung 4.30: Mutierbarer Bereich des Zielenzyms in der ersten Designphase
Die Abbildung zeigt die Struktur des Zielenzyms in der *Surface*-Darstellung, wobei die 53 mutierbaren Positionen der ersten Designphase orange eingefärbt sind. Der grüne Bereich ist als rotierbar definiert. Zusätzlich ist eine Auswahl von drei möglichen Ligandpositionen dargestellt (blau, *Ball & Stick*-Modell).

Die Ergebnisse der jeweils 400 Designoptimierungen wurden anhand ihres TransCent-Energiewertes geordnet und die besten 20 Vertreter einer visuellen Inspektion unterzogen. Bei der C-terminalen Variante war das beste Ergebnis mit minimal veränderter Rotamerkonfiguration mehrfach vertreten, was als deutlicher Hinweis für das erfolgreiche Auffinden des globalen Energieminimums gewertet werden kann. Außerdem enthielten die besten 20 Designmodelle lediglich drei verschiedene Ligandpositionen, welche in die zweite Designphase übertragen wurden.

Bei den Designergebnissen für die N-terminale Hälfte des Proteins waren unter den besten 20 ebenfalls nur drei unterschiedliche Ligandpositionen vertreten. Andererseits konnte in diesem Fall kein ausgeprägtes Energieoptimum beobachtet werden und auch insgesamt wurden die gefunden Lösungen signifikant schlechter bewertet als die C-terminalen Designmodelle ($E_{min}^{N-term} = -358$ gegenüber $E_{min}^{C-term} = -383$ TransCent-Energieeinheiten). Dies ist größtenteils darauf zurückzuführen, dass weit weniger Wechselwirkungen zwischen Ligand und Protein modelliert werden konnten und folglich nur ein Teil der Fingerprint-Potentiale erfüllt wurde. Der Vergleich in Abbildung 4.31 macht den Unterschied deutlich.

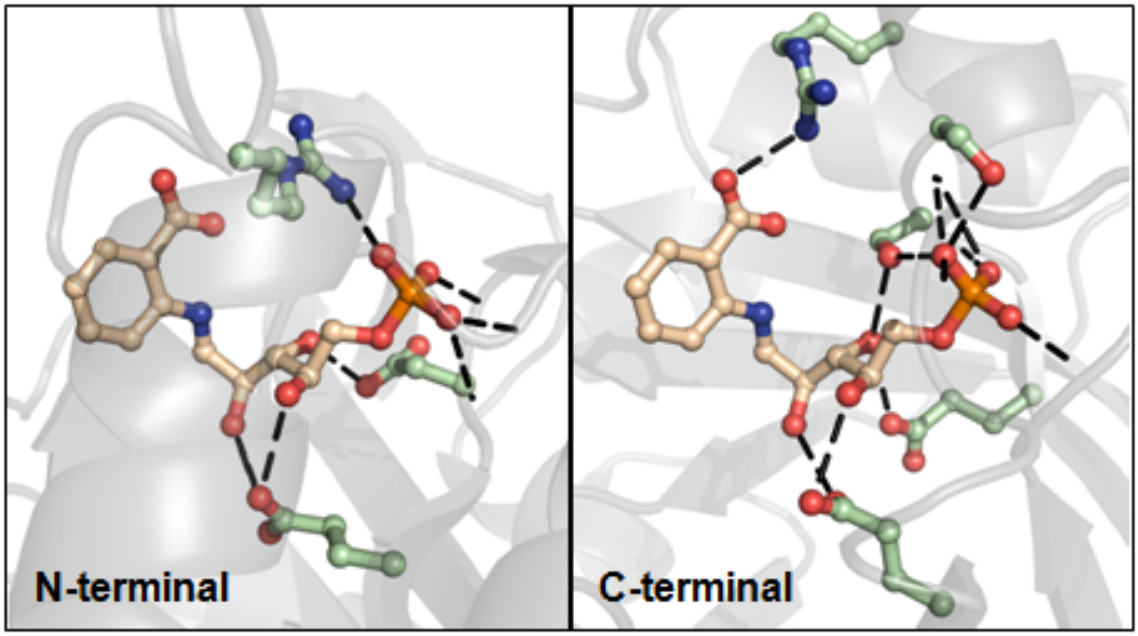


Abbildung 4.31: Vergleich der Protein-Ligand-Interaktionen bei Designmodellen mit N-terminaler bzw. C-terminaler Ligandenbindung
Die Abbildung zeigt die aktiven Zentren zweier Designmodelle mit N-terminaler bzw. C-terminaler Ligandenbindung, welche anhand des Liganden überlagert wurden. Das Ligandmolekül (cremefarben) und die mit ihm wechselwirkenden Seitenketten (grün) sind im *Ball & Stick*-Modell dargestellt. Die Wechselwirkungen selbst sind durch schwarze gestrichelte Linien angedeutet. Interaktionen mit dem Proteinrückgrat sind ebenfalls dargestellt, allerdings ohne Partneratom. Der Verlauf des Proteinrückgrates ist in Cartoon-Darstellung (grau) abgebildet.

Somit decken sich die *in silico* Ergebnisse und die experimentellen Beobachtungen, die alle darauf hindeuten, dass die N-terminale Phosphatbindestelle sich nicht für die Etablierung der TrpF-Aktivität eignet. Dementsprechend wurde die N-terminale Variante verworfen.

In der zweiten Runde der Designoptimierungen wurde neben der verkleinerten Auswahl an Ligandpositionen auch die Anzahl der mutierbaren Positionen reduziert. In der ersten Designphase hat TransCent beim energetisch besten Designmodell 32 Austausche bei 53 mutierbaren Residuen vorgenommen. Nachdem das gesamte aktive Zentrum von HisF als mutierbar definiert wurde, waren darunter auch mehrere Austausche, die nicht in unmittelbarer Umgebung des Ligandmoleküls lagen und somit nicht primär der Etablierung der Enzymaktivität dienten. Stattdessen hat TransCent versucht, die Kavität der jeweils „ungenutzten“ Hälfte des aktiven Zentrums durch den Einbau mehrerer großer Residuen zu schließen, um so die Stabilität des Proteins zu erhöhen.

Mit steigender Anzahl an Mutationen wächst aber auch die Wahrscheinlichkeit, dass es dadurch zu signifikanten Änderungen im Verlauf des Proteinrückgrats kommt, welche von TransCent nicht erfasst werden können. Daher war es ein Ziel der zweiten Designphase, die Zahl der Aminosäureaustausche auf ein notwendiges Minimum zu begrenzen. Folglich wurde die Anzahl der mutierbaren Proteinpositionen reduziert, indem Positionen mit einem Abstand von maximal 5 Å von einer der drei Ligandpositionen als mutierbar definiert

wurden. Hinzu kamen zwei Positionen, welche bei Punktrückmutationsrechnungen für die Ergebnisse der erste Runde mit einem hohen Energiegewinn bewertet wurden. Insgesamt waren damit an 20 Positionen im Proteingerüst Austausche erlaubt. Um eine feinere Abtastung des Suchraumes zu ermöglichen wurde für diese Positionen die Option `-ex2` verwendet (vgl. 3.2.2). Damit erzeugt die Modellierungseinheit zusätzliche Rotamervarianten, deren χ_2-Winkel leicht von den optimalen Werten abweichen.

In der zweiten Runde wurden 700 Designoptimierungen mit TransCent durchgeführt und die Ergebnisse wiederum entsprechend ihres Energiewertes sortiert. Unter den besten 100 Designmodellen war in diesem Fall ausschließlich die Ligandposition vertreten, welche bereits in der ersten Runde zum Modell mit der niedrigsten Energie geführt hatte. Außerdem wurde auch hier das beste Modell mehrfach erzeugt.

Abschließend wurden für die ausgewählten Designmodelle die Energiedifferenzen der Punktrückmutationen berechnet und alle Austausche mit einem Energiegewinn unter 3,5 TransCent-Energieeinheiten zurückgenommen. Die Ergebnisse der einzelnen Funktionsübertragungsdesigns sind im folgenden Abschnitt aufgeführt.

4.4.3 Experimentelle Überprüfung der TransCent-Vorhersagen

Zur Überprüfung des Designalgorithmus in Kombination mit dem TransLig-Modul wurden insgesamt fünf Designmodelle für drei unterschiedliche Kombinationen von Vorlage- und Zielenzym (vgl. 4.4.1) umgesetzt und experimentell auf Stabilität, Bindung und Aktivität hin untersucht.

4.4.3.1 Designergebnisse

Die Berechnung der Designergebnisse erfolgte in allen Fällen gemäß dem in Abschnitt 4.4.2 vorgestellten Designprotokoll. Auf Abweichungen vom beschriebenen Protokoll wird nachfolgend hingewiesen. Außerdem werden zur besseren Unterscheidbarkeit die Bezeichner für wildtypische Residuen im Folgenden kursiv gesetzt (z.B. *Asp147*).

PriA(HisA) Da das PriA-Enzym nur in einigen Actinobakterien vorkommt, wurde beim Erstellen des MSAs homologer Sequenzen auf eine obere Grenzen bei der Filterung nach Sequenzidentität verzichtet. Stattdessen wurde das „Roh-MSA“ manuell gefiltert um sicherzustellen, dass ausschließlich aus Actinobakterien stammende Sequenzen enthalten waren. Aufgrund der großen Ähnlichkeit zum Zielenzym HisA mussten zusätzlich alle HisA-Sequenzen aus dem MSA entfernt werden.

In der zweiten Designphase wurden 1000 Designoptimierungen durchgeführt. Dabei standen vier Ligandpositionen zur Auswahl und 17 Proteinpositionen in der C-terminalen Hälfte des aktiven Zentrums von *tm*HisA waren als mutierbar definiert. Das energetisch beste Designmodell PA620 enthielt neun Mutationen gegenüber der Wildtypsequenz. Aufgrund der Punktrückmutationsrechnungen wurden drei der neun Mutationen wieder zurückgenommen (vgl. Tabelle 4.5), sodass das endgültige Designmodell nur sechs Mutationen enthielt.

Position	*tm*HisA	PA620	ΔE	PRM
6	A	A	-	
8	D	E	-32,8	
48	H	H	-	
50	V	V	-	
79	G	G	-	
80	G	G	-	
100	I	S	-15,8	
125	S	R	-34,8	
127	D	E	-3,4	x
164	T	E	-46,8	
169	D	K	-12,1	
194	A	S	-10,3	
197	I	I	-	
198	S	S	-	
221	I	D	-2,0	x
222	V	V	-	
224	R	E	-0,4	x

Tabelle 4.5: Designmodell PA620 - Übertragung des PriA-Mechanismus auf *tm*HisA
In der Tabelle sind die 17 Proteinpositionen aufgeführt, welche in der abschließenden Designrunde als mutierbar definiert waren. Angegeben sind jeweils die wildtypische Aminosäure in *tm*HisA, die Aminosäure im Designmodell, der Energiegewinn (ΔE) einer Mutation relativ zum Designmodell und ob eine Position für eine Punktrückmutation ausgewählt wurde (x in Spalte PRM).

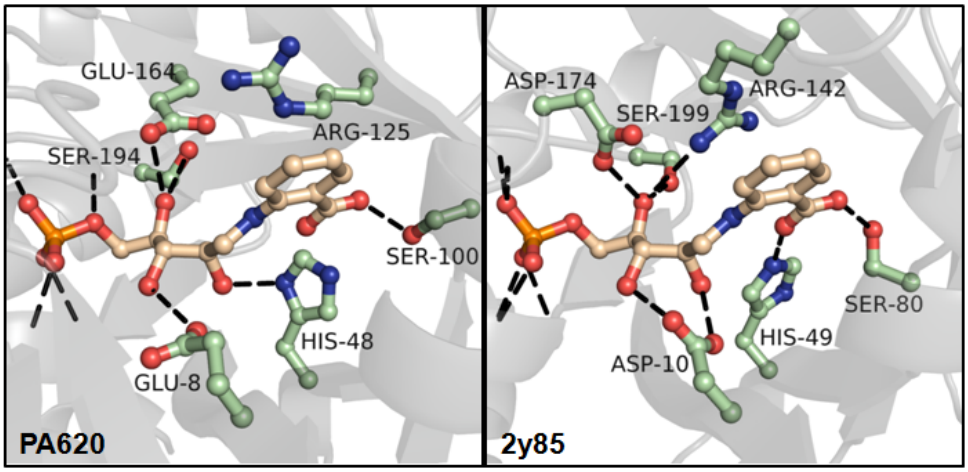


Abbildung 4.32: Vergleich der aktiven Zentren von PA620 und der PriA Wildtyp-Struktur 2y85
Die Abbildung zeigt die aktiven Zentren von PA620 und 2y85, welche anhand der Ligandposition überlagert wurden. Das Ligandmolekül (cremefarben) und die mit ihm wechselwirkenden Seitenketten (grün) sind im *Ball & Stick*-Modell dargestellt. Die Wechselwirkungen werden durch schwarze gestrichelte Linien zwischen Ligand- und Proteinatom angedeutet. Zusätzlich werden die äquivalenten Seitenkette von Arg125 und *Arg142* gezeigt. Interaktionen mit dem Proteinrückgrat sind als gestrichelte Linien ohne Partneratom dargestellt. Der Verlauf des Proteinrückgrates ist in Cartoon-Darstellung (grau) gezeigt.

Vergleicht man das aktive Zentrum von PA620 und der PriA-Struktur 2y85, so zeigt sich, dass alle mutmaßlich wichtigen Wechselwirkungen zwischen Protein und Ligand ausgebildet werden, oder zumindest die entsprechenden Seitenketten vorhanden sind (siehe Abb. 4.32). Das Histidin-Residuum His48, welches bereits wildtypisch in *tm*HisA vorhanden ist, wird, obgleich mutierbar, vom Designalgorithmus beibehalten und entspricht *His50* in *mt*PriA. Die Residuen *Asp10* und *Asp174* im Vorlageenzym werden durch Glu8 und Glu164 im Modell „ersetzt". Durch die Mutationen I100S und A194S werden die Serin-Seitenketten eingeführt, die mit dem Ligand wechselwirken. Auch das Residuum *Arg142*, welches in *mt*PriA den Carboxylat-Rest von rCdRP koordiniert, findet mit Arg125 eine Entsprechung im Designmodell. Allerdings ist aufgrund der Orientierung der Seitenkette der Abstand zwischen der Guanidinogruppe von Arg125 und dem Carboxylat-Rest zu groß (6,3 statt 4,1 Å), so dass keine Wechselwirkung zustande kommt.

TrpF(HisA) Für die Übertragung des TrpF-Mechanismus auf das *tm*HisA-Proteingerüst wurden in der zweiten Designphase 1000 Designoptimierungen berechnet. Dabei konnte TransCent zwischen zwei Ligandpositionen wählen, deren Phosphat-Gruppe in beiden Fällen in der C-terminalen Bindestelle von *tm*HisA platziert war. An den 15 mutierbaren Positionen wurden beim energetisch besten Modell (TA893) 13 Aminosäureaustausche eingeführt, von denen allerdings vier aufgrund des zu geringen Energiegewinns revertiert wurden. Zusätzlich wurde auch das beste Designmodell mit der zweiten Ligandposition (TA247) für eine experimentelle Überprüfung ausgewählt. Von dessen ursprünglich 14 Mutationen wurden drei zurückgenommen, so dass sich TA247 und TA893 an insgesamt sechs Positionen unterscheiden (siehe Tabelle 4.6).

Position	***tm*HisA**	**TA893**	**ΔE**	**PRM**	**TA247**	**ΔE**	**PRM**
6	A	A	-		E	-27,0	
8	D	E	-32,0		S	-29,2	
48	H	S	-17,0		A	-19,5	
50	V	Q	-14,7		Q	-13,6	
125	S	A	-3,5	x	A	-0,8	x
127	D	K	-13,5		I	-11,2	
162	V	H	-24,5		H	-23,3	
164	T	G	-41,0		A	-18,8	
166	I	T	-2,2	x	A	-4,6	
169	D	R	-26,8		R	-29,2	
194	A	G	-33,9		G	-31,4	
197	I	I	-		I	-	
198	S	Y	-1,6	x	F	-0,4	x
221	I	E	-25,3		E	-23,7	
224	R	W	-2,2	x	K	-0,6	x

Tabelle 4.6: Designmodelle TA893 & TA247 - Übertragung des TrpF-Mechanismus auf *tm*HisA
In der Tabelle sind die 15 Proteinpositionen aufgeführt, welche in der abschließenden Designrunde als mutierbar definiert waren. Angegeben sind jeweils die wildtypische Aminosäure in *tm*HisA, die Aminosäure im Designmodell, der Energiegewinn (ΔE) einer Mutation relativ zum Designmodell und ob eine Position für eine Punktrückmutation ausgewählt wurde (x in Spalte PRM).

Beim Vergleich der Modelle mit der Struktur des Protein-Ligand-Komplexes von *tm*TrpF fällt zunächst auf, dass das Ligandmolekül um ca. 90° um seine Achse gedreht in der Bindetasche liegt (Abb. 4.33). Diese Orientierung erlaubt, dass die Carboxylat-Gruppe des Liganden sowohl bei TA893 als auch bei TA247 durch Arg169 koordiniert wird. Das Residuum übernimmt somit die Rolle von *Arg36* im wildtypischen Enzym. Die Interaktionen der Residuen *Asp126* und *Asp178* werden bei TA893 durch die Seitenketten von Glu8 und Glu221 realisiert. Obwohl deren Seitenketten länger sind als die ihrer Aspartat-Äquivalente, ist infolge des Rückgratverlaufs der Abstand zum Liganden größer als in der Vorlage. Aufgrund der Ligandorientierung existieren im Proteingerüst von *tm*HisA keine Positionen, auf denen die Serin-Wechselwirkungen modelliert werden können.

Das Designmodell von TA247 unterscheidet sich von TA893 hauptsächlich durch den „Ringtausch“ der Aminosäuren auf den Positionen 6, 8 und 48 (vgl. Abb. 4.34). Außerdem wird an Position 127 die positive Ladung von Arg127 durch das aliphatische Ile127 ersetzt.

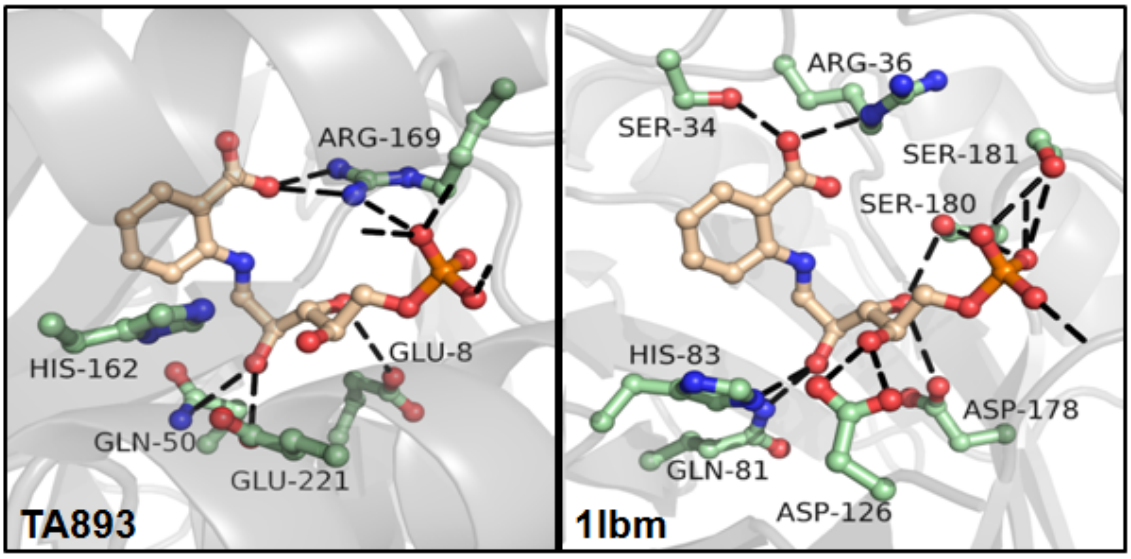


Abbildung 4.33: Vergleich der aktiven Zentren von TA893 und der TrpF Wildtyp-Struktur 1lbm
Die Abbildung zeigt die aktiven Zentren von TA893 und 1lbm, welche anhand des Liganden überlagert wurden. Das Ligandmolekül (cremefarben) und die mit ihm wechselwirkenden Seitenketten (grün) sind im *Ball & Stick*-Modell dargestellt. Die Wechselwirkungen zwischen Ligand- und Proteinatom werden durch schwarze gestrichelte Linien angedeutet. Bei TA893 ist zusätzlich das zu *His83* äquivalente Residuum His162 dargestellt. Interaktionen zwischen Ligandmolekül und Proteinrückgrat sind als gestrichelte Linien ohne Partneratom dargestellt. Der Verlauf des Proteinrückgrates ist in Cartoon-Darstellung (grau) gezeigt.

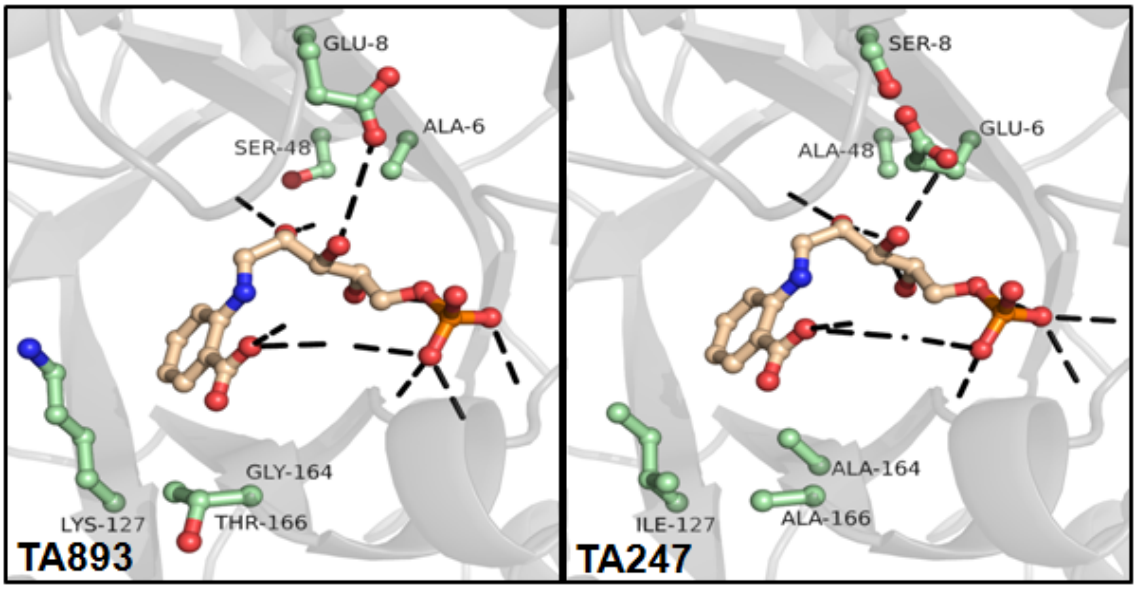


Abbildung 4.34: Vergleich der aktiven Zentren von TA893 und TA247
Neben dem Ligandmolekül (cremefarben) und den Protein-Ligand-Wechselwirkungen (schwarze gestrichelte Linien) sind die Seitenketten (grün) der sechs sich unterscheidenden Positionen im *Ball & Stick*-Modell dargestellt. Bei Gly164 von TA893 ist aufgrund des Fehlens einer Seitenkette nur dessen Position markiert. Die beiden Strukturen wurden anhand des Proteinrückgrats überlagert. Dessen Verlauf ist in Cartoon-Darstellung (grau) angedeutet.

TrpF(HisF) Wie bei der Kombinationen TrpF(HisA) wurde auch in diesem Fall jeweils das energetisch beste Designmodell der beiden besten Ligandpositionen für die Umsetzung im Labor ausgewählt. Die Modelle TF4 und TF148 enthalten nach Abzug der Punktrückmutationen 14 bzw. 13 Aminosäureaustausche bezogen auf die Wildtypsequenz von *tm*HisF. Untereinander unterscheiden sich die Modelle an vier Positionen, wie Tabelle 4.7 zu entnehmen ist.

Die aktiven Zentren von TF4 und TF148 stimmen größtenteils sehr gut mit den Vorgaben in der Struktur von *tm*TrpF überein. Fast alle Wechselwirkungen zwischen Enzym und Ligandmolekül sind an ähnlicher Position und in ähnlicher Orientierung modelliert worden. Wie bei den vorhergehenden Designmodellen sind auch hier die Residuen *Asp126* und *Asp178* durch entsprechende Glutamat-Residuen (Glu171 und Glu222) ersetzt. Das Modell TF4 verfügt mit Phe50 sogar über ein passend platziertes π-System welches mit dem aromatischen Ring von rCdRP wechselwirken kann, analog zu *Tyr31* in der Vorlage (siehe Abb. 4.35). Für die Residuen *Gln81* und *His83* existiert allerdings keine Entsprechung im aktiven Zentrum von TF4. Die Koordination der Carboxylat-Gruppe des Liganden durch Arg228 wird zusätzlich durch Lys18 unterstützt.

In TF148 wird die positive Ladung von Lys18 durch negative Ladung von Glu11 ersetzt, möglicherweise zur Optimierung des pK_a-Wertes von Arg228 (vgl. Abb. 4.36). Außerdem wird an Position 50 statt einer aromatischen Seitenkette in Form von Gln50 die zu *Gln81* analoge Wechselwirkung modelliert.

Position	*tm*HisF	TF148	ΔE	PRM	TF4	ΔE	PRM
9	C	C	-		C	-	
11	D	E	-7,2		S	-10,5	
18	V	Q	-9,1		K	-4,8	
48	V	H	-18,6		H	-19,5	
50	L	Q	-13,9		F	-5,6	
52	I	G	-25,1		A	-6,9	
171	T	E	-25,5		E	-27,8	
173	I	R	-5,4		R	-7,6	
175	R	K	-0,1	x	K	-2,7	x
176	D	A	-16,1		A	-14,6	
177	G	F	-7,6		F	-6,1	
201	S	G	-1,7	x	G	-7,8	
202	G	G	-		G	-	
204	A	G	-5,6		G	-5,2	
222	L	E	-25,3		E	-26,8	
223	A	A	-		G	0,0	x
224	A	S	-7,2		S	-3,7	
225	S	S	-		S	-	
226	V	P	-0,3	x	P	-3,1	x
228	H	R	-30,5		R	-26,9	

Tabelle 4.7: Designmodelle TF148 & TF4 - Übertragung des TrpF-Mechanismus auf *tm*HisF
In der Tabelle sind die 20 Proteinpositionen aufgeführt, welche in der abschließenden Designrunde als mutierbar definiert waren. Angegeben sind jeweils die wildtypische Aminosäure in *tm*HisF, die Aminosäure im Designmodell, der Energiegewinn (ΔE) einer Mutation relativ zum Designmodell und ob eine Position für eine Punktrückmutation ausgewählt wurde (x in Spalte PRM).

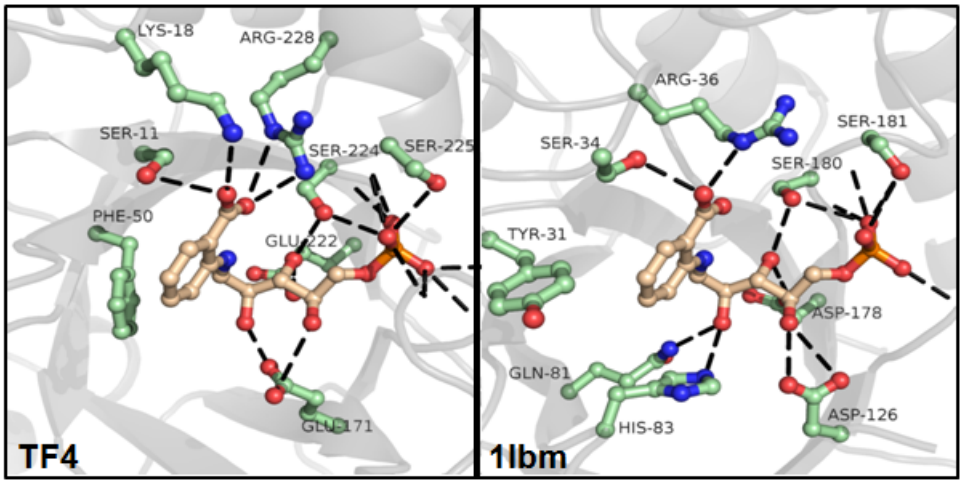


Abbildung 4.35: Vergleich der aktiven Zentren von TF4 und der TrpF Wildtyp-Struktur 1lbm
Die Abbildung zeigt die aktiven Zentren von TF4 und 1lbm, welche anhand des Liganden überlagert wurden. Das Ligandmolekül (cremefarben) und die mit ihm wechselwirkenden Seitenketten (grün) sind im *Ball & Stick*-Modell dargestellt. Die Wechselwirkungen zwischen Ligand und Protein werden durch schwarze gestrichelte Linien angedeutet. Zusätzlich sind die aromatischen Seitenketten Phe50 und *Tyr31* dargestellt die mit dem Ring des Liganden ebenfalls wechselwirken können. Interaktionen zwischen Ligandmolekül und Proteinrückgrat sind als gestrichelte Linien ohne Partneratom dargestellt. Der Verlauf des Proteinrückgrates ist in Cartoon-Darstellung (grau) gezeigt.

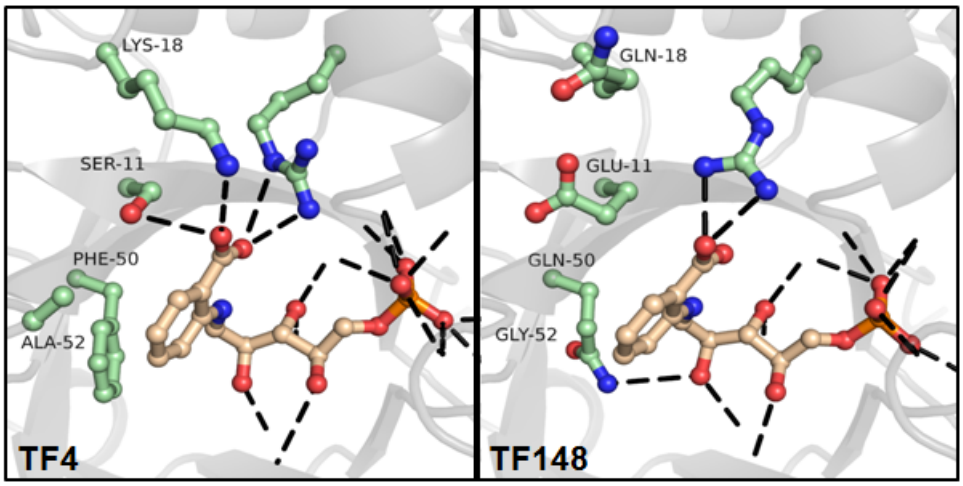


Abbildung 4.36: Vergleich der aktiven Zentren von TF4 und TF148
Neben dem Ligandmolekül (cremefarben) und den Protein-Ligand-Wechselwirkungen (schwarze gestrichelte Linien) sind die Seitenketten (grün) der vier sich unterscheidenden Positionen im *Ball & Stick*-Modell dargestellt. Bei Gly52 von TF148 ist aufgrund des Fehlens einer Seitenkette nur dessen Position markiert. Zusätzlich die unterschiedlichen Rotamere der Seitenkette von Residuum Arg228 gezeigt. Die beiden Strukturen wurden anhand des Proteinrückgrats überlagert. Dessen Verlauf ist in Cartoon-Darstellung (grau) angedeutet.

4.4.3.2 Laborergebnisse

Die experimentelle Überprüfung der Designergebnisse wurde von Bernd Reisinger aus der Arbeitsgruppe von Prof. Reinhard Sterner durchgeführt. Alle fünf getesteten Proteine erwiesen sich als stabil und konnten gut gereinigt werden. Die Aktivität der Enzyme wurde mittels *in vivo*-Komplementationsansatz in *Escherichia coli* untersucht (vgl. Methoden in [162]). Dabei wird ein *E. coli*-Stamm verwendet, bei dem das Gen deaktiviert wurde, welches für das Enzym mit der betreffenden Funktion kodiert. Da es sich bei der PRA-Isomerase Aktivität um einen essentiellen Schritt im Stoffwechsel von *E. coli* handelt, können die Organismen ohne funktionstüchtiges Enzym nicht wachsen. In diese Organismen wird über ein Plasmid das designte Enzym in Form des entsprechenden Gens eingeschleust. Dadurch kann festgestellt werden, ob das eingeschleuste Konstrukt komplementiert, d.h. ob es die fehlende Aktivität ersetzen kann und somit das Wachstum des Organismus ermöglicht. Da bei keiner der fünf Varianten Wachstum beobachtet wurde, muss daraus geschlossen werden, dass, zumindest bei physiologischen Bedingungen, keine messbare enzymatische Aktivität bei den Designmodellen vorhanden ist.

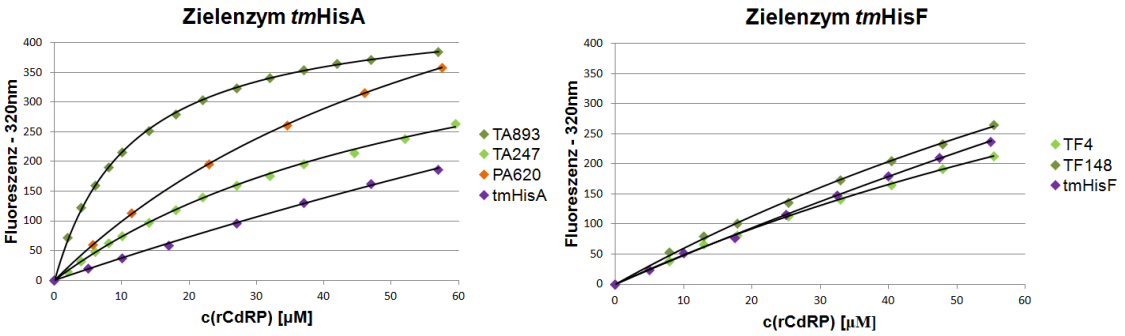


Abbildung 4.37: Titrationskurven zur Bestimmung der Dissoziationskonstanten
Die Abbildungen zeigen die Resultate der Fluoreszenztitrationsmessungen mit dem Produktanalogon rCdRP für die fünf getesteten Designmodelle und die beiden Zielenzyme *tm*HisA und *tm*HisF. Alle Kurvenverläufe wurden jeweils mit einer hyperbolischen Funktion gefittet (schwarze Linien), anhand derer die Werte der Dissoziationskonstanten bestimmt werden. Die Ergebnisse sind in Tabelle 4.8 aufgeführt.

Um die Gründe für die fehlende Aktivität zu untersuchen wurden alle Modelle zusätzlich auf Ligandenbindung getestet. Dazu wurden Fluoreszenztitrationsmessungen mit dem Produktanalogon rCdRP durchgeführt (vgl. Methoden in [28]) und jeweils die Dissoziationskonstante K_D bestimmt. Die gemessenen Werte der Modelle sind in Tabelle 4.8 angegeben. Es zeigt sich, dass die Bindung von rCdRP bei den drei Modellen, für die *tm*HisA als Zielgerüst verwendet wurde, merklich gegenüber der Ausgangssituation verbessert werden konnte. Vor allem TA893 liegt mit einem Wert von $K_D = 18$ µM fast im Bereich der wildtypischen Bindungsaffinität von *tm*TrpF. Dagegen lassen die beiden Trpf(HisF)-Varianten TF4 und TF148 keine spezifische Bindung erkennen (siehe Abb. 4.37). Die Ergebnisse werden ausführlich in Abschnitt 5.4 diskutiert.

4.4.3.3 In silico Nachbereitung

Auch wenn bei keinem der Designmodelle eine messbare PRA-Isomerase-Aktivität festgestellt wurde, so konnte doch in drei Fällen, mit unterschiedlichem Erfolg, die spezifische Bindung von rCdRP auf dem Proteingerüst etabliert werden. Dabei ist die Bindungsaffinität von TA893 deutlich besser als die von TA247. Dies ist auch die Reihenfolge, wie die Modelle von TransCent energetisch bewertet worden sind. Daraus ergibt sich die Frage, ob sich dieser schwache Trend auch bei den anderen Designmodellen fortsetzt. Da die TransCent-Energiewerte modellabhängig sind, können sie allerdings nur schwer direkt miteinander verglichen werden.

Verwendet man statt der gesamten Energiefunktion nur den DSX-Anteil so wird zum einen die Vergleichbarkeit erhöht und zum anderen ausschließlich die Ligandenbindung bewertet. Die in TransCent eingesetzte rotamerbasierte Version arbeitet allerdings mit einer reduzierten Scoring-Funktion. Daher wird die Bewertung der Modelle mit dem DSX-Server[1] durchgeführt, so dass auch Desolvatationseffekte bei der Bindung und die Torsionswinkel des Ligandmoleküls mit berücksichtigt werden können (vgl. 3.3.1). Die Ergebnisse sind zusammen mit den gemessenen Dissoziationskonstanten in Tabelle 4.8 aufgeführt. Wie man sieht, stimmen die Vorhersagen von DSX relativ gut mit den experimentellen Ergebnissen überein. Klammert man die beiden TF-Varianten aus, so ergibt sich eine fast perfekte Rangkorrelation zwischen DSX-Score und K_D-Wert.

Protein	**Dissoziationskonstante** K_D^{rCdRP} [µM]	**DSX-Score**
TF4	>100	-122
*mt*PriA	-	-119
TF148	>100	-114
HisF-D130V+D176V	0,18[a]	-114
*tm*TrpF	4,7[b]	-111
HisA-II	4,5[c]	-104
TA893	16±4	-95
TF72N	-	-90
TA247	60	-78
PA620	69	-59
*tm*HisA	>100	-
*tm*HisF	>100[d]	-

[a] Wert wurde übernommen aus [163]
[b] Wert wurde übernommen aus [159]
[c] Wert wurde übernommen aus [162]
[d] Wert wurde übernommen aus [164]

Tabelle 4.8: Vergleich mehrerer Dissoziationskonstanten und DSX-Scores für die Bindung von rCdRP
Bei den Berechnungen mit DSX wurden die Strukturen der Designmodelle bzw. die der Vorlageenzyme für die TransCent-Designs verwendet. Für HisF-D130V+D176V existiert eine Kristallstruktur mit gebundenem Produktanalogon (PDB-Code: 4ewn). Bei HisA-II musste der Ligand zunächst mit Hilfe von SwissDock [165, 166] in die Struktur (PDB-Code: 2w79) gedockt werden. Dissoziationskonstanten >100 µM entsprechen einer unspezifischen Bindung. Der Wert für TA893 ist das Mittel aus drei unabhängigen Messungen.

[1] http://pc1664.pharmazie.uni-marburg.de/drugscore/, Version 0.88

Um die Ergebnisse besser einordnen zu können sind zusätzlich die Werte für die wildtypischen Zielenzyme *tm*HisA und *tm*HisF und die Vorlageenzyme *mt*PriA und *tm*TrpF angegeben. Außerdem ist auch der DSX-Wert für TF72N aufgeführt, dem besten Designmodell mit Ligand gebunden in der N-terminalen Phosphatbindestelle. Vervollständigt wird die Liste durch zwei Enzyme aus früheren Designstudien [162, 163], auf denen die PRA-Isomerase-Aktivität etabliert werden konnte (*tm*HisF-D130V+D176V und HisA-II).

5 Diskussion

Im folgenden Abschnitt wird auf einen Teil der Ergebnisse aus dem letzten Kapitel nochmals eingegangen und deren Bedeutung auch im Zusammenhang mit aktuellen Forschungsergebnissen diskutiert.

5.1 TransLig - ein wichtiger Schritt hin zur flexiblen Ligandenpositionierung

Im Zentrum dieser Arbeit steht das Ziel, das Enzymdesignprogramm TransCent um die Fähigkeit zur flexiblen Positionierung des Ligandmoleküls zu ergänzen. Abgesehen von der Stabilität und der Löslichkeit eines Proteins hängen alle Bedingungen, welche beim Design von Enzymen berücksichtigt werden müssen, direkt oder indirekt mit der relativen Ausrichtung von Protein und Ligand zusammen.

Sowohl bei der Wahl bindungsvermittelnder Aminosäuren als auch bei der Platzierung der katalytischen Residuen ist die entscheidende Rolle des Liganden evident. Die entsprechenden Seitenketten im aktiven Zentrum des Enzyms sollen, zusätzlich zu eventuellen Wechselwirkungen mit dem Proteinrückgrat, die Bildung des Protein-Ligand-Komplexes ermöglichen und darüber hinaus den Übergangszustand des Liganden bei der katalysierten Reaktion stabilisieren. Dazu muss ein komplexes, fein aufeinander abgestimmtes und stabiles Gefüge von Aminosäureseitenketten im aktiven Zentrum etabliert und relativ zum Liganden ausgerichtet werden.

Auch die pK_a-Werte der titrierbaren Gruppen in der Ligandenbindetasche können mehr oder minder stark bei der Bindung beeinflusst werden, sei es durch das Aufbrechen eines Wasserstoffbrückennetzwerkes oder weil der Ligand geladen ist [167]. Hierbei ist ebenfalls die relative Lage von Protein und Ligandmolekül maßgeblich, so dass auch bei der pK_a-Wert-Berechnung die Position des Liganden berücksichtigt werden sollte.

Um all dem Rechnung zu tragen wurde zum einen der TransCent-Algorithmus erweitert, um bei der Designoptimierung mehrere Ligandpositionen verarbeiten zu können und zum anderen das TransLig-Modul zur Vorhersage geeigneter Ligandpositionen entwickelt. Die speziellen Anforderungen beim Enzymdesign machen den Einsatz „normaler“ Docking-Programme unmöglich, da das aktive Zentrum, in dem der Ligand binden soll, ja erst noch modelliert werden muss. Die Strategie, welche in TransLig verfolgt wird, hat daher nicht zum Ziel, die Ligandpose mit der besten Bindungsenergie zu finden. Stattdessen wird versucht, solche Positionen zu ermitteln, welche dem Designprozess größtmögliche Freiheit bei der Wahl der Seitenketten lassen, so dass die essentiellen Interaktionen zwischen Ligand und Enzym modelliert werden können. TransLig verwendet dabei ein vergleichsweise einfaches Kräftemodell, welches die Richtung der Translationen und Rotationen bei der iterativen Suche nach geeigneten Positionen vorgibt. Trotzdem ist das Programm in der Lage, für alle Enzyme des getesteten Datensatzes nativ-ähnliche Positionen vorherzusagen.

Legt man den RMSD-Schwellwert von 2 Å an, welcher bei normalen Docking-Programmen üblicherweise als Grenzwert für erfolgreiche Dockingvorgänge gilt [94], so ist nur in einem einzigen von 52 Fällen die am besten bewertete Ligandposition außerhalb des akzeptablen Bereichs.

Erleichternd kommt TransLig dabei zugute, dass die Bindetasche, in der der Ligand positioniert werden soll, vorgegeben werden muss. Andererseits verzichtet das Programm auf den Einsatz aufwändiger Atomtyp-spezifischer Potentiale und kann bei der Bewertung der Positionen keine charakteristischen Interaktionen wie Wasserstoffbrücken oder elektrostatische Wechselwirkungen zwischen Protein und Ligand berücksichtigen, da der Suchraum nur vom Proteinrückgrat definiert wird. Dennoch erreicht TransLig bei Rekapitulationsrechnungen eine beachtliche Performanz und liefert auch beim Design von Funktionsübertragungen geeignete Positionen.

Offensichtlich reicht also die Vorgabe der einzuhaltenden Abstände aus, um nativ-ähnliche Ligandpositionen zu identifizieren. Im Allgemeinen stellen diese dann die Positionen mit den niedrigsten cRMSD-Werten dar. Bei einigen wenigen Beispielen werden aber auch stark abweichende Ligandpositionen gut bewertet. Dies bedeutet, dass mindestens eine weitere Möglichkeit besteht, Protein und Ligand so anzuordnen, dass die Abstandsvorgaben (fast) perfekt eingehalten werden. Diese Fälle böten eine interessante Grundlage für Untersuchungen, ob dieselbe Enzymfunktion auf einem Proteingerüst auf zwei unterschiedliche Weisen etabliert werden könnte.

Limitationen bei der Suche nach geeigneten Ligandpositionen ergeben sich für TransLig aufgrund des Umstandes, dass das Proteinrückgrat, welches den Suchraum definiert, starr gehalten wird. Zum einen ist TransLig grundsätzlich darauf angewiesen, dass auf der Proteinoberfläche bereits eine Vertiefung vorhanden ist, welche als Ligandenbindetasche verwendet werden kann. Zum anderen muss diese auch ausreichend Platz für das Ligandmolekül bieten. Gerade bei großen Liganden kann dabei der Fall eintreten, dass keine passenden Ligandpositionen gefunden werden, obwohl möglicherweise nur kleine Anpassungen der Proteinstruktur notwendig wären. Im Gegensatz dazu ist der gewählte Ansatz aber geeignet, um unterschiedliche Ligandkonformationen bei der Berechnung zu verwenden. Diese müssen dem Programm vorgegeben werden und können unabhängig voneinander im aktiven Zentrum platziert werden.

Das Auffinden passender Ligandpositionen ist allerdings noch keine hinreichende Voraussetzung für ein erfolgversprechendes Enzymdesign. Neben der grundsätzlichen Qualität eines Designalgorithmus ist ebenso entscheidend, dass dieser geeignete Ligandpositionen auch als solche erkennt und beim Designprozess verwendet. Die Ergebnisse der Rekapitulationsdesigns mit einer Auswahl von jeweils 100 Ligandpositionen haben gezeigt, dass TransCent in der Lage ist, die native Position bzw. nativ-ähnliche Positionen zu identifizieren und darauf Designmodelle aufzubauen. Sowohl das DSX- als auch das Fingerprint-Modul tragen dazu bei, dass diese Positionen mit der niedrigsten und somit besten Energie bewertet werden. Darüber hinaus zeigen die Rekapitulationsrechnungen, dass es offensichtlich nicht beliebig viele gleichwertige Möglichkeiten gibt, die wichtigen Enzym-Ligand-Interaktionen bei gleichzeitiger Sicherung der Proteinstabilität zu etablieren. Stattdessen belegt der mittlere RMSD-Wert von nur 0,02 Å ein weiteres Mal, dass Enzyme hochspezialisierte Proteine sind, welche im Lauf der Evolution für ihre Aufgabe optimiert wurden.

Trotz der minimalen Abweichung von durchschnittlich 0,02 Å sinkt die mittlere Rekapitulationsrate im Vergleich zum Design mit der nativen Ligandposition um 3,9 Prozent-

punkte. Dieser deutliche Effekt unterstreicht, wie groß der Einfluss der Ligandposition auf das Designergebnis ist. Umso wichtiger erscheint dadurch die durchgeführte Erweiterung von TransCent, um den Liganden bei Funktionsübertragungen passend positionieren zu können.

Um die Erfolgsaussichten beim Design weiter zu erhöhen bietet sich der Einsatz von bereits vorhandenem Wissen über das Zielprotein bzw. die Enzymfunktion an. Dabei können unterschiedliche Strategien verfolgt werden, wie etwa die Wiederverwendung existierender Elemente des Zielproteins. Verfügen beispielsweise sowohl das native Ligandmolekül als auch der Ligand der zu übertragenden Enzymfunktion über eine Phosphatgruppe mit der sie an das Enzym binden, so kann bei geschickter Positionierung des Liganden die vorhandene Phosphatbindetasche in das Design eingebaut werden. TransLig bietet daher die Möglichkeit, mittels Spezifikation individueller Interaktionspartner einen Teil des Ligandmolekül bei der Suche nach geeigneten Positionen quasi zu „verankern“ während der Rest entsprechend den vorgegebenen Abstandsbedingungen frei bewegt werden kann. Die Nützlichkeit der Strategie hat sich bei den durchgeführten Funktionsübertragungen gezeigt. Besonders interessant erscheint diese Vorgehensweise, wenn man bedenkt, dass fast die Hälfte aller bekannten Proteine mit phosphathaltigen Liganden [124] interagiert. Die Anwendung ist aber nicht auf Phosphatgruppen beschränkt, sondern kann überall angewendet werden, wo dies sinnvoll erscheint.

Die Einbindung der flexiblen Ligandpositionierung erfolgt bei TransCent direkt in den Optimierungsprozess. Prinzipiell gibt es auch die Möglichkeit, jede Ligandposition unabhängig zu behandeln um dann alle Designmodelle zusammenzufassen und gemäß der erreichten TransCent-Energie zu bewerten. Diese Variante ist aber äußerst aufwändig, da für jede Position ein kompletter Designvorgang durchgeführt werden muss. Bei TransCent wird stattdessen das rotamerbasierte Designkonzept erweitert und die einzelnen Ligandpositionen werden wie Rotamere einer zusätzlichen „Pseudo-Proteinposition“ behandelt. Die besten Ligandpositionen werden dadurch völlig analog zu den Aminosäurerotameren anhand der zugehörigen Energiewerte in der *Simulated Annealing* Optimierung ermittelt.

Daraus ergeben sich zwei unmittelbare Konsequenzen. Da der Wechsel einer Ligandposition i.A. mit einer vergleichsweise hohen Energiedifferenz verbunden ist, sind Positionswechsel nur bei relativ hohen Pseudotemperaturen wahrscheinlich. Unterhalb eines gewissen Schwellwerts ist der Designalgorithmus also mehr oder minder auf eine Ligandposition festgelegt woraufhin die Wechselwirkungen zwischen Ligand und Enzym optimiert werden können.

Außerdem können die Interaktionsenergien sowohl für das DSX- wie auch das Fingerprint-Modul bereits vor der Optimierung berechnet werden. Dies wirkt sich positiv auf die benötigte Rechenzeit je Designoptimierung aus und ist aufgrund der Verwendung aufwändig zu berechnender Mehr-Körper-Energieterme fast unumgänglich. Erkauft wird dies allerdings durch eine Einschränkung der Flexibilität, was die Behandlung des Ligandmoleküls betrifft. Denn anders als bei Dockingverfahren oder dem Enzymdesignprotokoll von Rosetta [66] sind bei TransCent kleine Anpassungen der Ligandposition oder -konformation während der Optimierung nicht mehr möglich. Umso wichtiger ist es, bei der Auswahl der TransLig-Vorhersagen auf eine breite Abdeckung der realisierbaren Protein-Ligand-Geometrien zu achten.

5.2 Möglichkeiten und Grenzen des Designalgorithmus

Neben der Erweiterung um die Fähigkeit zur flexiblen Ligandpositionierung wurden weitere Modifikationen an TransCent vorgenommen. Dabei blieb der grundsätzliche modulare Aufbau des Programms allerdings unangetastet. Die Einflüsse der einzelnen Module beim Designprozess werden in der TransCent-Energiefunktion zusammengefasst und durch die Modulgewichte aufeinander abgestimmt. Die gewichtete Summation der Einzelenergien stellt in gewisser Weise die Situation in natürlichen Enzymen nach. Auch bei diesen existieren konkurrierende Anforderungen an die Positionen im Protein, welche durch den Prozess der natürlichen Selektion ausgeglichen werden. Die „Gewichtung“ erfolgt in diesem Fall in Abhängigkeit von den Umweltbedingungen, so ist z.B. Proteinstabilität bei hyperthermophilen Organismen ein weit wichtigeres Kriterium als bei mesophilen.

5.2.1 Wahrung der Proteinstabilität: das Rosetta-Modul

Das Modul für Proteinstabilität basiert auf der Rosetta Software Suite, einem Programmpaket welches zu den erfolgreichsten und leistungsfähigsten Verfahren auf dem Gebiet des Proteindesigns zählt. Es wird fortlaufend von mehreren Arbeitsgruppen weiterentwickelt und hat seit der Fertigstellung der letzten Version von TransCent einen Versionssprung von Rosetta++ auf Rosetta3 vollzogen. Um von den Neuerungen profitieren zu können, wurde folglich auch das Rosetta-Modul von TransCent aktualisiert. Die Umstellung auf die neue Energiefunktion *score12* führt zu einer signifikanten Verbesserung der Rekapitulationsrate bei den Designergebnissen für den Testdatensatz von 29,5 auf 33,4%. Dieser Wert liegt im Bereich der bereits publizierten Ergebnisse, die mit Rosetta erzielt werden konnten (vgl. [139, 168, 88]). Diese reichen in [139] von 22,7% für die Proteinoberfläche über 32,1% für den Grenzbereich bis hin zu einer Rekapitulationsrate von 49,5% im Proteininneren. Bei den mutierbaren Residuen des TransCent-Testdatensatzes handelt es sich um die Positionen in den aktiven Zentren der Enzyme und deren unmittelbare Nachbarschaft, weshalb ein Vergleich mit der Rate für den Grenzbereich zwischen Oberfläche und Proteininnerem angemessen erscheint.

Betrachtet man die Kombination aller vier Module, so fällt die Steigerung der Rekapitulationsrate mit 0,6 Prozentpunkten weit geringer aus (vgl. 4.2.3.2). Dies bedeutet, dass ein Teil der Veränderungen, welche zu einer Verbesserung der Rosetta-Energiefunktion geführt haben, redundant sind. Die Effekte, welche zusätzlich in *score12* berücksichtigt werden, sind größtenteils bereits durch eines der anderen drei Module abgedeckt.

Dennoch kann das Modul für die Proteinstabilität, dessen Energiefunktion für die Vorhersage stabiler Proteine optimiert wurde, die anderen drei Module nicht ersetzen. Gerade bei katalytischen Residuen und bindungsvermittelnden Proteinpositionen entspricht die Wahl der Aminosäure üblicherweise einem Kompromiss zwischen Stabilität und Funktionalität [77], so dass deren Einfluss auf das Enzym oft eher destabilisierend ist. Außerdem kann die Packungsdichte im aktiven Zentrum eines Enzyms nicht so hoch sein wie im Proteininneren, da sonst keine Ligandenbindung möglich wäre. Überdies sind dort häufig Wasserstoffbrückendonoren und -akzeptoren ohne entsprechenden Partner vorhanden, was sich negativ auf die Proteinstabilität auswirkt, die andererseits aber notwendig sind, um mit dem Ligandmolekül zu interagieren. Folglich führt ein Rekapitulationsdesign unter ausschließlicher Verwendung des Rosetta-Moduls dazu, dass das aktive Zentrum mit

großen Seitenketten besetzt wird um die Packungsdichte zu erhöhen, was sich in der relativ niedrigen Rekapitulationsrate niederschlägt. Dies zu verhindern ist unter anderem Aufgabe des DSX-Moduls, das für die Ligandenbindung sorgt.

5.2.2 Optimierung der Ligandenbindung: das DSX-Modul

Beim Modul für die Optimierung der Ligandenbindung wurde ebenfalls ein Versionswechsel durchgeführt. Es basiert nun auf einer modifizierte Version des Programms DSX [100], dem Nachfolger von DrugScore. Für die Verwendung im Rahmen von TransCent musste es wie sein Vorgänger für die Bewertung einzelner Rotamere angepasst werden. Wie schon bei DrugScore umfasst die Energiefunktion auch hier nur die abstandsabhängigen Paarpotentiale der DSX-Scoringfunktion. Der neu eingeführte Torsionswinkel-Score könnte zwar mit in die Berechnung einbezogen werden, da aber beim Design mit TransCent bislang nur eine Ligandkonformation verwendet wird, würde daraus lediglich eine konstante Verschiebung der Gesamtenergie für ein Modell resultieren. Diese wäre für die Bewertung der Designergebnisse irrelevant.

Die Desolvatationseffekte, welche bei der Bindung des Liganden auftreten, werden energetisch durch die SR-Potentiale erfasst. Allerdings muss bei der Berechnung immer der gesamte Enzym-Ligand-Komplex bzw. dessen Lösungsmittel-zugängliche Oberfläche verwendet werden. Es handelt sich also um eine Mehr-Körper-Energie, welche nicht einfach in eine Rotamerenergie zerlegt werden kann. Andererseits ist eine aufwändige Neuberechnung bei jedem *Simulated Annealing*-Schritt nicht praktikabel. Es existieren aber algorithmische Konzepte mit denen die Auswirkungen eines Rotamerwechsels auf die Solvatationsenergie näherungsweise erfasst werden können, wenn keine Heteroatome anwesend sind [169, 170, 139]. Die Adaption eines dieser Ansätze in Kombination mit den SR-Potentialen von DSX könnte eine Möglichkeit zur Verbesserung der TransCent-Energiefunktion darstellen.

In der aktuellen Version verfügt DSX über Paarpotentiale für 300 verschiedene Kontakttypen (PDB-Potentiale). Dadurch erlaubt das Programm eine detailliertere und präzisere Bewertung der Wechselwirkungen zwischen Ligand und Protein als DrugScore. Die bedeutsamere Änderung ist aber die Einführung der Gauß-Korrektur der Potentiale für kurze Abstände. Bei der neuen TransCent-Version kommt dem DSX-Modul nicht nur die Aufgabe zu, die Wahl der Seitenketten für die Bindung des Liganden zu optimieren. Da die Ligandposition nicht mehr fest vorgegeben ist, leistet es auch einen wichtigen Beitrag bei der Auswahl der passenden Ligandposition. Die Ergebnisse der Rekapitulationsrechnungen zeigen, dass DrugScore und DSX eine ähnliche Performanz aufweisen, sowohl was die Ligandpositionierung als auch die erzielten Rekapitulationsraten angeht. Der Wechsel ist also gerechtfertigt, da mit DSX aufgrund der Gauß-Korrektur Artefakte bei der Positionswahl verhindert werden können (vgl. 4.1.2). Bei Rekapitulationsdesigns kommt dies weniger zum Tragen, weil dabei die eindeutige Existenz einer optimalen Ligandposition gesichert ist. TransCent wurde allerdings nicht für die Rekapitulation von Enzymen sondern für die Übertragung enzymatischer Funktionen entwickelt. In diesen Fällen ist *a priori* nicht klar, wie der Ligand am günstigsten positioniert werden muss, wodurch der korrekten Platzierung des Liganden noch mehr Bedeutung zukommt.

5.2.3 Design spezifischer Enzym-Ligand-Wechselwirkungen: das Fingerprint-Modul

Im Allgemeinen reicht es beim Design von Enzymen nicht aus, stabile Proteine zu erzeugen, welche mit hoher Affinität das jeweilige Ligandmolekül binden. Zusätzlich zu diesen notwendigen Voraussetzungen müssen die Residuen im aktiven Zentrum sehr spezifische Anforderungen erfüllen, welche charakteristisch für die entsprechende Enzymfunktion sind. Deren Art und relative Ausrichtung zum Liganden wird durch die Fingerprint-Potentiale bzw. die dazugehörenden Ellipsoide definiert. Enthält ein Designmodell genau eine passende Seitenkette, welche ein Potential erfüllen kann, so wird dies energetisch belohnt.

Bestraft wird hingegen nicht nur das Fehlen einer Interaktion sondern auch das „Übererfüllen“ eines Potentials durch mehrere, üblicherweise benachbarte Rotamere. Somit sorgt das Fingerprint-Modul nicht nur dafür, dass die katalytisch wichtigen Residuen optimal im aktiven Zentrum platziert werden, sondern implizit auch dafür, dass in deren räumlicher Nachbarschaft unterstützende Mutationen möglich sind. Die Besetzung der benachbarten Positionen wird, je nachdem ob sie sich eher dazu eigen, die pK_a-Werte in der Umgebung einzustellen, oder die Stabilität des Proteins zu erhöhen, durch das PROPKA- bzw. Rosetta-Modul bestimmt.

Die Entwickler von Rosetta verfolgen eine andere Strategie, um die katalytisch wichtigen Seitenketten optimal relativ zum Ligandmolekül zu platzieren. Sowohl die Art der Aminosäuren als auch deren Positionen im Protein müssen vor der Designoptimierung festgelegt werden [171]. Um die korrekte Orientierung der katalytischen Residuen zum Liganden zu garantieren, werden für jede Seitenkette Winkel- und Abstandsbedingungen definiert, welche bei der Optimierung eingehalten werden müssen [69]. Die *a priori* Festlegung des katalytischen Motivs führt allerdings dazu, dass für mehrere Motiv-Variationen Designmodelle berechnet werden müssen, um den Suchraum der möglichen Lösungen ausreichend abzudecken.

Das von Lassila *et al.* vorgestellte Verfahren [172], welches auf dem Designalgorithmus FASTER [61] basiert, verwendet vom Benutzer vorgegebene Straf- bzw. Bonus-Energien, um sicherzustellen, dass die katalytisch wichtigen Aminosäuren korrekt modelliert werden. Diese Energien werden zu den Zwei-Körper-Energietermen addiert, welche die Wechselwirkungen zwischen Ligand und Seitenketten beschreiben. Dadurch wird die Modellierung spezifischer Protein-Ligand-Interaktionen energetisch begünstigt. Die Strategie wird bei der Identifikation geeigneter Ligandpositionen und beim Design selbst eingesetzt. Die veröffentlichten Ergebnisse für die Rekapitulation von drei Enzymen zeigen [172], dass dabei sowohl Übergangszustandsmodelle also auch Ligandkonformationen aus Kristallstrukturen verwendet werden können. Allerdings bleibt auch hier das Problem, dass die Positionen, welche Straf- bzw. Bonus-Energien erhalten sollen, im Vorhinein definiert werden müssen.

Die Ergebnisse der Rekapitulationsdesigns für unterschiedliche Modulkombinationen haben gezeigt, dass die Kombination aus Fingerprint- und Rosetta-Modul annähernd so gut abschneidet wie alle vier Module zusammen. Dies unterstreicht den Einfluss des Fingerprint-Moduls auf das Designergebnis, verdeutlicht aber auch, dass dessen Energiebeitrag sorgfältig auf die anderen Energieterme abgestimmt werden muss.

Das Fingerprint-Modul bewertet ausschließlich, ob ein Rotamer einer der erlaubten Aminosäuren entspricht und ob ein Donor- bzw. Akzeptoratom für die auszubildende Wasserstoffbrücke korrekt platziert wird. Dabei werden Wechselwirkungen mit anderen Atomen

zunächst völlig ignoriert. Prinzipiell können dadurch auch Rotamere energetisch begünstigt werden, welche aufgrund sterischer Kollisionen z.B. mit dem Proteinrückgrat gar nicht erlaubt sind, so lange sie die geforderten Bedingungen des Fingerprint-Potentials erfüllen. In solchen Fällen greift das Rosetta-Modul als Korrektiv ein, um sicherzustellen, dass nur Designmodelle mit physikochemisch sinnvollen Eigenschaften erstellt werden.

Einen Beleg für diesen Effekt liefert die Gridsuche zur Optimierung der TransCent-Modulgewichte. Der höchste mittlere PSSM-Score wird für eine Gewichtskombination erreicht, bei der das Fingerprint-Modul alle anderen dominiert ($\omega_{\mathrm{DSX}} = 0,25$, $\omega_{\mathrm{Fingerprint}} = 8,0$, $\omega_{\mathrm{PROPKA}} = 0,0625$). Es kann somit die anderen Module jederzeit „überstimmen“ und die Platzierung der entsprechenden Seitenketten durchsetzen, ohne Rücksicht auf mögliche negative Auswirkungen auf etwa die Proteinstabilität nehmen zu müssen. Bei der Wahl der Modulgewichte wurde diesem Umstand Rechnung getragen und das Fingerprint-Gewicht auf $\omega_{\mathrm{Fingerprint}} = 1,0$ gesetzt.

Ähnliches gilt bei der Positionierung des Liganden. Auch hier haben die Beiträge des Fingerprint-Moduls einen entscheidenden Einfluss, wie die RMSD-Werte der Rekapitulationsrechnungen mit flexiblem Liganden zeigen (vgl. Tabelle 4.3). In diesem Fall greift das DSX-Modul steuernd ein und verhindert Artefakte bei der Wahl der Ligandposition. Das Zusammenspiel der beiden Module sorgt für eine hervorragende Performanz bei der Erkennung der korrekten Orientierung zwischen Protein und Ligand, was der mittlere RMSD-Wert der gewählten Positionen von nur 0,02 Å belegt. Die Ergebnisse der Funktionsübertragungen lassen zwar keine statistisch signifikante Aussage über die Qualität der verwendeten Ligandpositionen zu, liefern aber doch Hinweise dafür, dass auch bei der Übertragung einer Enzymfunktion auf ein anderes Proteinrückgrat geeignete Positionen gefunden werden, welche das Design wichtiger Wechselwirkungen zwischen Enzym und Ligandmolekül ermöglichen.

An den Berechnungsroutinen des Fingerprint-Moduls wurden nur kleine Änderungen im Vergleich zur alten Version von TransCent vorgenommen. Das Referenzmodell für die Berechnung der Fingerprint-Scores wurde verfeinert, indem die angenommene Gleichverteilung der Aminosäuren durch die tatsächlich beobachteten Aminosäurehäufigkeiten aus der Swiss-Prot Datenbank [137] ersetzt wurden (siehe 3.4.2.2). Außerdem wurde die Methode zur Ableitung der Ellipsoid-Definitionen aus den Punktwolken der wechselwirkenden Schweratome überarbeitet, so dass der Bereich in dem die Donor- bzw- Akzeptoratome platziert werden sollen, noch präziser beschrieben wird (vgl. 4.1.3).

Ungelöst bleibt dadurch allerdings das Problem, welches sich ergibt, wenn die Verteilung in den Punktwolken nur unzureichend mit einer dreidimensionalen Normalverteilung angenähert werden kann. In vielen Fällen eignen sich die Ellipsoide gut für die Erfassung der räumlichen Anforderungen eines Fingerprint-Potentials. Gelegentlich wäre aber eine differenziertere Beschreibung notwendig, z.B. wenn eine Punktwolke zwei Häufungszentren besitzt (siehe Abb. 5.1). In diesen Fällen könnte beispielsweise mit einer Kombination mehrerer Gaußfunktionen eine genauere Beschreibung erzielt werden.

Um eine tatsächliche Neuerung am Fingerprint-Modul handelt es sich bei der Schnittstelle, welche eingebaut wurde, um die Fingerprint-Definitionen ausgeben, einlesen und überarbeiten zu können. Dadurch ist es auch möglich, völlig neue Fingerprints zu entwerfen und zu verwenden, um in der Natur nicht vorkommende Enzymfunktionen auf dem Zielenzym zu etablieren.

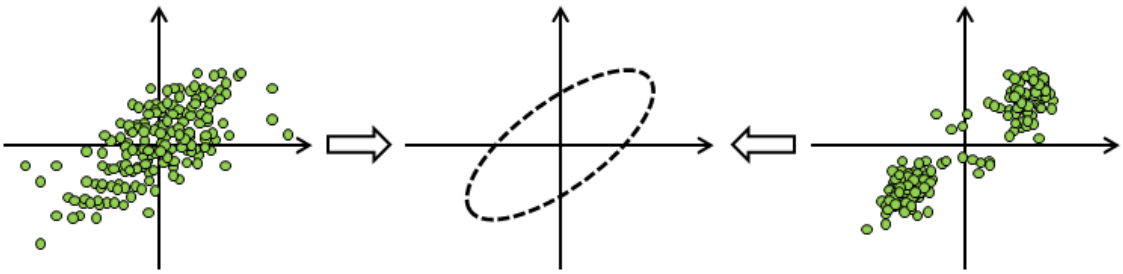


Abbildung 5.1: Ellipsoid-Definition für nicht normalverteilte Punktwolken
Obwohl sich die dargestellten Punktwolken in ihrer räumlichen Verteilung deutlich voneinander unterscheiden, wird für beide dieselbe Ellipsoid-Definition abgeleitet. Dies hat die unerwünschte Folge, dass im rechten wie im linken Fall die Platzierung eines passenden Schweratoms im Zentrum am besten bewertet wird, obwohl es im rechten Fall dort keine entsprechenden Beispiele in der Strukturbibliothek gibt.

5.2.4 Optimierung der Protonierungszustände: das PROPKA-Modul

Das PROPKA-Modul basiert weiterhin auf PROPKA 1.0, welches 2005 publiziert wurde [83], d.h. es wurde unverändert von der alten TransCent-Version übernommen, obwohl eine neuere Version PROPKA 3.1 existiert, welche sich von dieser im Wesentlichen in zwei Punkten unterscheidet. Zum einen wird nun der Einfluss gebundener Ligandmoleküle auf die titrierbaren Gruppen im aktiven Zentrum bei der Berechnung der pK_a-Werte berücksichtigt, wodurch diese dann den tatsächlich katalytisch relevanten Werten entsprechen und den Protonierungszustand beim Ablauf der Katalyse beschreiben. Zum anderen wurde die diskrete Einteilung in exponierte und vergrabene Residuen ersetzt durch ein Berechnungsmodell, welches eine lineare Interpolation der Werte für einen Übergangsbereich zwischen den beiden Zuständen durchführt. Dadurch ist das Programm weniger anfällig für Artefakte, die durch das willkürlichen Setzen einer Grenze zwischen den Zuständen „vergraben“ und „exponiert“ zwangsläufig entstehen. Die durchschnittliche Vorhersagequalität des Algorithmus wurde damit zwar nur unwesentlich verbessert, allerdings konnte die Anzahl grob falscher Werte reduziert werden [173].

Nachdem der Energieterm des PROPKA-Moduls nur Differenzen zwischen den vorhergesagten und den geforderten pK_a-Werten erfasst, sind „falsche“ Vorhersagen beim Design weniger kritisch als bei der pK_a-Wertbestimmung selbst. Im Allgemeinen handelt es sich dabei um systematische Fehler wie z.B. die oben angeführte strikte Unterscheidung zwischen exponiert und vergraben. Da sowohl beim Berechnen des Referenzwertes als auch bei der Designoptimierung derselbe Algorithmus zum Einsatz kommt, werden diese Fehler in der Energiefunktion korrigiert. Dies gilt auf jeden Fall bei Rekapitulationsdesigns und in eingeschränktem Maße auch bei Funktionsübertragungen.

Trotzdem könnte mit der Aktualisierung des PROPKA-Moduls wahrscheinlich eine Verbesserung der Performanz von TransCent erreicht werden, zumal aufgrund der flexiblen Positionierung des Liganden dessen Einfluss auf die pK_a-Werte nicht mehr vernachlässigt werden darf. Dem gegenüber steht der zusätzliche Rechenaufwand, den das komplexere Berechnungsmodell erfordert. Der Umstieg auf eine PROPKA 3.1 basierte Implementierung wäre also nur dann praktikabel, wenn dabei die benötigte Rechenzeit wie bei der ersten Version deutlich reduziert werden könnte.

5.3 Analyse der Stärken und Defizite von TransCent

Um die Performanz von TransCent zu überprüfen, wurden wie schon in [85] Rekapitulationsdesigns auf einem Datensatz mehrerer Enzyme durchgeführt. Da die Anzahl der in der Protein Data Bank hinterlegten Strukturen von hochaufgelösten Enzym-Ligand-Komplexen rasant anwächst, wurde dieser erweitert und umfasst nun 53 Enzyme mit Vertretern aus allen sechs EC-Klassen. Der Datensatz repräsentiert ein breites Spektrum von Ligandmolekülen und Proteingrößen und ist aufgrund seiner Diversität gut geeignet um damit die Leistungsfähigkeit des Designalgorithmus zu bewerten.

Wie sich zeigt, variiert der Designerfolg hinsichtlich der Rekapitulationsrate stark zwischen den einzelnen Enzymen. Während sich das beste Designmodell der Ribonuklease A aus *Bos taurus* (PDB-Code: 1o0h) unter Berücksichtigung der Punktrückmutationen nur noch an drei Stellen vom wildtypischen Enzym unterscheidet, liegt in anderen Fällen die Rekapitulationsrate teilweise deutlich unter 50%.

Insgesamt gilt aber weiterhin, dass trotz der Verbesserungen an der Energiefunktion des Rosetta-Moduls alle Module merklich zur Steigerung der Performanz beitragen und nur mit der Kombination aller vier Module die besten Ergebnisse erzielt werden können. Eine mittlere Rekapitulationsrate von rund 56% belegt aber auch, dass zusätzlich zu Proteinstabilität, Ligandenbindung, Enzymfunktion-spezifischen Wechselwirkungen zwischen Ligand und Protein und Optimierung der Protonierungszustände im aktiven Zentrum noch weitere Bedingungen existieren, welche die Wahl der Aminosäuren beeinflussen.

Mögliche Beispiele wären Anforderungen an die Beweglichkeit von Schleifenregionen (Induced-Fit-Modell, [11]) oder das Vorhandensein einer Phosphorylierungsstelle zur Regulation der Enzymaktivität [174]. Außerdem werden aufgrund des hohen Rechenaufwands Desolvatationseffekte bei der Bindung des Ligandmoleküls ignoriert, obwohl diese eine der entscheidenden Komponenten beim Bindungsprozess sind [175]. Abhilfe könnte eine Kombination der Ansätze aus [100] und [169] schaffen, mit der näherungsweise in angemessener Zeit Desolvatationseffekte in die Energiefunktion mit einbezogen werden könnten.

Die grundlegenden Zusammenhänge, welche beim Design von Enzymen berücksichtigt werden müssen, sind aber bereits in der TransCent-Energiefunktion erfasst. Ein Beleg dafür ist, dass ein deutlicher Zusammenhang zwischen Rekapitulationsrate und Konserviertheit einer Proteinposition beobachtet werden kann. Dem erfolgreichen Design von rund 67% der strikt konservierten Residuen steht eine Rekapitulationsrate von lediglich 23% bei den unkonservierten Positionen gegenüber. Die Designmodelle enthalten also i.A. die katalytisch essentiellen Reste, welche charakteristisch und unbedingt notwendig für die entsprechende Enzymfunktion sind.

Bei Enzymen handelt es sich jedoch um hochoptimierte komplexe Systeme, so dass auch die restlichen Positionen des aktiven Zentrums, korrekt modelliert werden müssen. Diese Residuen, welche z.B. das aktive Zentrum vom Lösungsmittel abschirmen, die Bindung des Liganden unterstützen oder für die korrekte Ausrichtung der katalytischen Seitenketten sorgen, sind erforderlich, damit ein Enzym tatsächlich auf wildtypischem Niveau arbeiten kann. Es existieren offenbar Effekte, die über Erfolg oder Misserfolg des Verfahrens entscheiden, welche aber in der TransCent-Energiefunktion nicht oder nur unzureichend zum Tragen kommen. Dies gilt nicht nur für TransCent sondern scheint allgemein ein Manko der bisher entwickelten Designalgorithmen zu sein, da es noch nicht gelungen ist, Enzyme zu erzeugen, deren Ratenbeschleunigung im Bereich natürlich vorkommender Enzyme liegt [69, 70, 71].

Die Auswertung der Rekapitulationsergebnisse offenbart auch, dass TransCent beim Design der 20 kanonischen Aminosäuren unterschiedlich gut abschneidet. Eklatante Schwächen offenbart der Designalgorithmus bei der korrekten Erkennung von Methionin-Residuen (vgl. Abb. 4.15). Diese sind nicht alleine mit Unzulänglichkeiten der Modellierungseinheit zu erklären (siehe 4.2.3.6) sondern sind auf Defizite der Rosetta-Energiefunktion zurückzuführen (vgl. [168]), welche durch die anderen drei Module nicht entsprechend ausgeglichen werden können. Gleiches gilt für die Schwierigkeiten beim Design von Cystein-Residuen und die Tendenz zur „Verwechslung“ von Phenylalanin und Tyrosin.

Ein anderer Punkt, den es zu beachtet gilt, ist die uneinheitliche Korrelation zwischen TransCent-Energie und Rekapitulationsrate (vgl. 4.2.3.4). Im günstigsten Fall entspricht das Designmodell mit der niedrigsten Energie auch dem, welches am besten mit dem Vorlageenzym übereinstimmt. Im ungünstigen Fall ist die TransCent-Energie als Maß ungeeignet, um gute von weniger guten Modellen zu unterscheiden. Wie die Rechnungen zeigen, hängt der Korrelationskoeffizient stark vom betrachteten Enzym ab. Beim Design einer Funktionsübertragung würde dies bedeuten, dass eine Vielzahl von Modellen tatsächlich umgesetzt werden müsste, um die bestmöglichen Varianten sicher zu erfassen. Beim Rekapitulationsdatensatz liegt der durchschnittliche Rang des besten von 1000 Designmodellen bei über 100. Diese Zahl relativiert sich allerdings in Anbetracht der Tatsache, dass sich die Modelle untereinander i.A. nur wenig unterscheiden. Zudem lässt die Aktivität des besten Modells erwarten, dass auch andere, diesem ähnliche Vorschläge aktiv sind.

Als möglicher Ausweg bietet sich hier die Strategie an, welche von Allen *et al.* [176] erfolgreich bei der Stabilisierung des Proteins Gβ1 aus *Streptococcus sp. GX7805* angewendet wurde. Statt einzelne Modelle umzusetzen und zu testen wurden mehrere unterschiedliche Designmodelle zusammengefasst und in eine Sequenzbibliothek übersetzt, welche ähnlich einem Sequenzprofil beschreibt, welche der erlaubten Aminosäuren an den mutierbaren Positionen der Designmodelle beobachtet wurden. Diese Bibliothek diente dann als Grundlage für einen Hochdurchsatz-Screening-Ansatz in dem die erzeugten Varianten auf Stabilität getestet wurden. Der Designalgorithmus wird dabei quasi als Filter eingesetzt, um für jede Position die ungeeigneten Aminosäuren auszusortieren, wodurch die Anzahl der möglichen Kombinationen auf ein zu bewältigendes Maß reduziert werden kann.

Die Problematik der mangelnden Korrelation dürfte nicht alleine auf Defizite der Energiefunktion zurückzuführen sein, sondern auch mit den Beschränkungen der Modellierungseinheit zusammenhängen. Die Rekapitulationsrechnungen haben gezeigt, dass vor allem bei großen Aminosäuren und vergrabenen Residuen die Rekapitulationsrate deutlich erhöht werden kann, wenn die nativen Seitenketten aus der Kristallstruktur als zusätzliche Rotamere verwendet werden und dadurch eine gründlichere Abtastung des Suchraums simuliert wird. Derselbe Effekt kann bei der Verwendung von Rosetta im Zusammenhang mit Protein-Protein-Docking beobachtet werden. Durch eine abschließende Energieminimierung der Seitenketten nach dem Dockingvorgang, was einer kontinuierlichen Abtastung des Suchraums um die berechnete Lösung entspricht, können korrekte Resultate besser von nicht natürlichen Komplexen unterschieden werden [177]. Dieses Verfahren der Seitenkettenoptimierung wird auch beim Design von Proteinen erfolgreich eingesetzt [69, 178].

Bei einer Energieminimierung werden durch kleine Bewegungen Verspannungen innerhalb einer Modellstruktur beseitigt. Dadurch verbessert sich die Energie der Modelle, obwohl Aminosäureaustausche nicht zugelassen sind. Der Energiegewinn ist bei Wildtyp-ähnlichen Modellen größer als bei unähnlichen. Dadurch verändert sich die Reihenfolge bei der Be-

wertung der Modelle und führt so zu einer Steigerung der Rekapitulationsrate und einer Verbesserung der Korrelation.

Eine weitere Schwäche der Modellierungseinheit ergibt sich aus dem Konzept der Rotamerisierung selbst. Bei den verwendeten Konformationen, die in den Rotamerbibliotheken abgelegt sind, handelt es sich um idealisierte Seitenketten, welche für sich genommen energetisch optimal sind. Im Allgemeinen können damit die tatsächlich beobachteten Seitenkettenkonformationen hinreichend genau angenähert werden. Im Inneren eines Proteins, wo eine hohe Packungsdichte herrscht, und insbesondere bei katalytisch wichtigen Residuen, gelten aber spezielle Anforderungen. So geht es im letzteren Fall nicht darum, die energetisch günstigste Konformation der jeweiligen Seitenkette zu finden, sondern darum, die Gesamtenergie der Kombination von Protein und Übergangszustand des Liganden zu minimieren. Daher können gerade bei katalytisch wichtigen Aminosäuren signifikante Abweichungen von den idealisierten Rotameren auftreten [179]. Einen möglichen Ausweg aus diesem Dilemma haben Gainza *et al.* mit dem Konzept der „kontinuierlichen“ Rotamere vorgestellt [60]. Dabei werden für jeden Torsionswinkel einer Seitenkette nicht nur einige wenige diskrete Werte betrachtet, sondern es wird das Optimum in einem Intervall um den idealisierten Wert bestimmt. Mit diesem Verfahren können bemerkenswerte Rekapitulationsraten von um die 80% erreicht werden. Proteinpositionen, welche mit Liganden oder Kofaktoren interagieren, wurden bei der Evaluation allerdings ausdrücklich ausgeschlossen. Außerdem verwendet der Designalgorithmus ein Dead End Elimination-Optimierungsverfahren, welches ausschließlich auf Ein- und Zwei-Körper-Energieterme anwendbar ist, so dass eine direkte Übertragung des Konzeptes auf TransCent unmöglich wird.

Aufgrund der eben beschriebenen Schwächen und Ungenauigkeiten ist TransCent mit einer durchschnittlichen Rekapitulationsrate von 56% relativ weit vom theoretischen Maximum entfernt. Andererseits schneidet das Programm bei strikt konservierten Residuen deutlich besser ab, so dass sich die Unterschiede zur Referenzsequenz überwiegend auf Positionen beschränken, die auch bei natürlich vorkommenden homologen Enzymen unterschiedlich besetzt sind. Daraus ergibt sich die Frage, wie nahe ein Designprogramm, dessen eigentliches Ziel die Etablierung einer vorgegebenen Enzymfunktion ist, einer perfekten Rekapitulationsrate überhaupt kommen kann bzw. muss.

Ein Versuch, diese Frage zu beantworten, wurde in Form der PSSM-Score-Auswertung für berechnete und wildtypische Enzymsequenzen unternommen. Die Darstellung der Ergebnisse in Abbildung 4.22 zeigt, dass die Designmodelle hinsichtlich Sequenzidentität zum Referenzenzym im Bereich natürlich vorkommender Proteine liegen. Die PSSM-Scores offenbaren jedoch, dass aktuell die Mehrheit der modellierten Enzyme sich noch merklich von diesen unterscheidet. Um in den Bereich Wildtyp-ähnlicher Sequenzen („Homolog typisch“) vorzustoßen, sind demnach Rekapitulationsraten um die 80% notwendig. Eine ähnliche Analyse in [139] kommt zum Schluss, dass die realistische obere Schranke für die erreichbare PSSM-Rekapitulationsrate, also der Anteil an Positionen mit positivem PSSM-Score, bei 80-90% liegt.

Zusammenfassend hat die Bewertung der Performanz von TransCent auf der Basis der Rekapitulationsdesigns ergeben, dass die modellierten Enzyme meist alle wichtigen Elemente enthalten, im Detail aber sichtlich von der Zielvorgabe abweichen. Trotzdem besteht immer noch die Möglichkeit, dass TransCent mit den berechneten Designmodellen Realisierungen der entsprechenden Enzymfunktion gefunden hat, welche zwar möglich sind, so aber in der Natur nicht vorkommen (vgl. [163]). Um dies zu überprüfen wurden insgesamt

fünf Designmodelle von drei unterschiedlichen Funktionsübertragungsdesigns ausgewählt und im Labor auf Stabilität, Ligandenbindung und Enzymaktivität getestet.

5.4 Ein wichtiger Schritt Richtung Aktivität: Bindung

Zur experimentellen Überprüfung der Ergebnisse von TransCent wurde die Isomerisierungsreaktion von Phosphoribosylanthranilat zu CdRP ausgewählt. Dabei handelt es sich um eine Ringöffnungsreaktion, welche von den Enzymen TrpF und PriA katalysiert wird. Die beiden Enzyme *tm*TrpF und *mt*PriA, welche als Vorlagestrukturen dienen, gehören genau wie die Zielstrukturen *tm*HisA und *tm*HisF zur Gruppe der $(\beta\alpha)_8$-Fässer. Sie weisen daher untereinander eine hohe strukturelle Ähnlichkeit auf (vgl. 4.4), wodurch sich die Chance deutlich erhöht, dass im Zielgerüst die essentiellen Seitenketten ähnlich wie in der Vorlage platziert werden können. In Kombination mit der Bestimmung geeigneter Ligandpositionen mittels TransLig ersetzt diese Strategie die aufwändige Suche nach passenden Proteingerüsten mit Verfahren wie RosettaMatch [171], PRODA_MATCH [179] oder ScaffoldSelection [180].

Beide Zielproteine stammen aus dem hyperthermophilen Organismus *Thermotoga maritima* und verfügen daher über eine sehr hohe Ausgangsstabilität. Sie bilden damit eine hervorragende Grundlage für Designexperimente, da sie eine relativ große Zahl von Mutationen tolerieren können, ohne dass sich große Änderungen im Verlauf des Proteinrückgrats ergeben sollten. Da beim Design mit TransCent das Rückgrat des Zielproteins starr gehalten wird, ist dies eine wichtige Eigenschaft. Für *tm*HisF haben Röthlisberger *et al.* bereits nachgewiesen, dass darauf eine fremde Enzymaktivität etabliert werden kann [69].

Für drei der vier möglichen Kombinationen von Ziel- und Vorlageenzym wurden insgesamt fünf Modelle umgesetzt und im Labor untersucht, welche sich durch sechs bis 14 Mutationen vom jeweiligen Zielenzym unterscheiden. Die Ergebnisse der Tests zeigen, dass alle Varianten stabil und gut zu reinigen sind. Allerdings kann unter physiologischen Bedingungen bei keinem der fünf Modelle eine messbare Aktivität festgestellt werden.

Bei den auf *tm*HisF basierenden Varianten scheint dies zunächst etwas überraschend, da bereits gezeigt wurde, dass mit lediglich zwei Austauschen (D130V und D176V) sowohl Bindung als auch katalytische Aktivität auf dem Proteingerüst etabliert werden können [163]. Betrachtet man die entsprechenden Positionen in den Designmodellen, so stellt sich heraus, dass TransCent mit D176A eine relativ ähnliche Mutation vorgeschlagen hat. Für Position 130 wurde in den Modellen die wildtypische Aminosäure beibehalten, da diese zu weit vom Liganden entfernt liegt und somit nicht als mutierbar eingestuft war. Darüber hinaus wurde in [163] gezeigt, dass der katalytische Mechanismus der Doppelmutante nicht dem wildtypischen Mechanismus von TrpF entspricht, so dass dieser auch nicht durch den Fingerprint beschrieben wird.

Die Titrationsmessungen für die Ligandenbindung haben ergeben, dass in drei von fünf Fällen die Bindungsaffinität für rCdRP gegenüber dem wildtypischen Zielenzym signifikant erhöht werden konnte. Während für PA620 und TA247 lediglich eine schwache Bindung ermittelt wurde ist der K_D-Wert von TA893 nur 3-4 mal so hoch wie bei vergleichbaren Enzymen, die darüber hinaus auch Aktivität zeigen (vgl. 4.8). Diese Laborergebnisse stimmen für die meisten Designmodelle bemerkenswert gut mit der *in silico* Analyse mittels DSX überein.

Interessanterweise werden aber gerade die Modelle TF4 und TF148, für die keine spezifische Bindung beobachtet wurde, durch DSX am besten bewertet, TF4 sogar besser als alle betrachteten wildtypischen Enzyme. Dies legt den Schluss nahe, dass es sich bei den Designmodellen zwar um hervorragende Modelle mit optimierter Ligandenbindung handelt, diese jedoch nicht der Wirklichkeit entsprechen. Der beste Weg, um diese Theorie zu untermauern oder zu widerlegen, ist die Aufklärung der Struktur von TF4, welche beim Abschluss dieser Arbeit allerdings noch nicht verfügbar war. Über die Ursachen potentieller Abweichungen kann daher nur gemutmaßt werden. Ein möglicher Grund ist, dass die vereinfachende Annahme, dass sich der Rückgratverlauf durch das Design nicht verändert, sich nur für eine kleine Anzahl von Mutationen als zulässige Näherung erweist. In der Tat liegt mit 13 bzw. 14 die Zahl der Austausche bei TF148 und TF4 deutlich über den neun Mutationen des besten Modells TA893.

Um das Manko des starr gehaltenen Proteinrückgrats beim Design zumindest etwas auszugleichen, wurden daher Punktrückmutationen eingeführt. Austausche, welche wenig oder gar keinen Energiegewinn gegenüber der wildtypischen Aminosäure bringen, werden zurückgenommen. Die Zweckmäßigkeit dieser Strategie wird anhand der Ergebnisse bei den Rekapitulationsrechnungen belegt (vgl. 4.2.3.7). Eine ähnliche Vorgehensweise wurde auch in [181] beim computergestützten Design eines $(\beta\alpha)_8$-Hybrid-Fasses aus Fragmenten zweier unterschiedlicher Proteine angewendet. Dabei wurden von den vorgeschlagenen Mutationen nur so viele eingeführt, dass der mittlere Energiegewinn je Austausch eine gewisse Schwelle nicht unterschritten hat (vgl. Supplement von [181]). In [182] wurden ebenfalls nur solche Mutationen beibehalten, die zusätzlich zum Energiegewinn noch einen weiteren Vorteil gegenüber der wildtypischen Aminosäure versprechen.

Eine weitere mögliche Ursache für Unterschiede zwischen Modell und realem Protein ist, dass der Designprozess ausschließlich unter Anwesenheit des Ligandmoleküls erfolgt. Dadurch wird die Orientierung der Seitenketten im aktiven Zentrum, welche für die Bindung optimal ist, auch energetisch am besten bewertet. Für die *Apo*-Form des Enzyms ohne gebundenen Liganden gelten diese Zwänge hingegen nicht, so dass die Seitenketten eine andere Konformation einnehmen können. Dadurch besteht die Gefahr, dass die Energiebarriere zwischen *Holo*- und *Apo*-Form des Proteins so groß ist, dass die Energie, welche bei der Bindung des Liganden frei wird, nicht ausreicht, um die Residuen in die für die Bindung notwendige Orientierung wechseln zu lassen. Um dieses Problem zu beheben, wird beim Design mit Rosetta die abschließende Energieminimierung (vgl. 5.3) ohne Ligand durchgeführt [66]. Dadurch wird sichergestellt, dass die Seitenketten auch in der *Apo*-Form des Enzyms die korrekte Orientierung beibehalten.

Insgesamt konnte bei drei von fünf Designmodellen statt der angestrebten katalytischen Umsetzung von PRA die Bindung des Produktanalogons rCdRP etabliert werden. Dies kann einerseits bedeuten, dass der Designalgorithmus noch nicht ausgereift ist, um korrekte Vorhersagen zu liefern. Andererseits kann die fehlende Aktivität aber auch darauf zurückzuführen sein, dass das beim Design verwendete Produktanalogon keine hinreichend gute Näherung für den Übergangszustand des Liganden darstellt und somit statt der Stabilisierung des Übergangszustandes die Bindung von rCdRP optimiert wurde. Betrachtet man zum Vergleich die erfolgreich durchgeführten Enzymdesigns von Röthlisberger *et al.* [69], Siegel *et al.* [70] und Jiang *et al.* [71], so wurden diese auf der Basis sogenannter Theozyme erstellt [75]. Dabei handelt es sich um eine Kombination aus einem Modell des Übergangszustandes des Liganden und den Seitenketten der katalytischen Residuen, welche *en bloc* im aktiven Zentrum des Zielproteins platziert wird. Beim nachfolgenden

Optimierungsprozess werden die Positionen in der Umgebung des Theozyms besetzt, um dieses bestmöglich zu stabilisieren.

Dadurch ist dieses Designprotokoll auf Funktionen beschränkt, für die ein passendes Theozym berechnet werden kann. Dies ist bislang nur für sehr einfache Reaktionen möglich und setzt die Kenntnis des genauen enzymatischen Mechanismus voraus. Im Gegensatz dazu können mit TransCent prinzipiell auch Übertragungen von Enzymfunktionen modelliert werden, bei denen der exakte Mechanismus bislang noch nicht bekannt ist.

Andererseits ist auch das Design von Ligandenbindung kein Problem, welches bislang für den allgemeinen Fall befriedigend gelöst werden konnte [183]. Selbst die erfolgreich künstlich erzeugten Enzyme weisen verglichen mit natürlich vorkommenden Enzymen eher bescheidene Bindungsaffinitäten im millimolaren oder hohen mikromolaren Bereich auf [69, 70, 71]. TransCent in seiner jetzigen Form könnte daher statt zur Übertragung von Aktivität als Werkzeug für das Design von Ligandenbindung verwendet werden.

6 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde das Enzymdesignprogramm TransCent erweitert, so dass beim Designprozess das Ligandmolekül flexibel im aktiven Zentrum des Zielproteins platziert werden kann. Für die Berechnung potentiell geeigneter Ligandpositionen wurde das Modul TransLig entwickelt. Darüber hinaus wurden drei der vier bestehenden Module von TransCent weiterentwickelt bzw. durch neuere Versionen ersetzt.

Anhand von Rekapitulationsdesigns auf einem neu zusammengestellten Datensatz konnte gezeigt werden, dass durch die Überarbeitung der Module die Performanz des Designalgorithmus verbessert wurde. Außerdem belegen die Ergebnisse, dass die Kombination von TransCent und TransLig in der Lage ist, nativ-ähnliche Ligandpositionen zu erzeugen und darauf aufbauend Designmodelle zu erstellen. Dadurch eröffnet sich auch ein neues Feld von Anwendungsmöglichkeiten, da die Übertragung von Enzymfunktionen nicht mehr grundsätzlich auf Proteingerüste mit sehr hoher struktureller Ähnlichkeit beschränkt ist.

Zur experimentellen Überprüfung der Vorhersagen von TransCent wurden insgesamt fünf Designmodelle für die Übertragung der PRA-Isomerase-Aktivität umgesetzt und getestet. Für drei der fünf Proteine, welche alle eine zufriedenstellende Stabilität aufweisen, konnte die Bindungsaffinität für das Produktanalogon rCdRP deutlich gegenüber dem Ausgangskonstrukt erhöht werden. Enzymatische Aktivität war allerdings bei keinem der Modelle messbar.

Die größten Aussichten auf Erfolg bei der Suche nach den Ursachen für das Fehlen von Aktivität verspricht eine detaillierte Analyse der Kristallstrukturen der erzeugten Proteine. Die Unterschiede zwischen den Designmodellen und den tatsächlichen Strukturen sollten Anhaltspunkte für den unterschiedlichen Erfolg der Designs liefern. Zum Zeitpunkt der Fertigstellung dieser Arbeit waren jedoch die Bemühungen zur Aufklärung der Struktur bei keinem der Proteine erfolgreich abgeschlossen. Zusätzliche wertvolle Hinweise könnten Moleküldynamik-Simulationen basierend auf den Designmodellen oder den Kristallstrukturen liefern (vgl. [50]), weil damit auch dynamische Effekte wie z.B. die Beweglichkeit bestimmter Seitenketten analysiert werden können.

Darüber hinaus könnte die Struktur des Protein-Ligand-Komplexes für das am besten bindende Modell TA893 als Ausgangspunkt für eine neue Designrunde dienen. Ähnlich wie bei Privett *et al.* könnte damit ein iterativer Prozess zur Verfeinerung der Designmodelle gestartet werden [51], um zunächst die Bindungsaffinität zu optimieren und schlussendlich Aktivität zu etablieren.

Ebenso könnten die Designmodelle die Basis für weiterführende Mutationsstudien bilden. Wie Khersonsky *et al.* zeigen konnten, ist es möglich, durch Zufallsmutagenese die enzymatische Aktivität eines am Computer modellierten Enzyms nochmals deutlich zu erhöhen [72, 74]. Allerdings verfügten dabei bereits die Designmodelle über eine schwache aber messbare Enzymaktivität.

Die Erkenntnisse aus dem Labor könnten und sollten auch dazu verwendet werden, die Defizite des Designalgorithmus aufzudecken um diesen anschließend weiterzuentwickeln.

Einige potentielle Ansatzpunkte zur Verbesserung von TransCent wurden bereits im letzten Kapitel angesprochen. Eine mögliche Strategie wäre, die bereits existierenden Module noch weiter zu verfeinern. So könnte z.B. das PROPKA-Modul überarbeitet und durch eine Adaption der aktuellen PROPKA-Version 3.1 ersetzt werden. Dadurch würde nicht nur die Vorhersagequalität des Moduls verbessert, sondern auch der Ligand bei der Berechnung der pK_a-Werte berücksichtigt.

Auch das Fingerprint-Modul bietet weitere Optimierungsmöglichkeiten. So könnten die Fingerprint-Definitionen um π-π-Wechselwirkungen oder Interaktionen zwischen Ligand und Proteinrückgrat erweitert werden. Bereits angesprochen wurde eine mögliche Verbesserung bei der Beschreibung der abgeleiteten Punktwolken, welche über die Verwendung von dreidimensionalen Gaußverteilungen hinaus geht. Aussichtsreicher als eine Optimierung bereits bestehender Module erscheint jedoch die Erweiterung des Designalgorithmus um zusätzliche Funktionalitäten. Die größte Einschränkung welche beim Design mit TransCent hingenommen werden muss, ist, dass das Proteinrückgrat während des Designs starr gehalten wird. Je mehr Mutationen für die Übertragung einer Enzymfunktion in das Zielprotein eingeführt wurden, desto eher kann diese Näherung zu Designartefakten führen, so möglicherweise geschehen bei den Varianten TF4 und TF148. Es wäre also überaus wünschenswert, wenn zumindest kleine Anpassungen des Rückgratverlaufs beim Design erlaubt wären. Ein vorstellbarer Weg, um dies zu erreichen, wäre die „Verschmelzung“ von TransCent mit dem Rosetta-Softwarepaket. Die Einbettung der TransCent-Module in das System von Rosetta würde den Zugriff auf eine breite Palette zusätzlicher Funktionen ermöglichen, vor allem vor dem Hintergrund, dass eines der Module ja bereits auf der Rosetta-Energiefunktion basiert. Dadurch könnte auch eine abschließende Energieminimierung der Designmodelle durchgeführt werden, was potentiell die Korrelation von TransCent-Energie und Modellqualität verbessern würde. Diese Vereinigung widerspräche zwar der allgemeinen Philosophie der Rosetta-Entwickler, sich auf Ein- und Zwei-Körper-Energien zu beschränken, könnte sich aber als notwendig erweisen, um die Qualität der Vorhersagen noch weiter steigern zu können.

Eine weitere Näherung, welche beim Design mit TransCent bisher gemacht wurde, ist die Verwendung der Struktur des Substrat-, Produkt- oder Übergangszustandsanalogons aus dem in der PDB abgelegten Enzym-Ligand-Komplex als Ersatz für ein Modell des Übergangszustandes. Ähnlich wie beim starr gehaltenen Proteinrückgrat könnte sich diese Näherung als unzureichend erweisen. Um diese Fragestellung zu klären, könnten Designmodelle für eine Enzymfunktion mit bekanntem Übergangszustand berechnet und getestet werden.

Darüber hinaus wurden bislang innere Freiheitsgrade des Ligandmoleküls nicht berücksichtigt, obwohl dies prinzipiell möglich wäre. Aufgrund der begrenzt verfügbaren Rechenleistung wird stattdessen die vereinfachende Annahme gemacht, dass es sich bei der Konformation in der Kristallstruktur um die katalytisch aktive Konformation des Liganden handelt. Eine Aufhebung dieser Beschränkung könnte zu einer verbesserten Modellqualität führen, müsste aber, wie fast alle aufgeführten Erweiterungen und Verbesserungen mit einer signifikanten Erhöhung des Rechenaufwands erkauft werden. Die Entwicklung der verfügbaren Rechenleistung sollte aber dazu führen, dass in Zukunft in gleicher Zeit immer aufwändigere Berechnungen möglich werden.

Zusammenfassend kann festgehalten werden, dass im Enzymdesignprogramm TransCent noch Entwicklungspotential steckt. Trotz der genannten Limitationen könnte TransCent

aber dennoch bereits in der Lage sein, erfolgreich Enzymaktivitäten zwischen verschiedenen Proteingerüsten zu übertragen. Grund zu dieser Annahme gibt die Feststellung, dass die Erfolgsquote auch bei den bislang erfolgreich durchgeführten Enzymdesigns mit Rosetta sehr niedrig ist. Von den 84 getesteten Designs für die Katalyse der Diels-Alder-Reaktion waren lediglich zwei aktiv [70]. Auch die bislang unerreicht hohe Erfolgsquote von 75% beim Design von Retro-Aldolasen durch Althoff *et al.* bedeutet, dass auch in diesem Fall für neun der 42 Enzymmodelle keine messbare Aktivität festgestellt werden konnte [182]. Um also die Hypothese zu bestätigen oder zu widerlegen, müssten noch weitere Designmodelle auch für zusätzliche Zielproteine berechnet, umgesetzt und getestet werden.

Die vorgestellten experimentellen Ergebnisse lassen aber den Schluss zu, dass TransCent in seiner jetzigen Form für das Design von Ligandenbindung verwendet werden kann.

Danksagung

An dieser Stelle möchte ich all denen meinen Dank aussprechen, die das Gelingen dieser Arbeit erst möglich gemacht haben.

Allen voran möchte ich meinem Doktorvater Prof. Dr. Rainer Merkl für die ausgezeichnete Betreuung dieser Arbeit danken. Er hatte stets ein offenes Ohr für meine Fragen und ich konnte dank zahlreicher spannender und hilfreicher Diskussionen viel von seiner langjährigen Erfahrung profitieren.

Ein besonderer Dank gilt Prof. Dr. Reinhard Sterner und Prof. Dr. Stephan Waack dafür, dass sie mich mit ihrem großen Erfahrungsschatz als Mentoren unterstützt und mir interessante und lehrreiche Einblicke in die Welt der Biochemie und der Algorithmen gewährt haben.

Prof. Dr. Elmar Lang möchte ich für die Übernahme der Funktion des Zweitgutachters danken und dafür, dass er mir die Tür zur Bioinformatik geöffnet hat.

Während meines Forschungsaufenthalts an der Vanderbilt University habe ich viel über das Leben und Forschen in den USA gelernt. Prof. Dr. Jens Meiler gilt mein besonderer Dank, dass er mir diese Möglichkeit geboten hat. Bei Sam Deluca bedanke ich mich für die Hilfe bei der Anpassung von Rosetta.

Friedemann Paulini danke ich für die spannende und fruchtbare Zusammenarbeit während seiner Diplomarbeit und viele interessante Skype-Konferenzen.

Großer Dank gilt auch Dr. André Fischer und Dr. Marco Bocola für die „Starthilfe“ bei der Einarbeitung in mein Projekt, viele weitere wertvolle Ratschläge und für so manche Spätschicht.

Meinen Zimmerkollegen Dr. Hermann Zellner und Jan-Oliver Janda und allen ehemaligen Untermietern möchte ich für eine tolle Arbeitsatmosphäre in den letzten Jahren, viele interessante und lehrreiche Gespräche und spannende Diskussionen über die hohe Kunst des Lupfens danken.

Ein besonderer Dank gilt auch meinem „Haus & Hof-Biochemiker“ Bernd Reisinger für die experimentelle Überprüfung meiner Vorhersagen und die geduldige Beantwortung unzähliger biochemischer Fragen.

Allen Sterners möchte ich für das angenehme und anregende Arbeitsklima und die große Hilfsbereitschaft bei biochemischen Problemen danken. Und für viele aufregende Kickerduelle.

Außerdem danke ich Dr. Vera Lechner für wertvolle Tipps bei der Gestaltung der Abbildungen und ihr geduldiges Ohr in mühsamen Momenten.

Großer Dank gebührt auch meiner ganzen Familie, insbesondere meinen Eltern und meinem Bruder, für die uneingeschränkte Unterstützung während der letzten Jahre.

Schließlich möchte ich meiner Freundin Caro für die letzten vier Jahre danken, für die Geduld, die Unterstützung, die Aufmunterungen zur rechten Zeit und ihre große Zuneigung.

Außerdem möchte ich diese Gelegenheit nutzen, ein vier Jahre währendes Versäumnis nachzuholen: Danke Caro!

Literaturverzeichnis

[1] Kirk, O., T. V. Borchert und C. C. Fuglsang: *Industrial enzyme applications.* Current Opinion in Biotechnology, 13(4):345–51, August 2002.

[2] Lopez-Gallego, F. und C. Schmidt-Dannert: *Multi-enzymatic synthesis.* Current Opinion in Chemical Biology, 14(2):174–83, April 2010.

[3] Bornscheuer, U. T., G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore und K. Robins: *Engineering the third wave of biocatalysis.* Nature, 485(7397):185–94, Mai 2012.

[4] Baker, D.: *An exciting but challenging road ahead for computational enzyme design.* Protein science, 19(10):1817–9, Oktober 2010.

[5] Barrozo, A., R. Borstnar, G. Marloie und S. C. L. Kamerlin: *Computational protein engineering: Bridging the gap between rational design and laboratory evolution.* International Journal of Molecular Sciences, 13(10):12428–12460, September 2012.

[6] Fleishman, S. J. und D. Baker: *Role of the biomolecular energy gap in protein design, structure, and evolution.* Cell, 149(2):262–73, April 2012.

[7] Kluge, F. und E. Seebold: *Etymologisches Wörterbuch der deutschen Sprache.* de Gruyter, 24., durchges. und erw. Auflage, 2002.

[8] Wolfenden, R. und M. J. Snider: *The depth of chemical time and the power of enzymes as catalysts.* Accounts of Chemical Research, 34(12):938–45, Dezember 2001.

[9] Radzicka, A. und R. Wolfenden: *A proficient enzyme.* Science, 267(5194):90–3, Januar 1995.

[10] Lichtenthaler, F. W.: *Hundert Jahre Schlüssel-Schloß-Prinzip: Was führte Emil Fischer zu dieser Analogie?* Angewandte Chemie, 106(23-24):2456–2467, Dezember 1994.

[11] Koshland, D. E.: *Application of a theory of enzyme specificity to protein synthesis.* Proceedings of the National Academy of Sciences of the United States of America, 44(2):98–104, Februar 1958.

[12] Webb, Edwin C. (Herausgeber): *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.* Academic Press, San Diego, Californien, 1992.

[13] Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi und L. L. Yeh: *UniProt: the Universal Protein knowledgebase.* Nucleic acids research, 32(Database issue):D115–9, Januar 2004.

[14] Murzin, A. G., S. E. Brenner, T. Hubbard und C. Chothia: *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of Molecular Biology, 247(4):536–40, April 1995.

[15] Govindarajan, S., R. Recabarren und R. A. Goldstein: *Estimating the total number of protein folds.* Proteins, 35(4):408–14, Juni 1999.

[16] Gerlt, J. A.: *New wine from old barrels.* Nature Structural Biology, 7(3):171–3, März 2000.

[17] Nagano, N., C. A. Orengo und J. M. Thornton: *One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.* Journal of Molecular Biology, 321(5):741–65, August 2002.

[18] Copley, R. R. und P. Bork: *Homology among* $(\beta\alpha)_8$*-barrels: implications for the evolution of metabolic pathways.* Journal of Molecular Biology, 303(4):627–41, November 2000.

[19] Banner, D. W., A. C. Bloomer, G. A. Petsko, D. C. Phillips, C. I. Pogson, I. A. Wilson, P. H. Corran, A. J. Furth, J. D. Milman, R. E. Offord, J. D. Priddle und S. G. Waley: *Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data.* Nature, 255(5510):609–14, Juni 1975.

[20] Wierenga, R. K.: *The TIM-barrel fold: a versatile framework for efficient enzymes.* FEBS Letters, 492(3):193–8, März 2001.

[21] Sterner, R. und B. Höcker: *Catalytic versatility, stability, and evolution of the* $(\beta\alpha)_8$*-barrel enzyme fold.* Chemical Reviews, 105(11):4038–55, November 2005.

[22] Caetano-Anollés, G., H. S. Kim und J. E. Mittenthal: *The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture.* Proceedings of the National Academy of Sciences of the United States of America, 104(22):9358–63, Mai 2007.

[23] Höcker, B., J. Claren und R. Sterner: *Mimicking enzyme evolution by generating new* $(\beta\alpha)_8$*-barrels from* $(\beta\alpha)_4$*-half-barrels.* Proceedings of the National Academy of Sciences of the United States of America, 101(47):16448–53, November 2004.

[24] Lang, D., R. Thoma, M. Henn-Sax, R. Sterner und M. Wilmanns: *Structural evidence for evolution of the* β/α *barrel scaffold by gene duplication and fusion.* Science, 289(5484):1546–50, September 2000.

[25] Söding, J., M. Remmert und A. Biegert: *HHrep: de novo protein repeat detection and the origin of TIM barrels.* Nucleic Acids Research, 34(Web Server issue):W137–42, Juli 2006.

[26] Richter, M., M. Bosnali, L. Carstensen, T. Seitz, H. Durchschlag, S. Blanquart, R. Merkl und R. Sterner: *Computational and experimental evidence for the evolution of a* $(\beta\alpha)_8$*-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds.* Journal of Molecular Biology, 398(5):763–73, Mai 2010.

[27] Gibson, D. G., J. I. Glass, C. Lartigue, V. N. Noskov, R. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z. Qi, T. H. Segall-Shapiro, C. H.

Calvey, P. P. Parmar, C. A. Hutchison, H. O. Smith und J. C. Venter: *Creation of a bacterial cell controlled by a chemically synthesized genome.* Science, 329(5987):52–6, Juli 2010.

[28] Jürgens, C., A. Strom, D. Wegener, S. Hettwer, M. Wilmanns und R. Sterner: *Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways.* Proceedings of the National Academy of Sciences of the United States of America, 97(18):9925–30, August 2000.

[29] Henn-Sax, M., B. Höcker, M. Wilmanns und R. Sterner: *Divergent evolution of $(\beta\alpha)_8$-barrel enzymes.* Biological Chemistry, 382(9):1315–20, September 2001.

[30] Cadwell, R. C. und G. F. Joyce: *Mutagenic PCR.* PCR Methods and Applications, 3(6):S136–40, Juni 1994.

[31] Arnold, F. H., P. L. Wintrode, K. Miyazaki und A. Gershenson: *How enzymes adapt: lessons from directed evolution.* Trends in Biochemical Sciences, 26(2):100–6, Februar 2001.

[32] Stemmer, W. P.: *DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution.* Proceedings of the National Academy of Sciences of the United States of America, 91(22):10747–51, Oktober 1994.

[33] Crameri, A., S. A. Raillard, E. Bermudez und W. P. Stemmer: *DNA shuffling of a family of genes from diverse species accelerates directed evolution.* Nature, 391(6664):288–91, Januar 1998.

[34] Schmidt-Dannert, C.: *Directed evolution of single proteins, metabolic pathways, and viruses.* Biochemistry, 40(44):13125–36, November 2001.

[35] Reetz, M. T., M. Bocola, J. D. Carballeira, D. Zha und A. Vogel: *Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test.* Angewandte Chemie (International ed. in English), 44(27):4192–6, Juli 2005.

[36] Russ, W. P., D. M. Lowery, P. Mishra, M. B. Yaffe und R. Ranganathan: *Natural-like function in artificial WW domains.* Nature, 437(7058):579–83, September 2005.

[37] Jäckel, C., J. D. Bloom, P. Kast, F. H. Arnold und D. Hilvert: *Consensus protein design without phylogenetic bias.* Journal of Molecular Biology, 399(4):541–6, Juni 2010.

[38] Janda, J., M. Busch, F. Kück, M. Porfenenko und R. Merkl: *CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure.* BMC Bioinformatics, 13:55, Januar 2012.

[39] Tokuriki, N., F. Stricher, L. Serrano und D. S. Tawfik: *How protein stability and new functions trade off.* PLoS Computational Biology, 4(2):e1000002, Februar 2008.

[40] Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard und D. Baker: *Design of a novel globular protein fold with atomic-level accuracy.* Science, 302(5649):1364–8, November 2003.

[41] Dahiyat, B. I.: *In silico design for protein stabilization.* Current Opinion in Biotechnology, 10(4):387–90, August 1999.

[42] Senn, H. M. und W. Thiel: *QM/MM methods for biomolecular systems.* Angewandte Chemie (International ed. in English), 48(7):1198–229, Januar 2009.

[43] Senn, H. M. und W. Thiel: *QM/MM studies of enzymes.* Current Opinion in Chemical Biology, 11(2):182–7, April 2007.

[44] Warshel, A. und F. Sussman: *Toward computer-aided site-directed mutagenesis of enzymes.* Proceedings of the National Academy of Sciences of the United States of America, 83(11):3806–10, Juni 1986.

[45] Frushicheva, M. P., J. Cao, Z. T. Chu und A. Warshel: *Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase.* Proceedings of the National Academy of Sciences of the United States of America, 107(39):16869–74, September 2010.

[46] Alexandrova, A. N., D. Röthlisberger, D. Baker und W. L. Jorgensen: *Catalytic mechanism and performance of computationally designed enzymes for Kemp elimination.* Journal of the American Chemical Society, 130(47):15907–15, November 2008.

[47] Brooks, B.R., R.E. Bruccoleri, D.J. Olafson, D.J. States, S. Swaminathan und M. Karplus: *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 4:187–217, 1983.

[48] Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel und P. Kollman: *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules.* Computer Physics Communications, 91(1-3):1–41, September 1995.

[49] Van Der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark und H. J. C. Berendsen: *GROMACS: fast, flexible, and free.* Journal of Computational Chemistry, 26(16):1701–18, Dezember 2005.

[50] Kiss, G., D. Röthlisberger, D. Baker und K. N. Houk: *Evaluation and ranking of enzyme designs.* Protein science, 19(9):1760–73, September 2010.

[51] Privett, H. K., G. Kiss, T. M. Lee, R. Blomberg, R. Chica, L. M. Thomas, D. Hilvert, K. N. Houk und S. L. Mayo: *Iterative approach to computational enzyme design.* Proceedings of the National Academy of Sciences of the United States of America, 109(10):3790–5, März 2012.

[52] Dunbrack, R. L. und M. Karplus: *Backbone-dependent rotamer library for proteins. Application to side-chain prediction.* Journal of Molecular Biology, 230(2):543–74, März 1993.

[53] Shifman, J. M. und S. L. Mayo: *Modulating calmodulin binding specificity through computational protein design.* Journal of Molecular Biology, 323(3):417–23, Oktober 2002.

[54] Suárez, M., P. Tortosa und A. Jaramillo: *PROTDES: CHARMM toolbox for computational protein design.* Systems and Synthetic Biology, 2(3-4):105–13, Dezember 2008.

[55] Merkl, R. und S. Waack: *Bioinformatik Interaktiv: Grundlagen, Algorithmen und Anwendungen.* WILEY-VCH, Weinheim, 2. Auflage, 2009.

[56] Sippl, M. J.: *Knowledge-based potentials for proteins.* Current Opinion in Structural Biology, 5(2):229–35, April 1995.

[57] Pierce, N. A. und E. Winfree: *Protein design is NP-hard.* Protein Engineering, 15(10):779–82, Oktober 2002.

[58] Desmet, J., M. De Maeyer, B. Hazes und I. Lasters: *The dead-end elimination theorem and its use in protein side-chain positioning.* Nature, 356(6369):539–42, April 1992.

[59] Goldstein, R. F.: *Efficient rotamer elimination applied to protein side-chains and related spin glasses.* Biophysical Journal, 66(5):1335–40, Mai 1994.

[60] Gainza, P., K. E. Roberts und B. R. Donald: *Protein design using continuous rotamers.* PLoS Computational Biology, 8(1):e1002335, Januar 2012.

[61] Desmet, J., J. Spriet und I. Lasters: *Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization.* Proteins, 48(1):31–43, Juli 2002.

[62] Allen, B. D. und S. L. Mayo: *Dramatic performance enhancements for the FASTER optimization algorithm.* Journal of Computational Chemistry, 27(10):1071–5, Juli 2006.

[63] Kirkpatrick, S., C. D. Gelatt und M. P. Vecchi: *Optimization by simulated annealing.* Science, 220(4598):671–80, Mai 1983.

[64] Loksha, I. V., J. R. Maiolo, C. W. Hong, A. Ng und C. D. Snow: *SHARPEN-systematic hierarchical algorithms for rotamers and proteins on an extended network.* Journal of Computational Chemistry, 30(6):999–1005, April 2009.

[65] Chowdry, A. B., K. A. Reynolds, M. S. Hanes, M. Voorhies, N. Pokala und T. M. Handel: *An object-oriented library for computational protein design.* Journal of Computational Chemistry, 28(14):2378–88, November 2007.

[66] Richter, F., A. Leaver-Fay, S. D. Khare, S. Bjelic und D. Baker: *De novo enzyme design using Rosetta3.* PLoS ONE, 6(5):e19230, Januar 2011.

[67] Leaver-Fay, A., M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. W. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker und P. Bradley: *Chapter nineteen - Rosetta3: An object-oriented software suite for the simulation and design of macromolecules.* In: Johnson, M. L. und L. Brand (Herausgeber): *Computer Methods, Part C*, Band 487 der Reihe *Methods in Enzymology*, Seiten 545 – 574. Academic Press, 2011.

[68] Meiler, J. und D. Baker: *ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility.* Proteins, 65(3):538–48, November 2006.

[69] Röthlisberger, D., O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik und D. Baker: *Kemp elimination catalysts by computational enzyme design.* Nature, 453(7192):190–5, Mai 2008.

[70] Siegel, J. B., A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael und D. Baker: *Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.* Science, 329(5989):309–13, Juli 2010.

[71] Jiang, L., E. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard und D. Baker: *De novo computational design of retro-aldol enzymes.* Science, 319(5868):1387–91, März 2008.

[72] Khersonsky, O., D. Röthlisberger, O. Dym, S. Albeck, C. J. Jackson, D. Baker und D. S. Tawfik: *Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series.* Journal of Molecular Biology, 396(4):1025–42, März 2010.

[73] Khersonsky, O., D. Röthlisberger, A. M. Wollacott, P. Murphy, O. Dym, S. Albeck, G. Kiss, K. N. Houk, D. Baker und D. S. Tawfik: *Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution.* Journal of Molecular Biology, 407(3):391–412, April 2011.

[74] Khersonsky, O., G. Kiss, D. Röthlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker und D. S. Tawfik: *Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59.* Proceedings of the National Academy of Sciences of the United States of America, 109(26):10358–63, Juni 2012.

[75] Tantillo, D. J., J. Chen und K. N. Houk: *Theozymes and compuzymes: theoretical models for biological catalysis.* Current Opinion in Chemical Biology, 2(6):743–50, Dezember 1998.

[76] Fischer, A., N. Enkler, G. Neudert, M. Bocola, R. Sterner und R. Merkl: *TransCent: computational enzyme design by transferring active sites and considering constraints relevant for catalysis.* BMC Bioinformatics, 10:54, Januar 2009.

[77] Beadle, B. M. und B. K. Shoichet: *Structural bases of stability-function tradeoffs in enzymes.* Journal of Molecular Biology, 321(2):285–96, August 2002.

[78] Gohlke, H., M. Hendlich und G. Klebe: *Knowledge-based scoring function to predict protein-ligand interactions.* Journal of Molecular Biology, 295(2):337–56, Januar 2000.

[79] Porter, C. T., G. J. Bartlett und J. M. Thornton: *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.* Nucleic acids research, 32(Database issue):D129–33, Januar 2004.

[80] Bartlett, G. J., C. T. Porter, N. Borkakoti und J. M. Thornton: *Analysis of catalytic residues in enzyme active sites.* Journal of Molecular Biology, 324(1):105–121, November 2002.

[81] Harris, T. K. und G. J. Turner: *Structural basis of perturbed pK_a values of catalytic groups in enzyme active sites.* IUBMB Life, 53(2):85–98, Februar 2002.

[82] Forsyth, W. R., J. M. Antosiewicz und A. D. Robertson: *Empirical relationships between protein structure and carboxyl pK_a values in proteins.* Proteins, 48(2):388–403, August 2002.

[83] Li, H., A. D. Robertson und J. H. Jensen: *Very fast empirical prediction and rationalization of protein pK_a values.* Proteins, 61(4):704–21, Dezember 2005.

[84] Sali, A. und T. L. Blundell: *Comparative protein modelling by satisfaction of spatial restraints.* Journal of Molecular Biology, 234(3):779–815, Dezember 1993.

[85] Fischer, A.: *TRANSCENT - ein Enzymdesignprogramm zum Transfer aktiver Zentren unter Wahrung Katalyse-relevanter Rahmenbedingungen.* Doktorarbeit, Universität Regensburg, 2007.

[86] Reisinger, B.: *Computerbasierte und vergleichende Ansätze zum Enzymdesign auf dem $(\beta\alpha)_8$-Barrel Proteingerüst.* Diplomarbeit, Universität Regensburg, 2009.

[87] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller und E. Teller: *Equation of state calculations by fast computing machines.* The Journal of Chemical Physics, 21(6):1087, 1953.

[88] Kuhlman, B. und D. Baker: *Native protein sequences are close to optimal for their structures.* Proceedings of the National Academy of Sciences of the United States of America, 97(19):10383–8, September 2000.

[89] Eijsink, V. G. H., A. Bjø rk, S. Gå seidnes, R. Sirevå g, B. Synstad, B. van den Burg und G. Vriend: *Rational engineering of enzyme stability.* Journal of Biotechnology, 113(1-3):105–20, September 2004.

[90] Rohl, C., C. E. M. Strauss, K. M. S. Misura und D. Baker: *Protein structure prediction using Rosetta.* Methods in Enzymology, 383(2003):66–93, Januar 2004.

[91] Bonneau, R., I. Ruczinski, J. Tsai und D. Baker: *Contact order and ab initio protein structure prediction.* Protein science, 11(8):1937–44, August 2002.

[92] King, N. P., W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. André, T. Gonen, T. O. Yeates und D. Baker: *Computational design of self-assembling protein nanomaterials with atomic level accuracy.* Science, 336(6085):1171–4, Juni 2012.

[93] Dantas, G., B. Kuhlman, D. Callender, M. Wong und D. Baker: *A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins.* Journal of Molecular Biology, 332(2):449–460, September 2003.

[94] Davis, I. W. und D. Baker: *RosettaLigand docking with full ligand and receptor flexibility.* Journal of Molecular Biology, 385(2):381–92, Januar 2009.

[95] Gray, J. J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. Rohl und D. Baker: *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* Journal of molecular biology, 331(1):281–99, August 2003.

[96] Smith, C. und T. Kortemme: *Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design.* PLoS ONE, 6(7):e20451, Januar 2011.

[97] Lazaridis, T. und M. Karplus: *Effective energy function for proteins in solution.* Proteins, 35(2):133–52, Mai 1999.

[98] Marino, M., M. Deuss, D. I. Svergun, P. V. Konarev, R. Sterner und O. Mayans: *Structural and mutational analysis of substrate complexation by anthranilate phosphoribosyltransferase from Sulfolobus solfataricus.* The Journal of Biological Chemistry, 281(30):21410–21, Juli 2006.

[99] Fischer, E.: *Einfluss der Configuration auf die Wirkung der Enzyme.* Berichte der deutschen chemischen Gesellschaft, 27(3):2985–2993, Oktober 1894.

[100] Neudert, G. und G. Klebe: *DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes.* Journal of Chemical Information and Modeling, 51(10):2731–45, Oktober 2011.

[101] Allen, F. H.: *The Cambridge Structural Database: a quarter of a million crystal structures and rising.* Acta Crystallographica. Section B, Structural Science, 58(Pt 3 Pt 1):380–8, Juni 2002.

[102] Neudert, G. und G. Klebe: *fconv: Format conversion, manipulation and feature computation of molecular data.* Bioinformatics, 27(7):1021–2, April 2011.

[103] Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman und R. D. Finn: *The Pfam protein families database.* Nucleic Acids Research, 40(Database issue):D290–301, Januar 2012.

[104] Schiller, L.: *Skriptbasiertes Erzeugen von multiplen Sequenzalignments unter Verwendung webbasierter Dienste.* Bachelorarbeit, FernUniversität Hagen, 2012.

[105] Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko und J. Ye: *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Research, 40(Database issue):D13–25, Januar 2012.

[106] Katoh, K., K. Misawa, K. Kuma und T. Miyata: *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Research, 30(14):3059–66, Juli 2002.

[107] Katoh, K. und H. Toh: *Recent developments in the MAFFT multiple sequence alignment program.* Briefings in Bioinformatics, 9(4):286–98, Juli 2008.

[108] Katoh, K. und H. Toh: *Parallelization of the MAFFT multiple sequence alignment program.* Bioinformatics, 26(15):1899–900, August 2010.

[109] Notredame, C., D. G. Higgins und J. Heringa: *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* Journal of Molecular Biology, 302(1):205–17, September 2000.

[110] Liu, Y., B. Schmidt und D. L. Maskell: *MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities.* Bioinformatics, 26(16):1958–64, August 2010.

[111] Notredame, C. und C. Abergel: *Using multiple alignment methods to assess the quality of genomic data analysis.* In: Andrade, Miguel (Herausgeber): *Bioinformatics and genomes: Current perspectives*, Seiten 30–50. Horizon Scientific Press, Wymondham, 2003.

[112] Eswar, N., D. Eramian, B. Webb, M. Shen und A. Sali: *Protein structure modeling with MODELLER.* Methods in Molecular Biology, 426(3):145–59, Januar 2008.

[113] Zhang, Y. und J. Skolnick: *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Research, 33(7):2302–9, Januar 2005.

[114] Baker, E. N. und R. E. Hubbard: *Hydrogen bonding in globular proteins.* Progress in Biophysics and Molecular Biology, 44(2):97–179, Januar 1984.

[115] Grüning, M.: *Untersuchungen zur Diskriminanzanalyse mit hochdimensionalen Daten.* Doktorarbeit, Otto-von-Guericke-Universität Magdeburg, 2007.

[116] Kuhn, H. W.: *The Hungarian method for the assignment problem.* Naval Research Logistics Quarterly, 2(1-2):83–97, März 1955.

[117] Kabsch, W.: *A solution for the best rotation to relate two sets of vectors.* Acta Crystallographica Section A, 32(5):922–923, September 1976.

[118] 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, September 1978.

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

[120] Verdonk, M. L., J. C. Cole, M. J. Hartshorn, C. W. Murray und R. D. Taylor: *Improved protein-ligand docking using GOLD.* Proteins, 52(4):609–23, September 2003.

[121] Kramer, B., M. Rarey und T. Lengauer: *Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking.* Proteins, 37(2):228–41, November 1999.

[122] Morris, G. M., R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell und A. J. Olson: *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* Journal of Computational Chemistry, 30(16):2785–91, Dezember 2009.

[123] Paulini, F.: *Translig: Ein Algorithmus zur Ligandenpositionierung beim Design aktiver Zentren von Enzymen.* Diplomarbeit, FernUniversität Hagen, 2012.

[124] Parca, L., P. F. Gherardini, M. Helmer-Citterich und G. Ausiello: *Phosphate binding sites identification in protein structures.* Nucleic Acids Research, 39(4):1231–42, März 2011.

[125] Kinoshita, K., K. Sadanami, A. Kidera und N. Go: *Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes.* Protein Engineering, 12(1):11–4, Januar 1999.

[126] Klein, R.: *Algorithmische Geometrie.* Addison-Wesley, Bonn, 1997.

[127] Shindyalov, I. N. und P. E. Bourne: *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 11(9):739–47, September 1998.

[128] Hasegawa, H. und L. Holm: *Advances and pitfalls of protein structural alignment.* Current Opinion in Structural Biology, 19(3):341–8, Juni 2009.

[129] Moreland, J. L., A. Gramada, O. V. Buzko, Q. Zhang und P. E. Bourne: *The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications.* BMC Bioinformatics, 6:21, Januar 2005.

[130] Holland, R. C. G., T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer und M. J. Schreiber: *BioJava: an open-source framework for bioinformatics.* Bioinformatics, 24(18):2096–7, September 2008.

[131] Murray, G.: *Rotation about an arbitrary axis in 3 dimensions, Java-Bibliothek.* http://inside.mines.edu/~gmurray/ArbitraryAxisRotation/, Juli 2011.

[132] Delano, W. L.: *The PyMOL Molecular Graphics System.* http://www.pymol.org, 2002.

[133] Sander, C. und R. Schneider: *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 9(1):56–68, Januar 1991.

[134] Needleman, S. B. und C. D. Wunsch: *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of Molecular Biology, 48(3):443–53, März 1970.

[135] Smith, T. F. und M. S. Waterman: *Identification of common molecular subsequences.* Journal of Molecular Biology, 147(1):195–7, März 1981.

[136] Henikoff, S. und J. G. Henikoff: *Amino acid substitution matrices from protein blocks.* Proceedings of the National Academy of Sciences of the United States of America, 89(22):10915–9, Dezember 1992.

[137] Bairoch, A. und R. Apweiler: *The SWISS-PROT protein sequence data bank and its new supplement TREMBL.* Nucleic Acids Research, 24(1):21–5, Januar 1996.

[138] Göbel, U., C. Sander, R. Schneider und A. Valencia: *Correlated mutations and residue contacts in proteins.* Proteins, 18(4):309–17, April 1994.

[139] DeLuca, S., B. Dorr und J. Meiler: *Design of native-like proteins through an exposure-dependent environment potential.* Biochemistry, 50(40):8521–8, Oktober 2011.

[140] Hubbard, T. J., B. Ailey, S. E. Brenner, A. G. Murzin und C. Chothia: *SCOP: a Structural Classification of Proteins database.* Nucleic Acids Research, 27(1):254–6, Januar 1999.

[141] Weng, Y., D. T. Chang, Y. Huang und C. Lin: *A study on the flexibility of enzyme active sites.* BMC Bioinformatics, 12 Suppl 1(Suppl 1):S32, Januar 2011.

[142] Wang, R., X. Fang, Y. Lu und S. Wang: *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* Journal of Medicinal Chemistry, 47(12):2977–80, Juni 2004.

[143] Wang, R., X. Fang, Y. Lu, C. Yang und S. Wang: *The PDBbind database: methodologies and updates.* Journal of Medicinal Chemistry, 48(12):4111–9, Juni 2005.

[144] Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi und M. Tasumi: *The Protein Data Bank. A computer-based archival file for macromolecular structures.* European Journal of Biochemistry, 80(2):319–324, November 1977.

[145] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov und P. E. Bourne: *The Protein Data Bank.* Nucleic Acids Research, 28(1):235–42, Januar 2000.

[146] Sander, C. und R. Schneider: *The HSSP data base of protein structure-sequence alignments.* Nucleic Acids Research, 21(13):3105–9, Juli 1993.

[147] Wang, G. und R. L. Dunbrack: *PISCES: recent improvements to a PDB sequence culling server.* Nucleic Acids Research, 33(Web Server issue):W94–8, Juli 2005.

[148] Laskowski, R. A.: *PDBsum new things.* Nucleic Acids Research, 37(Database issue):D355–9, Januar 2009.

[149] Laskowski, R. A., E. G. Hutchinson, A. D. Michie, A. C. Wallace, M. L. Jones und J. M. Thornton: *PDBsum: a Web-based database of summaries and analyses of all PDB structures.* Trends in Biochemical Sciences, 22(12):488–90, Dezember 1997.

[150] Xiang, J.: *A protein structure modeling package*, 2002.

[151] Zhang, Y.: *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 9:40, Januar 2008.

[152] Richter, F., R. Blomberg, S. D. Khare, G. Kiss, A. P. Kuzin, A. J. T. Smith, J. L. Gallaher, Z. Pianowski, R. C. Helgeson, A. Grjasnow, R. Xiao, J. Seetharaman, M. Su, S. Vorobiev, S. Lew, F. Forouhar, G. J. Kornhaber, J. F. Hunt, G. T. Montelione, L. Tong, K. N. Houk, D. Hilvert und D. Baker: *Computational design of catalytic dyads and oxyanion holes for ester hydrolysis.* Journal of the American Chemical Society, August 2012.

[153] Shimoni, L. und J. P. Glusker: *Hydrogen bonding motifs of protein side chains: descriptions of binding of arginine and amide groups.* Protein science, 4(1):65–74, Januar 1995.

[154] Borders, C. L., J. A. Broadwater, P. A. Bekeny, J. E. Salmon, A. S. Lee, A. M. Eldridge und V. B. Pett: *A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens.* Protein science, 3(4):541–8, April 1994.

[155] Liu, Y. und B. Kuhlman: *RosettaDesign server for protein design.* Nucleic Acids Research, 34(Web Server issue):W235–8, Juli 2006.

[156] Zhang, Y. und J. Skolnick: *SPICKER: a clustering approach to identify near-native protein folds.* Journal of Computational Chemistry, 25(6):865–71, April 2004.

[157] Hooft, R. W., G. Vriend, C. Sander und E. E. Abola: *Errors in protein structures.* Nature, 381(6580):272, Mai 1996.

[158] Joosten, R. P., J. Salzemann, V. Bloch, H. Stockinger, A. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A. L. Da Costa, G. Deleage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I. J. Tickle und G. Vriend: *PDB_REDO: automated re-refinement of X-ray structure models in the PDB.* Journal of Applied Crystallography, 42(3):376–384, April 2009.

[159] Sterner, R., G. R. Kleemann, H. Szadkowski, A. Lustig, M. Hennig und K. Kirschner: *Phosphoribosyl anthranilate isomerase from Thermotoga maritima is an extremely stable and active homodimer.* Protein science, 5(10):2000–8, Oktober 1996.

[160] Due, A. V., J. Kuper, A. Geerlof, J. P. von Kries und M. Wilmanns: *Bisubstrate specificity in histidine/tryptophan biosynthesis isomerase from Mycobacterium tuberculosis by active site metamorphosis.* Proceedings of the National Academy of Sciences of the United States of America, 108(9):3554–9, März 2011.

[161] Henn-Sax, M., R. Thoma, S. Schmidt, M. Hennig, K. Kirschner und R. Sterner: *Two* $(\beta\alpha)_8$*-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates.* Biochemistry, 41(40):12032–42, Oktober 2002.

[162] Claren, J., C. Malisi, B. Höcker und R. Sterner: *Establishing wild-type levels of catalytic activity on natural and artificial* $(\beta\alpha)_8$*-barrel protein scaffolds.* Proceedings of the National Academy of Sciences of the United States of America, 106(10):3704–9, März 2009.

[163] Reisinger, B., M. Bocola, F. List, J. Claren, C. Rajendran und R. Sterner: *A sugar isomerization reaction established on various* $(\beta\alpha)_8$*-barrel scaffolds is based on substrate-assisted catalysis.* Protein Engineering, Design, and Selection, 2012. (in Revision).

[164] Leopoldseder, S., J. Claren, C. Jürgens und R. Sterner: *Interconverting the catalytic activities of* $(\beta\alpha)_8$*-barrel enzymes from different metabolic pathways: sequence requirements and molecular analysis.* Journal of Molecular Biology, 337(4):871–9, April 2004.

[165] Grosdidier, A., V. Zoete und O. Michielin: *SwissDock, a protein-small molecule docking web service based on EADock DSS.* Nucleic Acids Research, 39(Web Server issue):W270–7, Juli 2011.

[166] Grosdidier, A., V. Zoete und O. Michielin: *Fast docking using the CHARMM force field with EADock DSS.* Journal of Computational Chemistry, Mai 2011.

[167] Bas, D. C., D. M. Rogers und J. H. Jensen: *Very fast prediction and rationalization of* pK_a *values for protein-ligand complexes.* Proteins, 73(3):765–83, November 2008.

[168] Kortemme, T., A. V. Morozov und D. Baker: *An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes.* Journal of Molecular Biology, 326(4):1239–1259, Februar 2003.

[169] Leaver-Fay, A., G. L. Butterfoss, J. Snoeyink und B. Kuhlman: *Maintaining solvent accessible surface area under rotamer substitution for protein design.* Journal of Computational Chemistry, 28(8):1336–41, Juni 2007.

[170] Pokala, N. und T. M. Handel: *Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation.* Protein science, 13(4):925–36, April 2004.

[171] Zanghellini, A., L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Röthlisberger und D. Baker: *New algorithms and an in silico benchmark for computational enzyme design.* Protein science, 15(12):2785–94, Dezember 2006.

[172] Lassila, J. K., H. K. Privett, B. D. Allen und S. L. Mayo: *Combinatorial methods for small-molecule placement in computational enzyme design.* Proceedings of the National Academy of Sciences of the United States of America, 103(45):16710–5, November 2006.

[173] Søndergaard, C. R., M. H. M. Olsson, M. Rostkowski und J. H. Jensen: *Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pK_a values.* Journal of Chemical Theory and Computation, 7(7):2284–2295, Juli 2011.

[174] Krebs, E. G. und J. A. Beavo: *Phosphorylation-dephosphorylation of enzymes.* Annual Review of Biochemistry, 48:923–59, Januar 1979.

[175] Gallicchio, E. und R. M. Levy: *Recent theoretical and computational advances for modeling protein-ligand binding affinities.* Advances in Protein Chemistry and Structural Biology, 85:27–80, Januar 2011.

[176] Allen, B. D., A. Nisthal und S. L. Mayo: *Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles.* Proceedings of the National Academy of Sciences, November 2010.

[177] Wang, C., O. Schueler-Furman und D. Baker: *Improved side-chain modeling for protein-protein docking.* Protein science, 14(5):1328–39, Mai 2005.

[178] Kellogg, E. H., A. Leaver-Fay und D. Baker: *Role of conformational sampling in computing mutation induced changes in protein structure and stability.* Proteins: Structure, Function, and Bioinformatics, Seiten 830–838, 2010.

[179] Lei, Y., W. Luo und Y. Zhu: *A matching algorithm for catalytic residue site selection in computational enzyme design.* Protein science, 20(9):1566–1575, Juni 2011.

[180] Malisi, C., O. Kohlbacher und B. Höcker: *Automated scaffold selection for enzyme design.* Proteins, 77(1):74–83, Oktober 2009.

[181] Eisenbeis, S., W. Proffitt, M. Coles, V. Truffault, S. Shanmugaratnam, J. Meiler und B. Höcker: *Potential of fragment recombination for rational design of proteins.* Journal of the American Chemical Society, 134(9):4019–22, März 2012.

[182] Althoff, E., L. Wang, L. Jiang, L. Giger, J. K. Lassila, Z. Wang, M. Smith, S. Hari, P. Kast, D. Herschlag, D. Hilvert und D. Baker: *Robust design and optimization of retroaldol enzymes.* Protein science, 21(5):717–26, Mai 2012.

[183] Schreier, B., C. Stumpp, S. Wiesner und B. Höcker: *Computational design of ligand binding is not a solved problem.* Proceedings of the National Academy of Sciences of the United States of America, 106(44):18491–6, November 2009.